

**UC Berkeley**

**Dissertations, Department of Linguistics**

**Title**

The Role of Dynamic Cues in Speech Perception, Spoken Word Recognition, and Phonological Universals

**Permalink**

<https://escholarship.org/uc/item/8f3393rd>

**Author**

Warner, Natasha

**Publication Date**

1998

**The Role of Dynamic Cues in Speech Perception, Spoken Word Recognition,  
and Phonological Universals**

by

**Natasha Lynn Warner**

**B.A. (University of Minnesota) 1991  
M.A. (University of California, Berkeley) 1995**

**A dissertation submitted in partial satisfaction of the**

**requirements for the degree of**

**Doctor of Philosophy  
in**

**Linguistics**

**in the**

**GRADUATE DIVISION**

**of the**

**UNIVERSITY OF CALIFORNIA, BERKELEY**

**Committee in charge:**

**Professor John J. Ohala, Chair  
Professor Sharon Inkelas  
Professor James Matisoff  
Professor Yoko Hasegawa**

**Fall 1998**

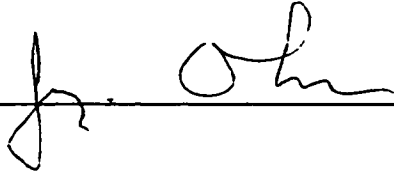
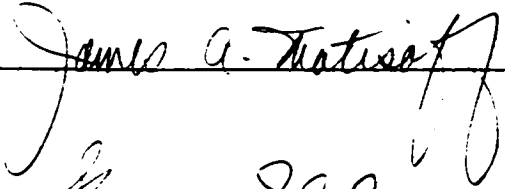
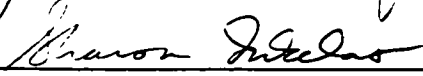

**The Role of Dynamic Cues in Speech Perception,  
Spoken Word Recognition, and Phonological Universals**

**Copyright 1998**

**by**

**Natasha Lynn Warner**

The dissertation of Natasha Lynn Warner is approved:

Chair		June 16, 1998
		Date
		June 4, 1998
		Date
		June 14, 1998
		Date
		June 15, 1998
		Date

University of California, Berkeley

Fall 1998

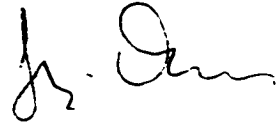
**Abstract****The Role of Dynamic Cues in Speech Perception,  
Spoken Word Recognition, and Phonological Universals****by****Natasha Lynn Warner****Doctor of Philosophy in Linguistics****University of California, Berkeley****Professor John J. Ohala, Chair**

**This study investigates whether information in the speech signal is distributed uniformly or whether it is concentrated in certain regions, and, if the latter, whether these regions correspond to areas of acoustic change. Traditionally steady states, which more closely approximate "targets," have been considered important, but because the auditory system is especially sensitive to stimulus modulations, speech might be perceived primarily through dynamic cues.**

**Expanding Furui's (1986) perceptual study of gated speech (but using an open response method), I located points in the acoustic speech signal of Japanese and English that were most effective in reducing the lexical cohort in word recognition and in allowing listeners to perceive segments correctly. I compared these points with points of maximal acoustic change (as defined by Furui), and found that they occur together significantly more often than by chance, showing that listeners use predominantly dynamic cues. I also found that additional phonetic information sometimes caused listeners to increase the word cohort instead of helping them to narrow it down, a result theories of spoken word recognition would not predict.**

**The results of this study show that the phonological system of a language affects how listeners use phonetic cues: because Japanese makes no place distinctions in coda position, Japanese listeners are slower than English listeners to make use of cues in a**

vowel-consonant transition. Furthermore, the importance of dynamic cues provides perceptual motivations for several common phonological patterns. Dissimilation, the special status of onsets relative to codas, and the common alternation of glides with high vowels may all relate to how long listeners require to perceive distinctions. For example, stops were perceived more quickly through CV than VC cues. Since stops in onset position have the rapid CV cues, while stops in coda position do not, and rapid changes are important for perception, this result offers a motivation for the special status of onsets. There was also long-lasting perceptual confusion between glides and high vowels, explaining the common alternations between these segments. These results further Stevens' (1985) and Steriade's (1997) work regarding qualitative differences between features and cues and their utilization in languages.

A handwritten signature in black ink, appearing to be 'J. J. J.' or similar, located in the lower right quadrant of the page.

## Table of Contents

List of Figures	ix
List of Tables	xiii
Acknowledgments	xv
1. Introduction	1
1.1. General introduction	1
1.2. The dynamic theory of speech perception	7
1.2.1. The traditional bias toward steady states	7
1.2.2. The importance of changes for auditory processing	8
1.2.3. Experimental studies of dynamic and steady information in speech perception	10
1.2.3.1. Perception of features of consonants	10
1.2.3.2. Perception of vowels	18
1.2.3.3. Perception of a variety of speech sounds	20
1.2.4. The importance of change in the signal seen through phonological universals	23
1.3. Perception of segments and recognition of words	25
1.4. The use of gating experiments	28
1.4.1. Overview of the gating method	28
1.4.2. Topics investigated through gating	29
1.4.3. A problem: Gating as altering rather than removing cues	31
1.4.4. Open versus closed class responses for gating	32
1.4.5. Individual versus successive presentation	33
1.4.6. Gating as a reflection of online processing	37
1.4.7. Ways of analyzing gating data	38
1.4.8. The gating interval	39
1.5. Language specific differences in perception and spoken	40

word recognition	
1.6. Summary	48
2. Methodology of the perception experiment	50
2.1 The English experiment	50
2.1.1. Choice of stimulus words	50
2.1.1.1. The two segment sequence for gating	50
2.1.1.2 Factors manipulated in the choice of transitions	51
2.1.1.3 The word list	53
2.1.1.4 Other potential effects	61
2.1.2 Production of the stimuli	66
2.1.2.1 Recording of the data	66
2.1.2.2 Gating	66
2.1.3 Subjects and procedures during the experiment	71
2.2. The Japanese experiment	78
2.2.1 Differences from the English methods	78
2.2.2 The word list	80
2.2.3 Production of stimuli	85
2.2.4 Subjects and procedures	86
3. Degree of spectral change	91
3.1. Furui's measure of degree of spectral change D	92
3.1.1. Calculation of the measure D	92
3.1.2. Method of locating the $D_{\max}$ point or points	95
3.1.2.1. Cases with only one peak of D in the gated area	95
3.1.2.2. Cases with more than one acoustic change expected in the gated area	97
3.1.2.3. Peaks associated with changes other than the transition of interest	101



3.1.2.4. Changes in the signal which are not always expected	107
3.1.2.5. Highest value of D not a peak	112
3.1.2.6. Summary of criteria for locating $D_{\max}$ points	112
3.2. Problems with the measure D	114
3.2.1. Changes the measure D does not reflect well	114
3.2.1.1. Insensitivity of D to linguistically relevant gradual changes	114
3.2.1.2. Sensitivity of D to linguistically irrelevant changes	124
3.2.1.3. Importance of sensitivity to amplitude and frequency changes	127
3.2.1.4. Criteria for selecting $D_{\max}$ when D does not reflect changes well	127
3.2.2. Temporal inaccuracy	128
3.2.3. Suggestions for alternative measures of spectral change	130
3.2.3.1. Possible alternative measures not yet implemented	130
3.2.3.2. The issue of sampling rate	134
3.2.3.3. Results of one alternative measure of degree of spectral change	136
3.3. Spectral change measurements for English and Japanese acoustic data	138
4. Results	143
4.1 Method of analyzing the data	143
4.1.1. Transcribing the responses	143
4.1.2. Conversion to numerical data	145
4.1.2.1. The number of responses measure	145
4.1.2.2. The percent correct measure	146
4.1.3. Data excluded for anomalous slopes	151
4.1.4. Fitting of ogival curves	153
4.1.5. Exclusion of linear data	154

4.2. Comparison of location of maximal change in the perceptual measures to location of point of maximal spectral change	159
4.2.1. Observed results	162
4.2.2. Probability of overlap by chance	173
4.3. Analysis of cases in which perceptual measures are not related to the D <sub>max</sub> point	182
4.3.1. Categories of exceptions	182
4.3.1.1. Postvocalic stops	182
4.3.1.2. Transitions into vowels	185
4.3.1.3. Vowel-Vowel transitions	188
4.3.1.4. Transitions into sonorants	190
4.3.1.5. Fricatives	190
4.3.1.6. Exceptions specific to Japanese	193
4.3.2. Classification of words by potential exception category	196
4.3.3. Classification of words' results into these categories	200
4.3.4. Statistical tests of results	208
4.3.4.1. Percent of words showing each effect	208
4.3.4.2. Distance of area of maximal perceptual change from D <sub>max</sub>	210
4.3.4.3. Results for words not in any category of exceptions	211
4.4. Analysis of cases in which the two perceptual measures differ	214
4.4.1. Rationale for comparing the timing of perception of segments and recognition of words	214
4.4.2. Overall results of comparing the two perceptual measures	215
4.4.3. Cases where maximal change in number of responses represents perception of something other than the target segment	220
4.4.4. Cases in which the cohort is too large	224
4.4.5. Perceptual mismatch due to problems with the #Resp measure	225

4.4.6. Cases in which curve fitting is too successful	228
4.4.7. The need for more subjects	230
4.5. Rise time of perceptual curves	232
4.5.1. Reasons for examining the rise time	232
4.5.2. Method of calculating rise time	234
4.5.3. Rise time results	235
4.5.4. Statistical analysis	240
4.6. Perception of word initial stops	243
5. Conclusions	247
5.1. Dynamic perception of segments	247
5.1.1. Distribution of information in the signal	247
5.1.2. Slow and fast changes and the sensitivity of the auditory system	247
5.1.3. Implications of the experiment for fast and slow changes	253
5.1.4. The meaning of alignment with Dmax	256
5.2. Dynamic perception and spoken word recognition	261
5.3. Implications for the cohort model	265
5.3.1. Recognition need not proceed from the first phoneme	265
5.3.2. Increases in the number of candidates over time	270
5.3.3. The size of the unit used to narrow the cohort	278
5.4. Effects of suprasegmental units	282
5.5. Language specific effects of phonology on perception	285
5.5.1. Effect of phoneme inventory	286
5.5.2. Effect of syllable structure constraints	289
5.6. Perceptual motivations for phonological universals or alternations	292
5.6.1. The preference for onsets over codas	293
5.6.2. Alternations between glides and high vowel nuclei	297

	viii
5.6.3. Dissimilation	303
5.7. Integrating speech perception and formal phonology	306
5.7.1. Perception by comparison to the UR	308
5.7.2. Perception as the basis for formal constraints	315
5.8. Overall conclusions	317
Works cited	322
Appendix A—Subjects' language backgrounds	335
Appendix B— Graphs of results	339
Appendix C—Graphs of results for word initial stops	392

## List of Figures

2.1	Gated stimuli for the word "Disney" /dɪzni/	69
2.2	The screen subjects for the English experiment used	74
3.1.	The quantity $a_i$	93
3.2	Waveform, spectrogram, and measure D for the word "chapel" /tʃæ/	96
3.3	Waveform, spectrogram, and D for the word /hakama/ 'Japanese style pleated skirt'	98
3.4	Waveform, spectrogram, and D for the word "skull" /kʌ/	100
3.5	Waveform, spectrogram, and D for the word "citizen" [ɪr]	102
3.6	Waveform, spectrogram, and D for the word "committee" [ɪr]	103
3.7	Waveform, spectrogram, and D for the word "fair" /eɪr/	105
3.8	Waveform, spectrogram, and D for the word /sakka/ 'author' [kk]	106
3.9	Waveform, spectrogram, and D for the word "ranch" /rʌntʃ/	109
3.10	Waveform, spectrogram, and D for the word "snow" /sn/	110
3.11	Waveform, spectrogram, and D for the word /syaberu/ 'to chat' [ʃa]	111
3.12	Waveform, spectrogram, and D for the word "caboose" /kə/	113
3.13	Waveform, spectrogram, and D for the word "biotech" /aɪoʊ/	116
3.14	Waveform, spectrogram, and D for the word /hatake/ 'field' /ak/	117
3.15	Waveform, spectrogram, and D for the word "elevator" /el/	118
3.16	Waveform, spectrogram, and D for the word /koNyaku/ [konjaku] 'engagement'	119
3.17	Waveform, spectrogram, and D for the word "custom" /kʌ/	120

3.18	Waveform, spectrogram, and D for the word "shell" /ʃe/	121
3.19	Waveform, spectrogram, and D for the word /seNmoN/ [semmon] 'specialization'	122
3.20	Waveform, spectrogram, and D for the word /maNneNhitu/ [mannençtsu] 'fountain pen'	123
3.21	Waveform, spectrogram, and D for the word "trail" /re/	125
3.22	Waveform, spectrogram, and D for the word "groan" /gr/	126
3.23.	A hypothetical case in which too long a window was used for calculating the slope of the regression line of the cepstral coefficient values over time	129
3.24	Waveform, spectrogram, and D for the word "academic" /kə/	131
4.1	Number of responses and percent correct data for the word "circle" /ə:k/	152
4.2	Parameters of the ogival curves	155-156
4.3	The screen used for fitting curves to data	157
4.4	Data, best linear fit, and fitted ogival curve for the number of responses data for /kiNtyoo/ [kintʃoo] 'nervousness'	160
4.5	Spectrogram, measure D, and the two perceptual measures for the word "band" /bænd/	171
4.6	Spectrogram, measure D, and the two perceptual measures for /baso/ [baso] 'place'	172
4.7	A hypothetical case in which there is exactly one $D_{\max}$ point	175
4.8	A hypothetical case in which there is exactly one $D_{\max}$ point, and it is near the endpoint of the first gate	175
4.9	A hypothetical case in which there is exactly one $D_{\max}$ point, and it is located between the penultimate and ultimate gate endpoints	176

4.10	A hypothetical case in which there is exactly one $D_{\max}$ point, and it is just before the endpoint of the penultimate gate	178
4.11	A hypothetical case in which there are two $D_{\max}$ points which are separated from each other by more than 25 ms	178
4.12	A hypothetical case in which there are two $D_{\max}$ points near each other	179
4.13	Spectrogram, measure D, and the two perceptual measures for "oats" /o <sup>◌</sup> ts/	183
4.14	Spectrogram, measure D, and the two perceptual measures for /hatake/ [hatake] 'field'	184
4.15	Spectrogram, measure D, and the two perceptual measures for "toad" /to <sup>◌</sup> d/	186
4.16	Spectrogram, measure D, and the two perceptual measures for /kato/ [kato] 'crossing'	187
4.17	Spectrogram, measure D, and the two perceptual measures for "diagonal" /da <sup>i</sup> ægə <sup>n</sup> əl/	189
4.18	Spectrogram, measure D, and the two perceptual measures for "fair" /fe <sup>i</sup> r/	191
4.19	Spectrogram, measure D, and the two perceptual measures for /kyaku/ [kjaku] 'guest'	192
4.20	Spectrogram, measure D, and the two perceptual measures for "leaf" /lif/	194
4.21	Spectrogram, measure D, and the two perceptual measures for /keego/ [keego] 'honorific language'	195
4.22	Spectrogram, measure D, and the two perceptual measures for /huyoo/ [ɸujoo] 'unnecessary'	197
4.23	Percent correct (%Corr) data for two words	217
4.24	#Resp and %Corr curves for "session" /ɛʃ/	219
4.25	#Resp data for /megumi/ 'grace'	229

4.26	Calculation of rise times for two examples (%Corr data)	236
5.1	The response of the measure D for some hypothetical two segment sequence with two peaks of D within the gated area	258
5.2	Waveform, spectrogram, D, and %Corr data for the word "committee" [ɪr]	260



## List of Tables

3.1.	Criteria for counting peaks of the measure D as $D_{\max}$ points	114
3.2.	Locations of $D_{\max}$ points for all words in the experiment	139
4.1.	Comparison of location of $D_{\max}$ point(s) and location of area of maximal slope of the two perceptual measures	163
4.2.	Number and percentage of words for which the area of maximal change in the fitted curve surrounds a $D_{\max}$ point	170
4.3.	Actual and predicted (chance) numbers and percentages of words with area of maximal perceptual change surrounding $D_{\max}$	181
4.4.	Categorization of words by effects which can lead to the area of most change in the %Corr measure not surrounding $D_{\max}$	201
4.5.	Summary of results for potential and actual effects	208
4.6.	Actual and predicted (chance) numbers and percentages of words which do not fall into any of the exception categories with area of maximal perceptual change surrounding $D_{\max}$ for each language	212
4.7.	Number of gates separating the area of maximal change in %Corr from the area of maximal change in #Resp	215
4.8.	Responses to "Italian" /rt/	220
4.9.	Responses to "mechanical" /ək /	221
4.10.	Responses to "crops" /kr /	222
4.11.	Responses to "soybean" /oʊb /	227
4.12.	Responses to "attempt" /ɛm/	228
4.13.	Rise times of the %Corr measure	237
4.14.	Rise time of the %Corr measure for word initial stops	244

4.15. Average rise time of %Corr data for stops in word-initial, postvocalic onset, and coda positions	244
5.1. Results for words with more than one $D_{\max}$ point, with regard to which $D_{\max}$ point the area of maximal perceptual change (in %Corr) surrounds	259
5.2. Responses to "asthma" /æz/	273
5.3. Responses to "unconcealed" /ns/	274
5.4. Responses to /himoto/ 'origin of a fire'	275
5.5. Responses to the stimulus "eon" with initial palatal glides	298
5.6. Responses to the stimulus /huyoo/ [ɸujoo] 'unnecessary' with syllable structures other than the intended one	299
5.7. Responses to the stimulus /mawari/ [mawari] 'surroundings' with syllable structures other than the intended one	301
5.8. Responses to English stimuli with postvocalic voiceless or voiced obstruents	313

## Acknowledgments

I am greatly indebted to the members of my committee, Sharon Inkelas, Jim Matisoff, Yoko Hasegawa, and especially to my advisor, John Ohala, for all the advice and help they gave me on this dissertation. John Ohala has taught me so much about speech, experimental phonology, scientific method, and how to design experiments. I am very grateful to him for all he has taught me throughout the time I have been in graduate school, and also for his encouragement. I have been very fortunate to study with him.

I would like to thank all of the people, in addition to my committee members, who discussed this material with me and helped me understand it better: Takayuki Arai, Chuck Fillmore, Steven Greenberg, Shawn Ying, Madelaine Plauché, Ani Patel, and María Josep Solé, as well as the many people who gave me helpful comments after hearing me present talks on this material. In addition to some of the people mentioned above, Khalil Iskarous, Stefan Frisch, and Herb Clark gave me helpful suggestions on how to devise a better measure of degree of spectral change. Any errors are, of course, my own. Furthermore, I am very grateful to Shawn Ying for writing computer programs for me (usually on short notice), rewiring the sound booth, and keeping all of the computers in the lab running.

Koji Nabeshima and Tomoko Smith were very gracious about helping me when a native speaker of Japanese was necessary. I am especially grateful to the two speakers for the experiment, as well as to the approximately 350 people who participated in either the English or Japanese experiment or a pilot version of this experiment. Thanks are also due to Susanne Gahl, John Mcwhorter, Tom Shannon, John Ohala, and Jim Matisoff for allowing me to recruit their students as subjects, and for finding ways to reward the students for their participation. The graduate student instructors of Linguistics 5 and 55 were also very helpful in allowing me to visit their sections to recruit subjects. I am extremely thankful for the help of Professor Kyoko Ohara of Keio University, Professor Seiya Matsumoto of Aichi Shukutoku University, and Setsuko Imatomi of Sophia University for recruiting the subjects in Japan and arranging the facilities for running the

experiment there. Without their help, and that of Takayuki Arai and Kimiko Akita, it would not have been possible for me to do the Japanese part of this project.

I was supported by the Berkeley Fellowship for Graduate Studies while doing this work, and the project in Japan was funded by a grant from the Vice Chancellor's Research Fund. I am very grateful for both of these sources of support, as well as for a grant from the Linguistics Department which allowed me to attend the LSA to present this work. I would also like to thank Belén Flores and Paula Floro for helping me with administrative issues.

Finally, I would like to thank my parents and my friends at Berkeley for helping me get through the process of writing the dissertation. Most of all, I would like to thank my husband, Keith Alcock, for discussing the results with me at length, writing numerous computer programs to make it easier to analyze the data, and in general being supportive.

## 1. Introduction

### 1.1. General introduction

One of the most general questions about speech perception is how information is distributed in the speech signal. Is information spread equally throughout the signal in time, or are there certain parts of the signal in which information is concentrated? If there are parts of the signal which carry more information than others, times of high information flow, then one important task of speech perception research is to determine what the characteristics of these high information flow regions are, and why these regions carry a disproportionate amount of information. One possibility is that parts of the signal with rapid acoustic change are areas with a high concentration of information, because transitions from one sound to another can carry information about both the sound before and the sound after them, and because the auditory system is especially sensitive to acoustic changes.

A speech signal contains periods of rapid change, periods of slower change, and periods during which a sound stays relatively steady. Rapid changes include the bursts of stops and affricates. Formant transitions into or out of a consonant or movement of formants during a diphthong are inherently changing, but change more slowly. The steady state portion of a monophthongal vowel or of a fricative such as [s] changes very little, even over a period of 100 ms or more. When listeners perceive speech, one may ask whether they accomplish the task of recognizing sounds (whether as distinctive features, phonemes, syllables, or some other unit) by attending primarily to the changing aspects of the signal—perception through dynamic cues—or by concentrating on the steady parts of the signal where they are available, or whether changing and steady aspects are equally useful. Perception in general (visual perception as well as auditory non-speech perception) is more sensitive to changes in signals than to static signals. One might expect that speech would be perceived the same way. Listeners might make use of both dynamic and steady information, perhaps weighting the two types of information in some way. Because of the

many types of distinctions to be perceived in speech, the type of sound to be perceived, as well as the environment in which it appears, may affect this weighting. Furthermore, listeners' use of dynamic or steady information might influence the timing of recognition of words as well as perception of individual sounds.

There may also be language specific effects in the choice of dynamic or steady cues. Such language specific differences might be caused by factors in the phonological system: for example, if there are important dynamic cues in vowel-consonant sequences, the extent to which a language's listeners use these cues might be affected by whether the language has CVC as well as CV syllables. Furthermore, a preference for dynamic or static cues might be reflected in phonological patterns which are common in the world's languages. For example, if dynamic cues are important for the perception of some type of segment, there might be common phonological rules which result in that segment appearing in environments where those dynamic cues are available, rather than in environments where only static cues would be present.

In this dissertation, I report on a gating experiment designed to determine whether listeners make more use of dynamic cues than static ones in speech perception. In a gating experiment, one presents listeners with parts of a word, in this case with the beginning of a word, cutting the word off at various points. One asks the listeners to say what word this might have been the beginning of. By examining the types of responses given when listeners were allowed to hear differing parts of the word, one can determine where in the signal crucial cues for distinctions are located.

The specific method applied here to determine whether listeners make disproportionate use of dynamic cues rather than steady ones is the method used by Furui (1986). Furui defined a measure of degree of spectral change, which he called D, based on

the amount of change in the cepstral coefficients over time<sup>1</sup>. He then identified the point at which listeners' identification of segments reached 80% correct, and found that this point was almost always within a few milliseconds after the point of maximal spectral transition (the point in the signal with the greatest value of D). He concluded that since listeners become able to identify segments just after hearing the part of the signal with the most change in the spectral information, they must be using dynamic cues. This was a step forward from previous work on dynamic cues, because the definition of an objective measure of degree of spectral change allows one to ask whether listeners' perception improves at the point of the signal with dynamic cues regardless of what those dynamic cues might be. Previous work on the subject had only tested dynamic versus static cues in a particular distinction (such as place of articulation in a stop-vowel sequence), but Furui could test use of dynamic cues more generally.

However, Furui's experiment still tested a very limited group of transitions, namely the CV and CyV<sup>2</sup> transitions which are possible in Japanese. One of the major advances in the experiment reported here is that I apply the test of dynamic cues to a very wide range of segment transitions, providing the most general test of dynamic versus static cues yet. This is important, because even if listeners do perceive a particular feature in a particular environment based on dynamic cues, such as place of articulation in stop-vowel sequences, this does not show that they also perceive manner of articulation in coda consonants through dynamic cues, for example.

To carry out a general test of the hypothesis of dynamic speech perception, I adopt Furui's method of calculating the degree of spectral change for the acoustic data, locating the point of maximal spectral change, and comparing the location of that point to the

---

<sup>1</sup> This is perhaps better termed a measure of estimated degree of spectral change, since the best method for calculating spectral change is uncertain. Furui's measure D does provide an objective measure, but further research might determine a better method of calculating degree of spectral change.

<sup>2</sup> For phonemic transcriptions of Japanese, /y/ will be used for the palatal glide instead of IPA /j/. Japanese phonetic transcriptions and both phonetic and phonemic transcriptions of English will use IPA.

location of a point which is important for perception. Furui used the point at which listeners reach 80% correct as the criterion for perception, but this alone does not really show that the improvement in perception takes place around the point of maximal spectral change, which is the prediction of dynamic speech perception. Furui further calculated the amount of change in correct identification over the 10 ms preceding the point at which listeners reached 80% correct. This does show that listeners in his experiment quickly became able to perceive segments around that point. Instead of using these two separate calculations, I located the area which was most important for perception not by any set criterion such as 80% correct, but rather by finding the part of the signal during which listeners most quickly improved in their perception of the target segment. I then compared the location of this area to the location of the point of maximal spectral change of the acoustic signal. If the area of the most perceptual improvement is at the point of maximal spectral change, one can conclude that listeners are probably using dynamic cues to perceive the segment.

This is not a direct test of the hypothesis that listeners make disproportionate use of dynamic cues. A more direct test would be provided by synthesizing stimuli with and without changing information. This has been done before, and will be discussed in section 1.2 below. However, this method can again only be applied to a particular distinction, since one must define what is dynamic and what is static about the signal for the particular distinction in order to remove dynamic aspects from the signal in synthesis. I therefore preferred to use an indirect test of the hypothesis. That is, my experiment tests whether the part of the signal which allows listeners to make the most progress toward perceiving a segment is also the part of the signal with the most change in the spectral information. If the perceptually important area overlaps the point with the most acoustic change, at least if it does so for a significant proportion of the cases tested, I will assume that this means listeners are using dynamic information to perceive the sound.



In a gating experiment, one can use either an open or a closed response test. If listeners are simply asked what word they heard the beginning of (open response), the experiment will provide information about the time course of spoken word recognition as well as about perception of individual segments. By evaluating the number of different responses given by the entire group of listeners to hear various portions of a word, one can see whether there are points at which the listeners quickly converge on a smaller number of responses, and whether these points are also near the point of maximal spectral change. One can thus compare the timing of perception of segments to the timing of recognition of the lexical item. Furthermore, while evaluating correctness of perception of a certain segment requires the experimenter to define what counts as a correct or incorrect response, examining the number of different responses given does not. The experiment reported here uses an open response method, and thus allows one to determine first, whether listeners perceive segments by using dynamic cues rather than steady ones, second, whether listeners' progress in recognizing the word also takes place where dynamic cues are available, and third, how closely related recognition of segments and recognition of the word are.

Another aspect of the current experiment which is new for the study of dynamic speech perception is that it is a cross-linguistic test, using both Japanese and English. Previous research (discussed in section 1.5 below) has shown that Japanese and English have very different properties for speech perception, at several levels. The two languages might also show some differences in the degree to which listeners use dynamic cues, and listeners of the two languages might use dynamic cues in different environments. The segmental and suprasegmental phonological structure of the two languages is quite different, and this might affect listeners' use of dynamic and static cues.

The major result of the experiment presented here is that listeners do quickly become able to perceive segments correctly right at the point of maximal spectral change (the  $D_{\max}$  point) in a significantly greater than chance proportion of the cases tested.

However, in approximately half the cases tested, they do not make the most progress in perception exactly at the point of maximal spectral change. In Section 4.3, I identify several categories of transitions which do not follow the prediction of perception at the point of maximal spectral change, and offer explanations based in well known phonetic facts for why these types of transitions fail to follow the prediction.

My results show several differences between Japanese and English in use of dynamic cues. In Chapter 5, I relate some of these differences to the phoneme inventories of the two languages, which result in different constraints on cues necessary for perception, and relate others to effects of the phonotactic system of the languages on perception. As for the comparison of spoken word recognition and perception of individual segments, I find that in most cases, the two do progress together—when listeners recognize a segment or some distinctive feature of a segment, this allows them to narrow in on the word they are perceiving. However, in Section 4.4, I identify some situations in which progress toward recognizing the word and toward perceiving the segment do not take place at the same time.

Finally, I have identified several cases in which listeners' use of dynamic cues may be reflected in phonological patterns. Some common phonological alternations (such as between glides and high vowels), some well known phonological universals (the preference for onsets over codas), and some historical phonological changes (such as dissimilation) may have roots in the use of dynamic cues. The results of this experiment offer evidence for the perceptual motivation of these patterns.

In sum, this dissertation reports on an experiment designed to test for use of dynamic versus static cues in speech perception and word recognition. It tests the hypothesis that dynamic cues are the more important in a far wider variety of environments than have previously been tested, and also offers cross-linguistic evidence and a comparison with spoken word recognition. I find evidence in favor of perception through dynamic cues, but I also find that the use of dynamic cues is strongly affected by the type

of distinction to be perceived and numerous factors of its environment. Finally, I propose that effects of dynamic cues can be seen in a variety of well known phonological patterns.

The structure of the dissertation is as follows: the remainder of this chapter provides the necessary background on several topics which are important to the study reported here, such as previous research on dynamic cues in perception, spoken word recognition, the gating methodology, and language specific differences in perception. The second chapter describes the methods used for the current experiment. The third chapter presents the analysis of the acoustic data for degree of spectral change. The fourth chapter comprises the analysis of the results of the perception experiment and the comparison of perceptual and acoustic results. The final chapter elucidates the implications of the experimental results for the hypothesis of speech perception through dynamic cues and a variety of other theoretical topics.

## 1.2. The dynamic theory of speech perception

### 1.2.1. The traditional bias toward steady states

The steady state portions of a signal have traditionally been considered more important than changing aspects of the signal for speech perception. This is especially true of vowels, because it is relatively easy to separate the steady portion of a vowel from the changing portions (formant transitions near surrounding consonants) on a spectrogram, and because the formant transitions in a CVC syllable are clearly caused by the adjoining consonants. The steady state portion of a vowel is thought to have its formants at the prototypical frequencies for that vowel, while the parts of the vowel near a consonant may have their formant frequencies altered drastically. Therefore, there is a tendency to assume that listeners perceive vowels by paying attention to the steady state portion of the vowel, and that while formant transitions might be useful in identifying the consonants, they are not of primary importance for recognizing the vowel itself. The traditional emphasis on the formant space defined by the first and second formants as the means of describing vowels

reflects this bias toward prototypical steady state formant values. A similar argument might be applied to the steady portion of a long fricative such as [s] or [ʃ] as compared to the onset or end of frication noise, or portions of frication noise showing effects of a neighboring vowel on the frequency of the noise.

The traditional belief that steady states are the most important perceptual cues is for the most part an assumption, not a tested hypothesis. Strange et al. (1983) document the traditional bias toward the steady state formant values as the main perceptual cue for vowel quality, and also review experimental literature which claims that vowels are perceived on the basis of the formants at one point in time (which could be interpreted as a claim for static cues). Furthermore, some non-experimental work proposes that perceptual cues can be perceived more easily if they continue through a longer part of the signal, that is, if their duration is increased (Flemming 1995), although increasing the duration of a cue by spreading it is likely to decrease the amount of change in the signal. Flemming (1995) claims that if the transitions into or out of a consonant extend through a large portion of a vowel, this will make it difficult for listeners to perceive the vowel. He states that "if overall syllable duration remains constant, any extension of formant transitions will be at the expense of shortening the interval during which the vowel alone is realized" (1995:55), and that "lengthening of consonant transition features is in conflict with maintaining vowel contrasts because the distinctiveness of a difference depends on its duration as well as its magnitude" (1995:56). This argument clearly depends on vowels being perceived from static spectral cues in their steady states, not from dynamic cues in their transitions.

### 1.2.2. The importance of changes for auditory processing

Contrary to linguists' traditional bias toward steady states, research on auditory perception suggests that changes in sounds should be of more use than steady states in perceiving speech. A large body of work on the responses of auditory nerve fibers (usually of cats) shows that auditory nerve fibers respond strongly when a sound near their

characteristic frequency begins, and that the response of the fiber drops off as the sound continues. This is called adaptation. (Delgutte 1980, Delgutte 1997, and Greenberg 1996a, 1997 provide useful reviews of this topic.) Each auditory nerve fiber has a characteristic frequency, the frequency to which it responds most strongly. Because the peripheral auditory system has auditory nerve fibers with many characteristic frequencies, as one hears a speech signal, the auditory nerve fibers with high characteristic frequencies have a peak response at the onset of speech sounds with high frequency energy (fricatives, bursts, affricates), followed by a drop off in response as the fibers adapt to the sound. At the onset of speech sounds with strong low frequency energy, such as vowels, the auditory nerve fibers with low characteristic frequencies show a peak response followed by a decrease in response as they adapt. Thus, different groups of auditory nerve fibers react strongly and then adapt as a speech signal progresses.

Adaptation is clear evidence that at the level of peripheral auditory processing, changes in sounds elicit a stronger response than steady states, since the auditory nerve fibers adapt during the first five to ten milliseconds from the onset of a signal (Greenberg 1997:1305-7), rather than continuing to respond strongly throughout a longer steady state sound. Although such data is obtained from auditory nerve fibers of cats, and not of humans, Delgutte (1997: 509) argues that the auditory systems of cats and some other animals are sufficiently similar to that of humans that such results are applicable to human speech processing.

Change is also important for visual perception, and this parallel with auditory perception suggests that dynamic cues, or change in the thing to be perceived, is important for perception in general. Some cells in the retina have a higher firing frequency when there is a sudden change in a visual stimulus (when something suddenly appears or disappears) and a lower firing frequency in between the onset and offset of the stimulus. This has some similarities with adaptation of the auditory nerve fibers. Furthermore, Yantis and Jonides (1984) and Jonides and Yantis (1988) show that attention is

automatically drawn to a visual stimulus which appears suddenly. They relate this to the fact that it is easier to notice something which moves than a part of a still scene. For example, one may not notice a rabbit or other animal which is standing still, but one sees it immediately if it begins running. Yantis and Jonides (1984) also mention the fact that lights on emergency vehicles are particularly salient because they flash, thus changing. (One might note also that most sirens and alarms use a sound which is modified over time, not continuous.) Thus, the hypothesized importance of change in auditory signals for speech perception has clear similarities to the importance of change in visual perception.

Delgutte (1997:508) following Stevens, points out that one of the defining characteristics of speech signals is that there is an alternation between high amplitude parts of the signal (vowels) and low amplitude parts (consonants), with this alternation taking place approximately three to four times per second (3-4 Hz). Delgutte (1997:513) argues that adaptation (the drop off in response of auditory nerve fibers as a sound continues) is important as a way for the peripheral auditory system to locate important points in the signal, such as the onsets of these high and low amplitude sounds, and is not simply an artifact of physical properties of the nerves.

### 1.2.3. Experimental studies of dynamic and steady information in speech perception

Research on responses of auditory nerve fibers, however, does not show what cues listeners use in perceiving speech. It only shows what characteristics of a signal might be likely to be useful perceptual cues because of a biologically determined sensitivity to them. Some past work in speech perception has addressed the question of whether listeners use dynamic or steady cues in perceiving speech, though.

#### 1.2.3.1. Perception of features of consonants

Several studies have investigated the importance of static and dynamic cues in the perception of place of articulation of initial stops (Stevens and Blumstein 1978, 1981,

Kewley-Port 1983, Kewley-Port et al. 1983, Strange et al. 1983). The primary question addressed by Stevens and Blumstein (1978) is what cues to place of articulation are invariant across a variety of vocalic environments. The question of whether there are perceptual cues which are invariant across environments or whether listeners must evaluate cues based on their environment is orthogonal to the question of whether the perceptual cues are dynamic or steady. However, the invariant cues Stevens and Blumstein (1978) identify are described as static, or steady.

In this experiment, Stevens and Blumstein (1978) synthesized a continuum between /b, d, g/ and another from /b/ to /g/, with /i, a, u/ as the possible following vowels. For each such continuum (for example, /ba/ to /da/ to ga/), they further generated one continuum which had only vocalic information (vowel formant transitions and steady states, but no initial burst), one continuum with an initial burst and the following vowel (with formant transitions), and one continuum with the initial burst and a following vowel, which however began immediately at the steady state values for its formants, without transitions. The bursts had spectra which are characteristic of bursts at that place of articulation. Stevens and Blumstein found that listeners were highly inconsistent in identifying the stimuli which had bursts but no formant transitions, which is not surprising, since these stimuli had conflicting cues for place in the burst and the (non-)transitions. However, listeners were rather consistent and showed typical categorical perception responses for both types of stimuli with formant transitions, those with and without bursts.

Stevens and Blumstein conclude that the invariant cue to place of articulation is the "gross shape of the spectrum sampled at the consonantal release" (1978: 1367), including both the burst and the beginning of voicing in the calculation of the spectrum. They emphasize the importance of this cue over the dynamically varying formant transitions, because it is relatively consistent regardless of vocalic context. They believe that the importance of the formant transitions is only to "join the onset spectrum to the vowel smoothly without introducing any additional discontinuities" (1978:1367), and that the

poor performance of listeners to the transitionless stimuli was the result of the discontinuity of those stimuli. They interpret the burst spectrum cue as one which does not vary over time ("...an *event*, ... rather than ... a sequence of events over time" (1978:1367)), and the bursts of their synthesized stimuli were static for the 5 ms duration of the burst. However, this is not a clear test of dynamic versus steady cues. First, their synthetic static bursts were not compared to dynamic bursts. Second, any event which lasts only 5 ms may be considered to be inherently dynamic. Third, since listeners identified the burst-less stimuli quite well, there is no way to show from this experimental design that the cue listeners used for those stimuli was the burst spectrum during one particular window near the onset (a static cue) rather than the movement of the formants during the transitions (a dynamic cue).

Stevens and Blumstein (1981) continue this work, by identifying spectral templates which they believe represent invariant<sup>3</sup> cues for place of articulation present near the release of a consonant. They compare a corpus of naturally spoken syllables to these templates, and find that the proposed invariant cues are present in a large proportion of the spoken syllables. They do not specifically address the issue of dynamic versus static cues here, but the proposed cues, characteristics of the short time spectrum at consonant release, are static. They do, however, mention the importance of rapid spectral changes as a characteristic of speech.

Stevens (1971), in a very interesting discussion of dynamic cues, divides the possible perceptual cues into three types: static, changes taking 50-100 ms, and changes taking 10-30 ms. He defines "acoustic transients" as having "abrupt changes in the short-time spectrum of the sound" (1971:96), and argues that these are important perceptual cues because of the response characteristics of the auditory system. That is, he emphasizes the importance of rapid changes in the signal, but only of rapid spectral changes. He specifically rules out rapid changes in amplitude only as comparable cues, although he

---

<sup>3</sup> This refers to invariance in different vocalic environments, not necessarily to lack of change over time.



states that the emphasis on abrupt spectral changes and not on abrupt amplitude changes is tentative. He presents data showing that the spectra near the beginning of a consonant-vowel transition demonstrate such rapid spectral changes, but only for [+consonantal] segments, not for [-consonantal] segments such as [h] or [ʔ].

In later work, Stevens (1985) argues again for the importance of brief portions of the signal with rapid change, although here he considers both rapid spectral and amplitude changes as important. He points out that these portions of the signal are often considered to be the boundaries between sounds, but says they are more than that: "a great deal of information concerning the phonetic features of segments is carried by various properties of the sound in the time interval that lies within 10 to 30 msec of these so-called boundaries. Rather than considering these regions as boundaries between segments, it may be appropriate to regard them as intervals that provide cues for a number of the phonetic features of the segments in an utterance" (1985:243). Stevens (1985) presents data on perception of several distinctions, such as the strident/non-strident distinction, the anterior/non-anterior distinction for fricatives, the nasal/stop distinction, etc., using synthesized CV syllables. For each distinction, he argues that the important cues are located within a window of approximately 10-30 ms surrounding the boundary between the consonant and vowel. For example, he proposes that the distinction between strident /s/ and non-strident /θ/ is perceived by comparing the amplitude of noise near the frequency of the vowel's fifth formant to the amplitude of that formant, and that one can best compare these at the boundary between frication and voicing. Similarly, the [±anterior] distinction is said to be perceived by comparing the amplitude of frication at the frequency of the vowel's F3 to the amplitude of F3, with this comparison done at the CV boundary.

However, this study does not provide a real test of the importance of these areas of rapid change, since amplitude of frication and of the vowel formants are static features in Stevens' synthesized syllables. Using the perceptual cues Stevens proposes, the only

reason for the area of rapid change to be important is that this is where the two things to be compared are close together. Stevens does, however, mention the sensitivity of the auditory system to rapid changes. Stevens' proposal regarding dynamic cues is most clear in this work: it is that listeners locate regions of rapid change in the signal and extract cues, which may themselves be static, from those regions. However, he does leave open the possibility that the cue might be the change in the spectrum during these regions (1985:253). Stevens' focus on regions of rapid change in the spectrum as areas of concentrated information is also apparent in his other work (Stevens 1980, Stevens and Blumstein 1978, 1981, Stevens and Keyser 1989, Stevens et al. 1986), but it is most clearly discussed in Stevens (1971, 1985).

Ohde and Ochs (1996) present data which bears on the issue of whether the part of the signal near a segment boundary simply contains static cues, or whether change during that part of the signal is a cue in itself. They excised 25 ms each of nasals and vowels, using the portion of the nasal or vowel which was immediately next to the boundary between the two in NV syllables. These short, excised signals correspond to the regions of rapid change at the boundary Stevens (1985) considers important. They presented these excised short stimuli to listeners, presenting in one condition the nasal or vowel portions separately, and in another the nasal portion followed by the vowel portion with a silent gap of various durations intervening between the two. They show that listeners can perceive the place of the nasal significantly better when presented with both the nasal and vowel stimuli than when they heard only one or the other, despite the intervening silent gap. However, perception was better for shorter gaps. Although the main purpose of their experiment was not to test dynamic versus static cues, they believe their results indicate that the change in the spectrum during the part of the signal surrounding the boundary between sounds is an important perceptual cue in itself, and that the boundary does not simply signal that there are static cues nearby.

Kewley-Port (1983) and Kewley-Port et al. (1983) argue against any static invariant cue and present evidence in favor of dynamically varying cues for place of articulation. Kewley-Port (1983) recorded productions of /b, d, g/ before each of several vowels from three speakers, and produced running spectra for the beginning of each CV syllable, calculated over a 20-25 ms window. The first spectrum was centered around the beginning of the burst, and eight spectra were calculated for each syllable, each one centered 5 ms later than the previous one. Thus, the series of 8 spectra represented the initial 40 ms of the syllable, and provided dynamically varying information about that 40 ms of the signal. Using a rather unusual methodology, Kewley-Port printed out the series of spectra and presented these visual representations to three phonetically trained judges. She defined three characteristics the judges were to use in evaluating the spectra, namely tilt of the spectrum at the burst, "late onset" (equivalent to a relatively long VOT, as /g/ would have compared to /b, d/), and presence of mid-frequency peaks. Judges were trained as to what combination of presence or absence of these features /b/, /d/, or /g/ should have. They were then asked to identify, based on the visually presented running spectra, what the onset consonant had been. This is quite unusual in that the experiment involved no listening—judges made their decisions based solely on the printed spectral representations. It represents an early strategy similar to that of modern connectionist modeling. Kewley-Port found that the judges were able to identify the consonant rather accurately (approximately 90% correct or more in most environments) based on these running spectra. She concludes that there are invariant cues to place of articulation, and that dynamically varying cues allow a more principled identification of place than static cues do. However, this unusual methodology is not conclusive, and she does not explicitly test dynamic cues against steady ones in this experiment.

Kewley-Port et al. (1983) present a more explicit test of dynamic versus static cues to place of articulation of initial /b, d, g/. In two experiments, they used stimuli consisting of the first 20, 30, or 40 ms of CV syllables, with /b, d, g/ as the consonants and /i, a, u/

as the vowels. (Not all combinations of those consonants and vowels were used, however). In one experiment, they tested three types of stimuli: real speech, synthetic speech with the spectral shape of the burst kept constant throughout the burst (and modeled on the spectral shape for real speech), and synthetic speech with spectral shape of the onset of the burst matched to real speech and then allowed to vary each 5 ms thereafter, with values interpolated to the beginning of voicing. Thus, the burst onset spectrum (Stevens and Blumstein's proposed static cue) was the same for both, and was matched to real speech, but remained the same throughout the burst for the static stimuli, and varied for the dynamic stimuli. Klatt synthesis was used for the synthetic stimuli. Except for a few cases which are explained as synthesis mistakes, subjects identified the dynamic stimuli more accurately than the static ones, and almost as well as the natural stimuli. The other experiment was very similar to the first, except that the synthetic stimuli were synthesized by LPC synthesis instead of Klatt synthesis. This time, the amplitude of the bursts was held constant for both stimulus types, and matched to natural speech. Again, subjects identified stimuli with changing spectral information more accurately than stimuli with static information, and also identified the dynamic stimuli better than in the Klatt synthesis experiment.

The authors conclude that one cannot divide CV perceptual cues into static cues in the burst and dynamic cues in the vowel transitions, but that the burst also contains dynamic cues. They also conclude that listeners make use of those dynamically varying cues, and do not treat the spectrum of the burst as static. While this study is rather convincing, one could argue that the dynamic stimuli were identified more accurately not because they contain changing spectra, but because they more closely mirror the real speech they were modeled on. The synthesis procedure involved altering the static stimuli more drastically from natural speech than the dynamic stimuli were altered. The synthesis could have introduced some artifacts into the static stimuli, causing the static stimuli to be identified less accurately even if listeners use only static burst cues. However, this study

provides strong evidence for the use of dynamic cues in the perception of place of articulation of initial stops, and provides a very direct test of dynamic versus static cues.

However, while Kewley-Port et al. (1983) argue explicitly that Stevens and Blumstein's hypothesis of static cues is incorrect, it is not clear that Stevens and Blumstein entirely support static rather than dynamic cues themselves. In Stevens and Blumstein (1981:4-6), they delineate three characteristics of speech signals which distinguish them from other sound signals (mentioned above with regard to Delgutte's work). These characteristics are first, that the amplitude of a speech signal rises and falls from consonants to vowels to consonants. In a general way, this is the phonetic correlate of syllable structure and the sonority sequencing principle. Second, there are changes in the spectrum of the signal which happen more rapidly than the overall amplitude fluctuations, for example formant transitions. Third, there are peaks and valleys in the spectra of speech sounds (along the frequency axis, not over time—these are the formants). (Tillmann (1980) similarly divides prosody into changes at the level of the utterance, the level of the syllable (lexical stress), and within segments.) Stevens and Blumstein believe that these characteristics of speech signals are related to the ability of the auditory system to perceive certain types of signals, that is, that speech is structured to exploit the abilities of the auditory system. They state: "...the auditory system may be predisposed to produce a variety of distinctive responses *when the properties of the sound are characterized by change or lack of constancy*: changes in spectral amplitude over frequency at a fixed time, changes in amplitude over time, and changes in spectral peaks and valleys over time" (1981:5, emphasis added). Of the three types of change they mention, the latter two are both changes over time, while the first is a "change" only along the dimension of frequency, not in time. While the cues Stevens and Blumstein suggest as invariant cues for place of articulation are described only as the static onset spectra, it certainly seems that Stevens and Blumstein consider time varied aspects of the signal to be important to perception. As discussed above, Stevens (1971, 1985) emphasizes the importance of

rapidly varying parts of the signal, but proposes that the cues located at these areas may be static themselves.

#### 1.2.3.2. Perception of vowels

Strange et al. (1983) test dynamic versus steady state cues, in addition to the cue of duration, for vowels. The authors recorded /bVb/ syllables with each of 10 English vowels, and removed various portions of the signal in order to selectively remove the regions hypothesized to contain steady and dynamic cues. They presented listeners with the entire natural speech syllable, the syllable with the steady state portion of the vowel replaced by silence (duration not altered), or the syllable with steady state of the vowel replaced by silence and the duration of the silence altered so that all syllables had the same duration as the shortest or the longest token. They also created stimuli from the same syllables with all but the steady state portion of the vowel removed (but the duration of the steady state unaltered), and the same with the steady state vowel duration altered to equal the duration of the shortest or longest token. Finally, they also created stimuli with only the initial stop and vocalic transition or only the final vocalic transition and stop. Thus, one can separate the contributions of the transitions, the steady state portion of the vowel, and duration as perceptual cues for vowel quality.

Results showed that listeners could identify the vowel almost as well from either the initial and final transitions or the steady state portion as from the entire intact syllable. Removing durational cues did decrease correct identification of the vowel in some cases. Listeners could not identify vowels well at all when presented with only the initial or only the final portion of the syllable. The authors conclude that vowel perception cannot depend solely on locating a steady state target, since listeners can identify vowels accurately even when the steady state portion has been completely removed. However, their data does not show that vowels are necessarily perceived on the basis of dynamic cues, since listeners identified vowels equally well from either the steady state portion or the transitions. It

shows only that listeners are able to use dynamic cues, and that the traditional view that transitions are irrelevant to vowel perception and only steady states are important is incorrect.

Fujisaki and Sekimoto (1975) also test changes in formant frequency as perceptual cues, but in V1V2, glide-vowel, and stop-vowel sequences. They find that listeners, hearing a sequence which is truncated during a formant transition, can extrapolate to the intended endpoint of the formant transition, making similar categorization decisions about the stimulus regardless of whether they heard as little as 70% or as much as 95% of the transition. They find that a very similar effect occurs with non-speech stimuli as well, and conclude that the extrapolation they found is the result of peripheral auditory processing, not of higher level knowledge about speech. Furthermore, they find that a preceding vowel, glide, or stop has different influences on the category boundary between /*u*/ and /*a*/, and conclude that there are phonological influences on the extrapolation effect as well as auditory influences. There are several problems with the design of this study, and it used a small number of highly experienced subjects (two of them the authors). However, whether the authors' conclusion about the extrapolation effect being based in auditory processing is correct or not, they do indicate that listeners make use of dynamic cues (formant movements) in categorizing vowel phonemes.

Nearey and Assmann (1986) test the importance of change in formant frequencies in the perception of vowels produced in isolation. In English, even most vowels which are usually considered monophthongal, such as /*i*, *e*, *æ*/, have some change in their formants over the course of the vowel. This is true for these vowels in isolation, and this change is in addition to the transitions imposed by neighboring consonants. Nearey and Assmann (1986:1297) refer to this as "inherent spectral change." In this study, Nearey and Assmann presented listeners with two small sections of each vowel, either the vowel nucleus and offglide in that order (the natural order), the offglide and then nucleus (reverse order), or the nucleus twice (representing a completely steady state vowel). They found that listeners

could identify vowels in the natural order condition approximately as well as they could identify naturally produced unmodified vowels, but that identification in the reverse order and completely steady state conditions was significantly worse. They conclude that the slight movement of formants during even monophthongal vowels provides important dynamic perceptual cues.

These experiments have laid a strong groundwork for the question of dynamic versus steady cues in speech perception, but they do not constitute a complete answer. They explicitly test dynamic and static cues for only two distinctions in only a few environments: place of articulation in initial, prevocalic position, and vowel quality in the /bVb/ environment or in isolation<sup>4</sup>. Furthermore, most of these studies have used only English speaking listeners and either English stimuli or synthetic stimuli modeled on English. They have also used monosyllabic nonsense syllables as stimuli for the most part. It is quite possible that the weighting of dynamic versus steady cues could be influenced by stress, that it could be different for consonants in VC position rather than CV position, that it could vary by language, and that it could be different for different manners of articulation. It is important that the initial tests of a hypothesis use stimuli which are as controlled as possible, as the stimuli in these experiments were, but it is also important to extend the tests of a hypothesis to more realistic environments in subsequent work.

#### 1.2.3.3. Perception of a variety of speech sounds

Furui's (1986) work on dynamic perception of segments by Japanese listeners tests the hypothesis of dynamic cues in a novel way, and simultaneously tests it in more diverse environments than the previous literature did. Furui defined a measure of degree of spectral transition, which he called D, based on the average slope of the cepstral coefficients over a time window. He also performed a perception test, in which he used all

---

<sup>4</sup> The studies discussed here which tested other environments or distinctions did not explicitly test dynamic against steady stimuli.



possible Japanese CV and CyV syllables (the set of monomoraic syllables in the language), and truncated (gated) them from either the beginning or the end of the syllable. He found that in initial gating (beginning removed), the vowel is correctly identified at all durations of truncation. However, consonant identification reaches a "critical point," which he defines as 80% correct, just before the point of maximal spectral transition,  $D_{\max}$  (when that point is still present). For final gating (end of syllable removed), consonant and vowel identification improve together as listeners are allowed to hear more of the syllable, but the critical point for both is just after the point of maximal transition.

Although he also presents other results in this article, the important point for dynamic cues is that listeners could identify an entire CV syllable correctly if they heard from the beginning of the syllable through the point of maximal spectral transition, or from before the point of maximal spectral transition through the end of the syllable. If the point of maximal spectral change was gated out of the stimulus, the listeners could not identify the syllable correctly. Furui shows that improvement in identification of the vowel (final gating) or consonant (initial gating) takes place very rapidly around the point of maximal spectral change. That is, identification of the vowel, for example, is not very accurate even when listeners are allowed to hear up until almost the point of maximal spectral change, but as soon as the stimulus is long enough to include that point, identification improves sharply.

Furui concludes that listeners use dynamic cues to perceive speech sounds, since they quickly become able to identify sounds when they hear the point of maximal spectral change. He found that the improvement in perception happened within approximately 10 ms after the point of maximal spectral change, which provides rather convincing evidence that accurate perception and spectral change are related. This is a very different approach to testing dynamic cues from that taken in the previous literature. The previous literature defined dynamic information relative to particular segments, considering as dynamic cues the formant transitions of a vowel or as the time varying signal during a burst, for example.

Because Furui defines a signal processing based measure of degree of spectral transition, his work is not limited to a few segment types, but rather offers an objective way to determine the point of maximal spectral change for any speech signal. However, as will be discussed in Chapter 3, there are problems with the measure D.

It is important to consider that the point of maximal spectral change identified using Furui's measure D may not correspond to what would traditionally be thought of as a transition. For example, much of the previous literature on dynamic cues in stop-vowel sequences has considered the formant transitions in the early part of the vowel to be the dynamic cue. The measure D, however, would identify the onset of voicing, not the period during which formants move toward their steady state value, as the point of maximal spectral change. This is because the change from aspiration noise to voicing is a more sudden change than the formant transitions. The issue of dynamic cues as slower changes within a signal versus dynamic cues as the onset of a sound will be addressed more thoroughly in Section 5.1.2. Still, the measure D is valuable as an objective measure of degree of spectral change which can be applied to any speech signal.

Furui's work extends the previous investigations of dynamic cues greatly in that he tests all of the possible Japanese monomoraic syllables, so he is not restricted to a single contrast (such as place of articulation in a CV or vowel quality). The fact that he offers an objective measure of degree of spectral change, which can be applied to any signal, is very important for this extension of the topic. Furthermore, his use of Japanese adds a cross-linguistic aspect to the literature on dynamic cues. However, his study is still quite limited: because he uses only monomoraic Japanese syllables, his results are limited to CV and CyV transitions, and include no VC, VV, or CC transitions. (Although Japanese has very few allowable VC syllables, it is quite possible to test such transitions, since there is no reason why the transitions to be investigated must be within a syllable.) Also, Furui used the entire set of Japanese monomoraic syllables, which corresponds to the set of syllabary symbols, since Japanese orthography is really based on the mora, not the syllable. The

subjects, who were asked to respond with the syllabary symbol they thought they heard, knew the set of possible answers, since all educated native speakers of Japanese know the set of syllabary symbols. This means that this experiment was similar in the responses to the closed class forced response, nonsense word stimulus methods of the previous literature. Such methods investigate only perception of individual segments in a very controlled situation, not perception of actual speech, which requires an open class response (discussed further in section 1.2 below). Also, Furui's experiment still used only stimuli which constituted the only syllable of the nonsense word, so his work does not address possible effects of position in the word (first versus second syllable for example), location of pitch accent, etc. Finally, Furui provided the subjects for his experiment, of whom there were only four, with extensive training on the stimuli before the data was collected. His results therefore represent an unrealistically careful listening situation. However, his experiment is an important step in advancing the study of dynamic cues.

#### 1.2.4. The importance of change in the signal seen through phonological universals

Kawasaki (1982) and Ohala and Kawasaki-Fukumori (1996) argue that changing aspects of the signal are important through an entirely different method: comparison to phonological universals. Kawasaki (1982) identifies many sequences of segments which are impossible in a large number of unrelated languages. For example, labials are not followed by labial glides in many languages (\*pw, \*bw, \*mw in English except for recent borrowings, as well as in many other languages), alveolars often cannot be followed by a lateral (\*tl), etc. Kawasaki recorded productions of a variety of consonant-glide-vowel, consonant-vowel, and vowel-consonant sequences, calculated the formant trajectories of these sequences over time, and defined a measure of the degree of change in formants during these sequences. She showed that many of the sequences which are cross-linguistically dispreferred are those with the least change over time in their formants. She concludes that acoustic modification of a signal over time is important for perception, and

that languages are therefore more likely to make use of segment sequences which involve great acoustic change. Although this study does not explicitly test listeners' perception of any sound, it adds a different type of evidence to the argument that dynamic cues are important in speech perception. Lindblom (1984) also discusses the need for acoustic change between consecutive segments, and argues that this could be a central principle in viewing phonemes and syllables as a self-organizing system.

Bladon (1987) examines several phonological or phonetic universals and searches for plausible explanations for them in the psychophysical properties of the auditory system. He emphasizes the importance of onsets of acoustic energy rather than offsets, which is a result of adaptation: since the auditory nerve fibers respond strongly when a sound begins, but have adapted by the time it ends, one sees strong reactions only at the onset of energy (at a particular frequency), not at the cessation of energy. Bladon argues that this fact can explain the cross-linguistic tendency for nasalization of vowels to be stronger before a nasal than after one, the tendency for postvocalic but not pre-vocalic laterals to become glides (as in English), and the cross-linguistic rarity of preaspiration and postvocalic [h]. A vowel-nasal or vowel-lateral transition mostly involves a reduction in energy, with only a few frequencies increasing in energy, whereas a nasal-vowel or lateral-vowel transition involves an increase in energy. Therefore, he claims, one should find little reaction of the auditory nerve fibers at the onset of a postvocalic nasal or lateral, since they have already adapted, whereas one should find a strong auditory nerve response at the onset of a vowel after a nasal or lateral. Similarly, preaspiration (after a vowel) and postvocalic [h] involve primarily a reduction in amplitude (energy) with little change in the frequency of energy. Although Bladon's (1987) proposals are based only on plausibility and not on any experiment designed specifically to test his hypothesis, they provide another interesting argument for the importance of changes in the signal, and suggest that the influence of auditory processing can be seen in phonological universals.

### 1.3. Perception of segments and recognition of words

There are several well known models of spoken word recognition, and the literature on various aspects of how listeners recognize spoken words based on phonetic information they extract from the signal is voluminous. Reviewing the literature on models of spoken word recognition would require a book in itself, and there are several books on this or related topics available (Frauenfelder and Tyler 1987, Altmann 1990, Marslen-Wilson 1989, among others). Therefore, I will give here only a brief discussion of the relation of spoken word recognition to perception of segments, restricted to topics which are important for the experiment presented in subsequent chapters.

Early models of spoken word recognition are well represented by early versions of the cohort model (introduced by Marslen-Wilson and Welsh (1978)). In this model, listeners, when they have recognized approximately the first phoneme of an utterance, form a cohort consisting of all the words in the lexicon which begin with that phoneme. Then, as each successive phoneme is recognized, words which do not match the segmental content of the recognized signal are eliminated from the cohort. This process continues until the word is recognized, ideally because only one word remains in the cohort. Using an example from Shillcock, when a listener hears /mar/, the cohort includes words such as "mar, mark, market, marlin, marquis<sup>5</sup>" and others (1990:26). However, once the listener has heard /markət/, the cohort is narrowed to "market" and words derived from it by suffixation, such as "marketing, markets" etc. Some words are uniquely identified (become the only member of their cohorts) before reaching the final segment of the word.

This is a useful way to think about what possibilities a listener has available when processing a speech signal, but there are several serious problems with this simple model. The most serious is that it presupposes that listeners know where the beginning of the word is. If all words could be identified as unique before reaching the end of the word, this

---

<sup>5</sup> "Marquette" would also be in the cohort if stress is not considered at this stage of word recognition.

would be possible, but several researchers have shown that this is not the case (Shillcock 1990 presents a review). Listeners must often locate word boundaries after the fact, as the end of one word often cannot be recognized until a point in the following word.

Furthermore, the early version of the cohort model assumes that listeners are always able to correctly recognize each phoneme of a signal, and can only recognize words by comparing the signal to the initial portions of lexical items. If a segment early in the word is misperceived, this would imply that the listener forms a completely incorrect cohort. For example, if a listener misperceived the /m/ in "market" as /n/, this would lead to the formation of the cohort "narcissus, narcotics" etc. (at recognition of the /r/) instead of the cohort given above for "market."

Subsequent modifications to the cohort model (Marslen-Wilson 1987) have at least partially addressed these problems, so that locating word boundaries is now suggested to come about through the correct recognition of the word, rather than being a prerequisite to correct recognition. More importantly for the experiment in this dissertation, in the newer cohort model, words can be recognized based on similarities with the signal, even if not all of the segments of the beginning of the word are identical.

Several other possibilities for the group of words listeners choose among during spoken word recognition have been proposed. The TRACE model (McClelland and Elman 1986) has much in common with the cohort model, but it allows features perceived later in a signal to activate words even if the initial segments are not the same, solving the early misperception problem. The neighborhood activation model (represented by Luce et al. 1990) involves choosing among the words in a lexical similarity neighborhood, which is calculated based on phonetic similarity of the signal to lexical items, the frequency of the word which is eventually recognized, and the frequency of all the words in the lexical similarity neighborhood. This method allows for recognition of words even if all segments are not perceived correctly, and accounts for effects of word frequency on recognition (faster recognition of high frequency words) much more directly than the cohort model can.

Extensive experimental evidence supports this hypothesis. Amano (1997) suggests that spoken words are recognized by choosing from a group of words which share the rhyme of the stimulus syllable (at least for monosyllabic English words). He claims that this pool of candidate words accounts for perceptual data more accurately than either the cohort or the lexical neighborhood does. However, it is difficult to see how this would apply to polysyllabic words, or to languages with few codas, such as Japanese: the rhyme-based cohort for the word /ka/ 'mosquito' would include all words with the vowel /a/, surely not the most efficient way to exclude lexical possibilities.

Another problem with early models of spoken word recognition is that some of them assumed perception, and the resultant narrowing of the cohort, happened one phoneme at a time. Extensive research on speech perception has shown that most phonetic cues are used as they become available, even though cues for one segment often occur during other segments, distributed throughout the signal (Warren and Marslen-Wilson 1987, Lahiri and Marslen-Wilson 1991, 1992, Repp 1980, several studies cited in Cutler and Otake, to appear). It is clear that a detailed model of spoken word recognition must allow for the integration of cues which are distributed in time, and for a narrowing of the pool of words from which one is recognized based on perception of some distinctive features of a subsequent segment before the entire phoneme is recognized. For example, such a model should allow for a postvocalic nasal to be perceived as nasal based on nasalization during the vowel, which would restrict the possible choices for the word to be recognized, before the place of articulation of that nasal is recognized (the example discussed in Lahiri and Marslen-Wilson 1991, 1992 and Ohala and Ohala 1995).

There has been much interesting research on many other issues of spoken word recognition. However, the topics which will be relevant for the experiment reported in this dissertation are those discussed here: the concept of the cohort, recognition of words despite mistakes in segment perception, the effect of word frequency on spoken word recognition, and the continuous use of perceptual cues which are distributed in time. For

further information on these and other issues of spoken word recognition, see Altmann (1990). One further point to keep in mind with regard to spoken word recognition is that one can never know exactly what words a particular listener will consider as candidate words when hearing a particular sequence of sounds. This is because individual listeners have different lexicons, and one can never know exactly which words a particular listener has in her active or passive vocabulary. Both the existence of some words in a listener's lexicon and the frequency or familiarity of a word depend to some extent on the individual listener's education, interests, hobbies, etc. This problem is unavoidable in most studies of spoken word recognition, however. One can choose words to use as stimuli which are common enough that all listeners are likely to be familiar with them, but one cannot know what additional words may compete with the stimulus word when a particular listener hears a particular stimulus.

#### 1.4. The use of gating experiments

##### 1.4.1. Overview of the gating method

Gating is a method which is used moderately often in studies of speech perception. t' Hart and Cohen (1964) provide an early description of gating, and Grosjean (1980) further developed the method. In a gating experiment, one presents listeners with one portion of a signal (a part of the duration, not of the spectrum) and asks them to respond, usually with the whole word they think the stimulus might have been a part of. This very simple manipulation, presenting only part of a signal, is intended to elucidate which parts of the signal contain the cues necessary for perception of some feature. Usually, a natural speech signal is gated, with no further manipulations to it. By gating, cutting the signal off at a certain point, one removes all potential cues which fall in the portion of the signal which is gated out. This truncation is the manipulation to the perceptual cues, so there is usually no need to manipulate the signal in other ways, such as by altering some parameter and resynthesizing, by filtering the signal, etc. (Warren and Marslen-Wilson (1987) and



Takehi et al. (1996), however, gate syllables which they created by splicing, thus adding an additional manipulation in the creation of the stimuli.)

In a gating experiment, a series of stimuli is generated from the same natural speech token. In final gating, which is the most common variety, the end of the signal is gated out, or end truncated, by ever larger amounts at a set interval. For example, the shortest stimulus might include only the first 50 ms of the word, the next longest stimulus would include the first 100 ms of the word, the next the first 150 ms, etc., with the longest stimulus consisting of the entire word. In initial gating, the same process is used, but ever larger amounts are removed from the beginning of the signal instead of the end. One can manipulate the gating interval (the number of milliseconds separating successive gates), as well as the boundaries in time of the area through which one wishes to gate.

#### 1.4.2. Topics investigated through gating

Gating has been used to examine the timing of perceptual cues for a variety of distinctions. Lahiri and Marslen-Wilson (1991, 1992) use final gating of Bengali and English CVC, CVN, and (for Bengali only) CVC̃ syllables to investigate the perception of nasal and oral consonants based on nasalization of the preceding vowel. Ohala and Ohala (1995) replicate their experiment with Hindi and English. By gating out successively longer portions of the end of the syllable, they can determine how much of the vowel and/or consonant listeners need to hear in order to recognize a following segment as nasal or oral. Lang and Ohala (1996) gate English CV syllables to explore patterns of vowel quality neutralization based on duration. Öhman (1966) investigates perception of place and manner of articulation and voicing in Swedish CV and VC syllables through initial and final gating. Furui's (1986) work using initial and final gating to analyze the perception of most segments of Japanese has already been discussed in section 1.2.3.3. Roengpitya (in press) employs initial gating to determine the timing of cues to the voicing contrast in Thai CV sequences. Efremova et al. (1963) use final gating to study the timing of cues for

perception of the pitch accent contrast in Norwegian, in an unusual but interesting application of gating to suprasegmentals. Cutler and Otake (to appear) use gating to determine whether Japanese listeners take pitch accent location into account during spoken word recognition, and where in the signal perceptual cues to pitch accent become available. The experiment by Kewley-Port et al. (1983) discussed above is also, in a way, a gating experiment, since they present listeners with the first 20, 30, or 40 ms of signals.

In all of these investigations, by presenting listeners with various portions of a signal, one can determine which part of the signal contains the necessary perceptual cues for a particular segment or contrast. If listeners hearing a particular gate give some responses with a voiced consonant and some with a voiceless consonant, for example, but at the next longer gate give only responses with a voiced consonant, then one can conclude that the part of the signal which is present in the longer gate but not in the shorter one contains some useful perceptual cue for the voicing of the consonant.

Gating has also been used to investigate spoken word recognition and the contributions of top down and bottom up processing. Many studies on Marslen-Wilson's cohort model (discussed in the previous section) make use of the gating method. Grosjean (1980) uses final gating of nouns, which could appear either in isolation, in a short sentence context, or in a long sentence context. Examples are "bog" for isolation, "He walked into a bog" for short context, and "Lost in the Scottish Highlands, he walked into a bog" for long context. In each case, only the final noun was gated through, not the entire sentence. Grosjean was able to identify several distinct sources of errors in recognition of the final word, some based on lack of phonetic information, some based on lexical status of a part of the target word (the response "cap" for "captain"), and some based on semantic interpretation. Tyler (1984) presents a similar gating experiment for Dutch, which she uses to compare the timing of top down and bottom up processing in spoken word recognition. She also investigates more thoroughly the tendency for listeners to respond with high

frequency words, showing that low frequency words are not completely excluded as responses even when there is also a possible high frequency response.

#### 1.4.3. A problem: Gating as altering rather than removing cues

Such experiments as these assume that the pool of responses given by all of the listeners to a particular stimulus reflects the cohort (see section 1.3 above) from which an individual listener recognizes a word. That is, in spoken word recognition, listeners are not choosing from all words of their lexicon, but rather from a cohort of words which have some information in common with the signal the listener has heard. The cohort is assumed to change over time as the listener hears more phonetic cues for the segments of the word. These experiments are an attempt to identify the cohort of words from which listeners are choosing at a given point in time, the point at which the signal is gated out. There may be some problems with this. The goal in gating is to remove some perceptual cues by gating out part of the signal. However, in some cases, gating actually alters some of the remaining cues instead of just removing others.

This is the case when the endpoint of the gate falls during a segment whose duration is a cue. If one gates /a/ out at a relatively early point in the vowel, English listeners will identify it as /ʌ/ instead of /a/. This probably does not mean that listeners perceiving connected speech, when hearing an /a/, initially parse it as /ʌ/ and then change their perception to /a/ as the vowel continues. Rather, the duration of the vowel is a cue to its identity, so gating it out alters that cue instead of just removing subsequent cues. Similarly, when a signal is gated out relatively early in a vowel, listeners perceive the word as having a voiceless obstruent following the vowel, because voiceless obstruents cause shortening of their preceding vowels. When the endpoint of the gate falls late in a relatively long vowel, however, listeners perceive the word as having a postvocalic voiced obstruent. (These results are discussed in Section 5.7.1 below.) This probably does not mean that

listeners hearing connected speech decide early in each vowel that there will be a following voiceless obstruent, reject words without postvocalic voiceless obstruents from the cohort, and then have to change this percept as the vowel goes on. Rather, gating the signal out during a segment which offers duration as a cue for some distinction alters listeners' perceptions in a way that does not represent the processes of perception of natural speech. Thus, one must be careful in interpreting the results of gating experiments, and must consider the specific types of cues which are removed or altered by gating. Gating does show where in the signal adequate cues become available for listeners to recognize particular individual segments, even when durational cues are altered. However, it may not always show what cohort of words listeners are considering as candidates during online spoken word recognition, particularly when durational cues are involved. The question of what cohort of words listeners are considering at a given point in the signal while perceiving speech is separate from the question of when phonetic cues to a segment become available, although gating is frequently used to investigate both these questions.

#### 1.4.4. Open versus closed class responses for gating

There are several methodological issues one must address in designing a gating experiment. One which has been alluded to above is whether to ask listeners for an open class or a closed class response. In a closed class design, such as those reported by Kewley-Port et al. (1983), Lang and Ohala (1996), or Roengpitya (in press), listeners are only allowed to choose between the possibilities the experimenter is testing, with possible responses such as /ba, da, ga/ (Kewley-Port et al.) or a list of the vowels of English (Lang and Ohala). The experimenter usually gives the listener the choice of all possible values for the distinction the experimenter intends to investigate, but if the listener believes the stimulus sounded like something other than the possible responses, this information will be lost. In an open class design, such as those by Lahiri and Marslen-Wilson (1991, 1992) or

Grosjean (1980), listeners are asked what whole word they might have heard the beginning of (or end of in initial gating).

In order to investigate spoken word recognition, the open class response method is essential. One cannot find out what words listeners succeed in accessing from the lexicon if one specifies the possibilities. However, open response data tends to be rather noisy, and thus difficult to interpret. If 5-10% of listeners give a particular type of response, or if a particular word is given as a response at one gate but not at another, is that meaningful, or is it chance? Ohala and Ohala (1995) argue against the open response method, at least for gating experiments such as Lahiri and Marslen-Wilson's, which are intended to evaluate listeners' ability to use certain phonetic cues at various points of the signal. They argue in part that it is quite difficult to perform statistical analysis on open response gating data. However, if spoken word recognition is a topic of investigation, open response experiments are required. It is possible to perform statistical analysis of such results with sufficient care for how the statistical tests are set up. One can also interpret the importance of minority responses by considering change over time: if early gates have 5-10% of a certain type of response, but later gates have 0% of that type, this result is meaningful. Ideally, open class response gating experiments should use very large numbers of subjects in order to reduce the noisiness of the data.

#### 1.4.5. Individual versus successive presentation

One of the most important issues in designing a gating experiment is the question of whether to present the same listeners with all gates of a word (successive presentation) or whether to have different groups of listeners hear each gate (individual presentation) so that an individual listener never hears more than one gate of the same word. In successive presentation, gating stages are usually presented successively from the shortest (most severely gated) to the longest (entire token, with nothing removed) to all listeners. Since an entire experiment involves gates of some number of different words, the stimuli are

presented in blocks by gate. This means that all subjects hear first a small portion of each stimulus word, then in the next block, a slightly longer portion of each of the same words, then in the next block, a yet longer portion of each of the same words, and so on until they hear the entire word for each of these same words. Many experimenters prefer to use this method rather than individual presentation because individual presentation greatly increases the number of subjects required (the number of subjects one would like to use multiplied by the number of gates).

It is easy to imagine that the experience of having heard previous gating stages of the same words could influence subjects' responses to subsequent gating stages. Ohala and Ohala (1995), based on data from Frederiksen (1967), discuss the possibility of a "hysteresis" effect, in which listeners who form an incorrect hypothesis about the identity of a word at an earlier stage of gating might maintain that hypothesis at later stages despite conflicting new information. In particular, when an open response method is used or spoken word recognition is being investigated, successive presentation is highly questionable, because open response experiments require listeners to perform lexical access. Once a listener has accessed a particular lexical item, it will be easier for the listener to give that response to subsequent stimuli than to access a different word. Alternatively, listeners who have had several opportunities to listen to the early part of a word might be able to make better use of ambiguous or secondary cues from early parts of the word when hearing a later gating stage of the same word than they could if they only heard the word once. If this is the case, they would give more accurate responses in successive presentation than in individual presentation. If either of these possibilities is true, it would skew the results from a successive presentation gating experiment in undesired ways.

Cotton and Grosjean (1984) test the successive and individual presentation formats against each other by replicating parts of the successive presentation experiment in Grosjean (1980) with the individual presentation method instead, and comparing the results of the two experiments. They claim to find no difference in results except in the subjects'

confidence in their answers at short gates. That is, the responses given by subjects in the two presentation style conditions were similar, but subjects who heard the individual presentation style were more sure of their answers for short gates. Many other researchers cite this experiment as evidence that it is safe to use the successive presentation method in gating experiments, but I believe Cotton and Grosjean's results are inconclusive and not thoroughly analyzed. At the longest gate duration, more subjects give the correct response (the word used for the stimulus) in the successive presentation format than in the individual presentation format. Perhaps subjects in the successive presentation experiment were able to use the repetition of earlier information to resolve ambiguities better. There was no significant difference between individual and successive presentation results when they were tested across all gates. Since the authors did not do a statistical analysis of data at only the last gate duration, we cannot tell whether this effect at the last gate is significant, but it appears in their graphs both for words in isolation and in short sentence contexts, both of the conditions tested.

The fact that such a discrepancy in the individual and successive presentation data appears only at one gate is perhaps support for the use of successive presentation. Given the obvious ways in which hearing parts of the same stimulus repeatedly could influence listeners, it is perhaps surprising that there is not more difference between the individual and successive presentation results. However, I do not believe that these results are sufficiently clear to justify the use of successive presentation in most gating experiments, especially open class response ones involving spoken word recognition (even though the experiment by Cotton and Grosjean is such an experiment). Furthermore, even if there was no real difference between individual and successive presentation results for Cotton and Grosjean's experiment, this lack of a difference might not extend to gating tests of all distinctions or to differently designed gating experiments. Most gating studies using successive presentation cite the study by Cotton and Grosjean (1984) as justification for using that method, but even if there were no doubt about Cotton and Grosjean's results,

their test would not be sufficiently general to assume that the results would be the same for all gating tests. This is unfortunate, since individual presentation requires many times more subjects than successive presentation.

Individual presentation data has some other disadvantages besides needing extremely large numbers of subjects. Especially with an open class response, it is likely to be somewhat more noisy than successive presentation data, because using different listeners for each gate increases the subject error component. (In general, between subjects designs have lower power than repeated measures, or within subjects, designs, because of the larger subject error component (Keppel 1991).) In spoken word recognition, different listeners have slightly differing lexicons and differing frequencies for words in their lexicons because of differences in their hobbies, education, etc., so individual presentation will result in some different responses at each gate based on the individuals' particular lexicons. Including more subjects in each gate is the obvious way to solve this problem, but since the individual presentation method already requires so many subjects, this may be impractical.

Thus, for now, the best way to approach a gating experiment with an open class response method and an interest in spoken word recognition is to use the individual presentation method, use as many subjects in each gate as possible, and develop additional strategies for analyzing noisy data. If one is using a closed response method, especially with nonsense word stimuli and responses, it may be safe to use the successive presentation method. Ohala and Ohala (1995) argue against successive presentation even with closed class responses (which were real words, however), but Lang and Ohala (1996) use successive presentation for a closed class nonsense word experiment. Lang and Ohala present the various gating stages of stimuli in random order, so that listeners do not hear the stimuli in order from the shortest gate to the longest. The dangers of successive presentation are certainly much less in such a case than for open response spoken word recognition studies.



#### 1.4.6. Gating as a reflection of online processing

A further methodological question for gating experiments is whether they reflect online processing of the signal, even though listeners are usually asked to write down their answers and given several seconds to do so. Because listeners have a large amount of time relative to the duration of the signal to decide on and write their answers, responses might reflect "postperceptual metalinguistic processes" (Tyler and Wessels 1985:218) rather than the online perception of the signal. Tyler and Wessels (1985) replicate the experiment in Tyler (1984), in which listeners were given 8 seconds to write down their responses, but in this version, they require listeners to say their responses out loud as quickly as possible, making this a naming version of the gating paradigm. Tyler and Wessels find that the reaction times in this test are similar to the reaction times found in other timed experimental methods which are assumed to be online tasks. They also find that responses (by all the measures commonly used in gating experiments, which will be discussed further below), as well as effects of various context conditions, are similar in the timed and the usual untimed gating tasks. From the similar responses and fast reaction times, they conclude that the gating task, whether timed or not, does represent online processes.

While there are some slightly questionable points in the interpretation of Tyler and Wessels' results, these results seem more clear than Cotton and Grosjean's results in support of successive presentation. In addition to being logistically inconvenient, using spoken responses introduces an additional source of error, since subjects' responses have to be recorded on tape and analyzed by the experimenter afterwards. If the experimenter is ever unsure, for example, of whether the subject's spoken response began with a labial or a velar, there is no way to check. Written responses introduce some experimental error too, as discussed in Chapter 4, but are probably preferable. There are difficulties resulting from listeners' attempts to interpret incomplete signals (such as durational cues which are altered by gating, as discussed above), but these would probably be equally present in a naming

version of a gating experiment. Careful consideration in interpretation of results, rather than a change to the experimental method, is in order for such cases.

#### 1.4.7. Ways of analyzing gating data

Results for gating experiments can be analyzed in several ways, yielding a wealth of information. Most gating experiments which focus on timing of perception of particular segments analyze only the percent of listeners giving responses with that segment correct, or the percent giving responses with particular features. Lahiri and Marslen-Wilson (1991, 1992), for example, analyze the percent of listeners at each gate giving responses with a final nasal consonant or a final oral consonant, or for Bengali, with a phonemically nasalized vowel or a non-nasalized vowel. This is clearly the most important information for evaluating listeners' ability to perceive a particular distinction. Some researchers choose some arbitrary percent correct (for example 80% for Furui (1986)), and define the point at which responses first exceed that level as a recognition point.

Most gating experiments which investigate spoken word recognition or bottom up and top down processing, however, use a variety of other measures and usually do not calculate the percent of responses with a particular feature or segment correct. Grosjean (1980), for example, locates an isolation point and a recognition point for each word. The isolation point is the point at which a listener gives the correct whole word response (the same word as the stimulus) and then does not give any wrong answer at subsequent gates. The recognition point is based not on the actual response, but on listeners' judgments about how certain they are of the response. Furthermore, he calculates the number of different responses given to each stimulus, and plots for each word the number of subjects who give each response at each gate. This is a useful way to see overall patterns in the data, but is difficult to analyze numerically. Asking subjects to state their degree of certainty about their responses, and evaluating this as well as the responses themselves, is relatively common in such studies (Cotton and Grosjean 1984, Tyler and Wessels 1985, for

example). The choice of measures, however, clearly depends on the type of information the experimenter hopes to gain from the study. Furthermore, in order to use the "isolation point" measure, one must gate through the entire word: if listeners only hear a part of the word even at the longest gates, they may never agree on the one "correct" response.

#### 1.4.8. The gating interval

A final issue in the design of a gating experiment is the interval at which to gate the stimuli, that is, how many milliseconds the endpoints of successive gates should be separated by. Most work on spoken word recognition and bottom up versus top down processing uses a rather large gating interval, as long as 50 ms (Salasoo and Pisoni 1985) or entire diphones (Cutler and Otake, to appear, defining the gating interval relative to acoustic landmarks rather than as a set interval of time). Even work on timing of perceptual cues for a particular segment sometimes uses long gating intervals, such as 40 ms in Lahiri and Marslen-Wilson (1992). Grosjean (1980) uses 30 ms gating intervals. Among shorter gating intervals are 20 ms (Efremova et al. 1963, Öhman 1966), and Furui (1986) uses the shortest gating interval I have seen, 10 ms.

The purpose of the experiment should dictate the gating interval: an interval of 50 ms can only give information about how listeners recognize words as the amount of phonetic information available increases by a phoneme or a large part of a phoneme with each gate. To test use of dynamic versus static cues, as Furui does, one needs an extremely short gating duration. In order to test how perceptual cues become available over time, the gating interval should be shorter than any proposed linguistically significant unit. Thus, the gating interval should be shorter than the aspiration period of a stop, and no longer than the burst of a stop, so that the contributions of the burst and the aspiration noise to perception of the voicing or aspiration distinction can be evaluated separately. If both burst and aspiration fall within the same gate, responses to the burst information will also include aspiration information. As flaps, bursts, and some glides are quite short, a gating

interval of no more than 10 ms is ideal. However, if one uses the individual presentation method, a gating interval of 20 ms would halve the number of subjects required, so some compromises on gating interval may be necessary.

#### 1.5. Language specific differences in perception and spoken word recognition

Research in phonetics and psycholinguistics using a variety of paradigms has shown that the methods speakers of different languages use in perceiving speech and recognizing words differ at several levels. Japanese and English, in particular, show many differences. At the level of perception of individual segments, Fujimura et al. (1978) and Kakehi et al. (1996) show that Japanese and English (Fujimura) or Japanese and Dutch (Kakehi) listeners weight the cues for place of articulation in a VCV sequence differently. Fujimura et al. (1978) spliced Japanese VC1V and VC2V stimuli together, where C1 and C2 are voiced stops with different places of articulation. They created VC1C2V stimuli with the duration of the closure too short for listeners to perceive the stimuli as having more than one consonant. They also manipulated accent placement, using both high-low and low-high stimuli. They presented these stimuli to both Japanese and English listeners. One should note that English listeners will perceive a Japanese word with first mora pitch accent, the high-low pattern, as having initial stress.

The results show that both Japanese and English listeners tend to perceive the VC1C2V stimuli as VC2V. That is, listeners of both languages weight the cues in the CV transition more strongly than those in the VC transition. However, English listeners make more use of the VC transition (responding VC1V somewhat more often) when they perceive the first vowel as stressed, whereas Japanese listeners' responses are the same regardless of pitch accent placement. Japanese listeners do not make differential use of transitions depending on pitch accent patterns. The authors conclude that experience with one's native language affects strategies used for perception, because final C's in Japanese

syllables are never specified for place, whereas English requires the use of VC transitions for its many closed syllables.

Takehi et al. (1996) performed a similar splicing experiment for Japanese and Dutch, but further manipulated the stimuli by combining splicing with gating. After splicing together VC1C2V stimuli as in Fujimura et al. (1978), they gated out portions of the signal beginning with the burst of the C2 and continuing part way through the second vowel. This simply removes the CV cues to varying degrees, instead of using stress to manipulate the weighting of available cues, as Fujimura et al. did. They asked listeners to identify both consonants (C1 and C2) in this experiment, which is a surprising methodological choice, as previous research such as that by Fujimura et al. shows that listeners perceive such spliced signals as having only one consonant. Indeed, both Japanese and Dutch listeners found it very difficult to identify two consonants in these stimuli. However, Takehi et al. did find that Dutch listeners were more accurate in identifying C1 than Japanese listeners were.

Takehi et al. performed several variations on this experiment, many using gating without splicing. For example, they gated out the burst and varying portions of the second vowel in natural VCV (unspliced), and CV stimuli, and also presented VC only stimuli, which were final gated. They found that Japanese and Dutch listeners were equally accurate in perceiving the C of a VCV stimulus, even when all CV cues had been removed by gating. Japanese and Dutch listeners also performed equally on gated CV stimuli, although neither group was very accurate, as few cues to place of articulation were left. However, Dutch listeners were more able than Japanese to perceive the place of the C in gated or intact VC stimuli. The results for the VCV stimuli show that Japanese listeners are not entirely unable to use VC cues, but the results for the VC stimuli and the spliced VC1C2V stimuli may confirm Fujimura's finding (comparing to English) that Japanese listeners are less likely to use VC cues than listeners of languages with more possibilities for coda consonants. However, both VC and VC1C2V sequences are impossible in

Japanese if the C is a stop, and listeners for the spliced condition are told to identify two consonants. These syllable types presumably are possible in Dutch, though, which clearly biases the results. As a supplement to Fujimura's work, this article does strengthen the result that Japanese listeners are even less likely to use VC cues than are speakers of languages with more coda contrasts, but it is not conclusive in itself because of these methodological issues.

Takehi et al. (1996) mention the restrictions on coda consonants in Japanese, but go even further to conclude that Japanese listeners parse the speech signal by units larger than the phoneme (probably moras), since they only succeeded in using VC cues in natural VCV stimuli, where the consonant is the onset of a CV mora. This argument is quite weak, especially given the confound of phonotactically inadmissible stimulus types. Other experimental paradigms, discussed below, do indicate that Japanese listeners use the mora in speech processing, but this experiment cannot test that hypothesis.

At a different level, extensive work by Cutler, Mehler, Norris, Segui, McQueen, Otake, and their colleagues has shown considerable differences in what suprasegmental units listeners use in perceiving speech, and in how listeners locate word boundaries. The paradigm which has yielded the most thorough results on this subject is a version of target spotting, in which listeners are given a target and told to press a button as quickly as possible when they hear the target. For the purpose of investigating effects of suprasegmental units on perception, the targets are either a CV or a CVC sequence. The stimuli (aside from distractor items) all begin with the entire CVC sequence, but in some stimuli, the second consonant is the onset of the next syllable, while in others it is the coda of the first syllable. Thus, in some cases, the target exactly matches the first syllable of the stimulus, and in others, the target is segmentally present but comprises either less or more than the first syllable of the stimulus. An example for English appears in (1).

(1) Target	Stimulus with matching syllable structure	Stimulus with unmatched syllable structure
"BA"	balance	balcony
"BAL"	balcony	balance
"PA"	palace	palpitate
"PAL"	palpitate	palace

Similar pairs of stimuli can be constructed in other languages. The measures are response time (time from onset of the word until the subject presses the button) and miss rate (percent of stimuli containing the target segmental material to which subjects fail to respond at all).

Early work with this paradigm was on French, using the French words equivalent to "balance, balcony" as one of the stimulus pairs (Mehler et al. 1981). French listeners in this experiment responded more quickly and accurately when the target was exactly equal to the first syllable of the stimulus. They responded more slowly both when the target was less than the entire syllable of the stimulus ("balcony" for the "BA" target) and when the target crossed a syllable boundary of the stimulus ("balance" for the "BAL" target). These results for French have since been replicated several times, and are highly reliable. It appears that at some stage of processing the speech signal, French speakers segment the signal into syllables.

Cutler et al. (1986) extend this work to English, using pairs such as the ones shown in (1). However, they found that English listeners do not show the "syllabic effect" of the French experiment at all. Rather, English listeners respond more quickly to CVC.CV stimuli than to CV.CV stimuli, regardless of which type of target they are asked to listen for<sup>6</sup>. The authors present several variations on this experiment, and Cutler et al.

---

<sup>6</sup> The use of /l/ as the second consonant for CV.CV and CVC.CV words may seem unwise, as the phonetic realization of English /l/ in onset and coda position is quite different, and the perceptual cues to the two allophones of /l/ are also rather different (Gesuato 1996). /l/ in onset position is an alveolar lateral, while in coda position it is velarized, and is often a back glide with no alveolar contact. However, since listeners were shown the target orthographically ("BAL" for the CVC target), they were probably monitoring for the phonemic sequence /bæl/, which is present regardless of which allophone of /l/ appears. The timing of availability of cues for the two allophones of /l/ might introduce some difference, though.

(1992) present results of the same experiment with English-French bilingual listeners. The authors conclude that infants acquiring language learn a strategy for segmenting the signal which is effective for the language they are learning. They relate these results to the predominance of stress in English, and the fact that English is said to be stress timed, while French is said to be syllable timed. While this experimental method does not show effects of stress in English, but only the lack of an effect of the syllable, the importance of stress in English speech perception has been well documented with other experimental paradigms (Cutler and Norris 1988, Cutler and Butterfield 1992, Cutler and Carter 1987).

Otake et al. (1993) and Cutler and Otake (1994) adapted this paradigm to Japanese, using the stimuli in (2) (Otake et al. 1993:263).

(2) Targets	CV.CV stimulus	CVC.CV stimulus
"MO/MON"	/monaka/ 'bean jam wafer'	/moNka/ 'in someone's tutelage'
"KA/KAN"	/kanoko/ 'fawn'	/kaNko/ 'lagoon; cheer'
"SA/SAN"	/sanaka/ 'in the midst of'	/saNka/ 'participation'
"NA/NAN"	/nanoka/ 'seventh day'	/naNka/ 'something'
"KI/KIN"	/kinori/ 'be interested in'	/kiNri/ 'interest rate'
"HA/HAN"	/haneda/ place name	/haNda/ 'solder'
"SHI/SHIN"	/sinigao/ 'face when dead'	/siNgao/ 'newcomer'

They predicted that if Japanese listeners used a syllable-based segmentation approach, as French listeners do, their reaction times would be faster for CV targets in CV.CV stimuli and for CVC targets in CVC.CV stimuli. However, if they used a mora-based strategy, they would respond equally quickly to CV targets in either stimulus type, and would respond more slowly to CVC targets in the CVC.CV stimulus type because they have to recognize two moras. Responses to CVC targets in CV.CV stimuli would be either slow or inaccurate, since the target does not comprise a whole number of moras.

These are indeed the results Otake et al. (1993) found: the Japanese listeners responded equally quickly and accurately to CV targets regardless of which type of stimulus they were in, responses to CVC targets in CVC.CV stimuli were accurate but slower, and the miss rate for CVC targets in CV.CV stimuli was extremely high (nearly 70%). (This pattern of a strong effect of target type is the opposite of the English results, which had an effect of stimulus type regardless of target type.) The authors conclude that



the Japanese listeners show no sign of a syllable-based strategy, and instead are using a mora-based strategy. They relate this to the claim that Japanese is mora-timed, while French is syllable-timed. Whether mora-, syllable-, and stress-timing are the source of these effects or not, there is certainly a striking difference between the results from French, English, and Japanese listeners to the same types of stimuli.

However, there are several problems with this experiment which may not be avoidable within Japanese phonology. In the English and French cases, the third phoneme of the CVC target was the same phoneme (/l/ in the English example above) whether it occurred at the end of a syllable or the beginning. In Japanese, however, the final "N" of the CVC target is the moraic nasal /N/ in the CVC.CV stimuli, but the non-moraic phoneme /n/ in the CV.CV case. This may seem to have been exactly the goal the authors had in constructing the stimuli, but in English and French, there is no phonemic contrast between /l/ at the end of a syllable and /l/ at the beginning of one. The contrast for Japanese is shown by (near) minimal pairs such as those in (3).

(3) /tani/	'valley'	/taNi/	'unit'
/kana/	'kana syllabary'	/kaNaN/	'take into account'
/mone/	'Monet'	/moNee/	'porter'
/kani/	'crab'	/kaNi/	'simplicity; official rank'
/kaneN/	'combustible'	/kaNeN/	'hepatitis'

The targets were presented to the Japanese listeners in Romanization, because writing the targets in either one of the *kana* syllabaries would require specifying whether the final "N" of the target was the moraic or the non-moraic nasal phoneme. The authors hoped that writing the targets in Romanization would leave the phonemic (and hence moraic) status of "N" ambiguous. Thus, they interpret the high miss rate for CVC (=CVN) targets in the CV.CV words as showing that subjects found it difficult to detect a target which does not begin and end at mora boundaries.

However, I do not believe it is possible to present the target in such a way that it remains phonemically ambiguous for Japanese subjects. I suspect that most, if not all, subjects immediately converted the Romanized targets into a representation which could be

written in Japanese, that is, into *kana*. The authors address this problem and argue that this is not the case, but their arguments are unconvincing. Even if subjects do not convert the Romanized target into an orthographic Japanese representation, they must process the Romanized target as some string of Japanese phonemes. *Kana* conversion, or even conversion into a string of phonemes, would make the "MON" type target into /moN/, which is not ambiguous with the /mon/ of /monaka/. Thus, the high miss rate for CVC targets in CV.CV words is not surprising, since the target is not phonemically present in the stimuli at all. This is a very different situation from that which faced the English and French listeners looking for "BAL" in "balance." This problem is nearly unavoidable for this experimental paradigm in Japanese. In addition, when the following consonant is a velar (as in five of the eight stimulus pairs), the N in the CVNVCV stimulus will assimilate to the place of the velar, while the non-moraic /n/ in the CVCVCV will remain alveolar. This introduces a further difficulty for listeners listening for the CVC targets<sup>7</sup>. I do not think that the results for CVC targets in CV.CV stimuli can be considered as evidence for mora-based segmentation. The various difficulties introduced by Japanese phonology do not affect the conditions with CV targets, however. The slow responses to CVC targets in CVC.CV words are probably also reliable, as converting the target into *kana* would not create a phonemic conflict with these stimuli.

Cutler and Otake (1994) substantiate the results of the CV/CVC target spotting experiments with an individual segment target spotting experiment on English and Japanese, in which the segment to be spotted sometimes comprises an entire mora and sometimes comprises only the onset or nucleus of a mora. Results from this experiment, although there are some difficulties, also support the hypothesis that Japanese listeners

---

<sup>7</sup> In the English experiment, there is allophonic variation in the realization of the second consonant (when it is /l/). In the Japanese experiment, there is both a phonemic mismatch between the CVC target and the CV.CV stimuli and also allophonic variation in the realization of the second consonant when the third consonant is velar. The Japanese stimuli share the allophonic variation problem the English stimuli have, and have an additional problem at the phonemic level.

parse the signal into moras. In sum, not all of the results for the Japanese CV/CVC target spotting experiment can be taken at face value, but there is clearly a difference in the suprasegmental units listeners of Japanese, French, and English use in perceiving speech.

Cutler and Otake (to appear) show another way in which the processes of speech perception and spoken word recognition are language dependent. They gated Japanese words which could have either high-low or low-high tones on the first two moras, and presented these to listeners in an open response gating experiment. They evaluated the percent of responses with pitch accent location matching that of the stimuli (at least as to having or not having pitch accent on the first two moras; pitch accent could diverge after the gated area). They found that Japanese listeners (at least speakers of the Tokyo dialect) gave responses with the correct pitch accent pattern significantly more often than chance (nearly 80% correct) even when the word was gated out at the middle of the first vowel. Thus, Japanese listeners use pitch accent information to narrow the cohort for spoken word recognition quite early in the word. They compare this result to a previous study on English by Cutler (1986), in which she found that English listeners do not consider stress in spoken word recognition at early stages of the word unless the difference in stress is accompanied by a difference in vowel quality. Specifically, hearing "forbear" with second syllable stress will initially activate the word "forbear" (first syllable stress) as well as the target word, and vice versa. Cutler and Otake propose that this is because in English, there are very few pairs of lexical items which are distinguished only by stress with no difference in vowel quality. However, they cite Sugitoo (1995) as showing that the Tokyo Japanese lexicon has a large split between words beginning with a high-low sequence (40%) and words beginning with a low-high sequence (60%). Thus, paying attention to suprasegmental information early in the word would be of more use in recognizing spoken words in Japanese than in English.

In sum, many researchers, using a variety of experimental paradigms, have identified language specific differences in speech perception and spoken word recognition.

Such differences are evident in the weighting of potential cues to a distinction (Fujimura et al. 1978, Kakehi et al. 1996), in the use of suprasegmental units, perhaps to parse the signal (Cutler and colleagues), and in attention to potential cues during spoken word recognition (Cutler 1986, Cutler and Otake to appear). While differences between English and French, Dutch and Spanish (Costa et al. in press), Dutch and Japanese, and other languages have also been documented, the two languages for which the most differences in perception and spoken word recognition have been tested and found are Japanese and English. Therefore, Japanese and English are a good choice for languages to compare in cross-linguistic speech perception research.

#### 1.6. Summary

In this chapter, I have provided an overview of the reasons, based in auditory processing, for thinking that changing portions of the speech signal may be more important for speech perception than steady states. In Section 1.2, I also reviewed previous work on the issue of dynamic versus static cues in speech perception. Section 1.3 gave a brief introduction to the cohort model of spoken word recognition and introduced several issues in the field of spoken word recognition which will be relevant to the study reported here. Section 1.4 discussed methodological issues regarding the gating method. In Section 1.5, I reviewed literature which shows that there are language specific differences in speech perception, spoken word recognition, and word segmentation, and that these differences are often founded in the phonological systems of the languages.

The rest of the dissertation will be structured as follows: in Chapter 2, I provide the details of the experimental design for a large scale experiment designed to test English and Japanese listeners' use of dynamic and static cues. This chapter addresses questions such as how the stimulus words were chosen, the characteristics of the subjects, and how the experiment was run. In the third chapter, I discuss Furui's measure of degree of spectral change  $D$ , and give the results of the acoustical measurements for the experiment. Since

the measure D is not yet well known, I intend this chapter to provide enough information and examples of this measure to make it more accessible for others. Chapter 4 presents the results of the perceptual experiment, and includes several different analyses of the data.

Finally, in Chapter 5 I discuss the results of the experiment with regard to the use of dynamic cues in speech perception and spoken word recognition, and elucidate the importance of the rate of change in dynamic cues. I show that the experimental results support some aspects of theories of spoken word recognition, but also require some changes in those models. The results further demonstrate that the phonological system of a language (facts about its phoneme inventory and syllable structure constraints) affect the timing of how listeners use perceptual cues in the speech signal. Based on the experimental results, I offer perceptual motivations for several well known phonological universals or common phonological alternations, such as the preference for consonants in syllable onset position rather than coda position, dissimilation, and alternations between glides and high vowels. I argue that the presence or lack of rapidly varying cues is an underlying factor for all three of these patterns. I discuss some ways previous researchers have attempted to involve formal phonological models in questions of speech perception, and show that my results support one such proposal but contradict another. Finally, I argue for the importance of such experimental studies of the timing of speech perception as an important way to investigate perceptual motivations for phonological universals, including syllable structure phenomena.

2. Methodology
- 2.1. The English experiment
- 2.1.1. Choice of stimulus words
- 2.1.1.1. The two segment sequence for gating

I designed a gating experiment to test the use of dynamic cues in speech perception and spoken word recognition. In order to gain information about spoken word recognition, stimuli must consist of gated portions of real words, not of short nonsense syllables, as discussed in Section 1.4.4 above. Although many previous gating experiments (Grosjean 1980, Tyler and Wessels 1985) gate through the entire word (that is, the shortest gate consists of only the very beginning of the word, and the longest gate includes the entire word), these experiments usually use a relatively long gating interval, such as 50 ms. Furui (1986), who uses a short gating interval (10 ms), uses only monomoraic isolated syllables as stimuli. Because testing the dynamic theory of speech perception through gating requires that the gating interval be shorter than the shortest segment to be tested, for my experiment, I chose a gating interval of 20 ms. Whole words must be used, but gating through entire words at a 20 ms interval would result in a very large number of stimuli for each word. This is extremely impractical if one uses individual instead of successive presentation of stimuli (as discussed in Section 1.4.5 above) because of the number of subjects required.

For these reasons, I chose to gate each word only through a two segment section of the word. That is, the shortest gate allowed the listener to hear from the beginning of the word up to a point part way through the first segment of interest, and the longest gate allowed the listener to hear from the beginning of the word up to a point part way through the subsequent segment (the second segment of interest). (Selection of the points during these segments to use for the first and last gate endpoints will be discussed below.) The two segment sequence of interest often does not comprise the first two segments of the word, so listeners usually hear more than two segments, but the part of the word during

which endpoints of stimuli can fall comprises a two segment sequence. For example, in the word "attempt," of which the /ɛm/ sequence was of interest, the shortest gate allowed the listener to hear from the beginning of the word to a point part way through the /ɛ/, and the longest gate allowed the listener to hear from the beginning of the word to a point part way through the /m/.

Gating through the transition between two segments (a diphone) instead of through an entire word is acceptable, because the question is the timing of listeners' perception of segments and how that influences spoken word recognition. There is no reason why perception of several segments must be investigated within the same word. By choosing to investigate a particular two segment transition in each word, one can better manipulate the characteristics of the segments in question without multiplying the number of subjects required. In the case of "attempt," ([ə<sup>h</sup>tɛmpt]) the transitions /ət/, /tɛ/, /mp/, /pt/ might not be the transitions one is most interested in investigating. Even if one does wish to investigate those transitions, one can do so more efficiently by choosing different words containing them, so that the same subjects can be used for all the words.

#### 2.1.1.2. Factors manipulated in the choice of transitions

Within the two segment sequences which are of interest, I manipulated several factors. Most importantly, both consonants and vowels appeared as both the first and second segments of interest, that is, the test transitions included CV, VC, CC, and VV sequences. Previous work on dynamic cues in speech perception has usually been limited to highly constrained sequences of phonemes: Kewley-Port et al. (1983), Strange et al. (1983), and Stevens and Blumstein (1981) all use stop-vowel or stop-vowel-stop stimuli which do not generally form real words. Even Furui's (1986) gating study examines only CV and CyV syllables produced in isolation. Therefore, his study addresses only CV, Cy, and yV transitions, and does not investigate any effects of position in the syllable or the

word, or of accent placement. Furui does use all possible Japanese CV and CyV syllables, but most previous studies on English use only a few carefully selected consonants and vowels. Thus, none of these studies shed light on the role of dynamic cues in the perception of a consonant following another consonant or on the effect of stress on the timing of segment perception. Most do not even address the possible effects of manner of articulation or vowel quality on the timing of perception, and while Furui's (1986) study was groundbreaking, it does not consider VC transitions, which are known to have very different perceptual properties from CV transitions (Fujimura et al. 1978, Kakehi et al. 1996).

The words used for the current study allow for manipulation of the place, manner, and voicing of the consonant(s) in CV, VC, and CC transitions and of the vowel quality in CV, VC, and VV transitions. In addition, I manipulated the relationship of the two segment sequence to the stress in the word: CV and VC transitions could appear with the vowel in either stressed or unstressed syllables, and VV transitions could have stress on either the first or the second vowel. The manipulation of stress for CC transitions was more complicated, but has the potential to provide very interesting results: the two segments of the CC transition could be both in the same syllable, split across a syllable boundary but within the same stress foot, or split across a foot boundary. Where phonotactic constraints allowed, word-initial, final, and medial position was also varied for CC transitions. Finally, position in the word was also manipulated for CV and VC transitions, which appear in either the first or the second syllable of the word. I made no attempt to include words with these transitions in the third or later syllable of words, as it is difficult to find words for which there is still ambiguity about what the word is at points later than the second syllable. For CC transitions, the stress and syllable manipulation replaced the first/second syllable manipulation, and for VV transitions, no attempt was made to position the VV sequence any later than the first and second syllable. For consonants which have well known allophonic variations, many of the allophones were



already varied by means of varying the stress (aspiration or lack thereof, flapping, etc.). I included some additional words in order to have each allophone of segments such as /t/ appear, even if all allophones are not conditioned by stress placement. For example, the word "stiff" [stɪf] was included in order to have the sequence /tɪ/ with unaspirated [t] conditioned by the preceding /s/ as well as examples of unaspirated [t] which is conditioned by a following unstressed vowel.

It is neither possible nor practical to include all combinations of all of the factors discussed above (place, manner, voicing, or vowel quality of both segments, consonantal or vocalic status of both segments, and position relative to stress and/or syllable, foot, and word boundaries). Phonotactic constraints prevent the juxtaposition of many segment types or limit their placement relative to word boundaries, and the number of stimuli provides a practical constraint, since a different word must be used for each combination of factors. Therefore, I did not attempt to make all possible combinations of the factors, but instead chose a few words which differed only by each of the factors, so that these small groups of words can be compared. CC transitions which can only appear word medially (and perhaps only when crossing a morpheme boundary) were usually avoided.

#### 2.1.1.3. The word list

128 English words, shown in (1), were chosen using the machine searchable Carnegie Mellon Pronouncing Dictionary (1995). For each word, the two segment sequence of interest is shown in bold in the transcription, and the manipulated factors are described after the transcription. Since the consonant in a VC transition is frequently not in the same syllable as the vowel, VC transitions are described as being in the first or second syllable and stressed or not based on their vowels.

## (1) English word list

## Stop-vowel and vowel-stop transitions

## /t/

tip	[t <sup>h</sup> ɪp]	First syllable, stressed, aspirated
stiff	[stɪf]	First syllable, stressed, unaspirated
Tibet	[t <sup>h</sup> ɪbet]	First syllable, unstressed, aspirated
petition	[pə'tɪʃən]	Second syllable, stressed, aspirated
attic	['ærɪk]	Second syllable, unstressed (flap)

## /k/

custom	[ˈkʰʌstəm]	First syllable, stressed, aspirated
skull	[skʌl]	First syllable, stressed, unaspirated
accompany	[əˈkʰʌmpəni]	Second syllable, stressed, aspirated
caboose	[kʰə'bus]	First syllable, unstressed, aspirated
academic	[ækə'demɪk]	Second syllable, unstressed, unaspirated

## /d/

duck	[dʌk]	First syllable, stressed, voiceless unaspirated
diploma	[dʒə'ploʷmə]	First syllable, unstressed, voiceless unaspirated

## /t/

citizen	['sɪrəzn]	First syllable, stressed, flap
fitness	['fɪt <sup>h</sup> nəs]	First syllable, stressed, glottalized/unreleased
Italian	[ɪt <sup>h</sup> æliən]	First syllable, unstressed, aspirated
committee	[kə'mɪtɪ]	Second syllable, stressed, flap
unity	['ju:nɪtɪ]	Second syllable, unstressed, flap

## /ʌk/

bucket	['bʌkət]	First syllable, stressed, unaspirated
mechanical	[mə'kʰænəkəl]	First syllable, unstressed, aspirated
indicate	['ɪndəkeɪt]	Second syllable, unstressed, unaspirated
induction	[ɪn'dʌkʃən]	Second syllable, stressed, unaspirated

## /ʌd/

muddy	['mʌdɪ]	First syllable, stressed, flap
cadenza	[kə'denzə]	First syllable, unstressed, (possibly voiceless)

## Nasal-vowel and vowel-nasal transitions (all are stressed)

## /mɛ/

medicine	['medɪsn]	First syllable
immense	[ɪ'mens]	Second syllable

## /ɛm/

remedy	['remədi]	First syllable
attempt	[ə'tempt]	Second syllable

/nɛ/	negative	['nɛgətɪv]
/ɛn/	tenants	['tɛnənts] <sup>1</sup>

#### Fricative-vowel and vowel-fricative transitions

/sæ/	saddle	['særl]	
/æs/	master	['mæstə]	
/zæ/	Zachary	['zækəri]	
/æz/	asthma	['æzmə]	
/ʃɛ/	shell	[ʃɛl]	
/ɛʃ/	session	['sɛʃən]	
/fɪ/	fees	[fɪz]	First syllable
	unfeeling	[ən'fi:lɪŋ]	Second syllable
/lɪ/	leaf	[lɪf]	First syllable
	relief	[rɪ'lɪf]	Second syllable
/væ/	vacuum	['vækjʊm]	
/æv/	ravish	['rævɪʃ]	

#### Approximant-vowel and vowel-approximant transitions

/reɪ/	trail	[treɪl]
/eɪr/	fair	[feɪr]
/lɛ/	lever	['lɛvə]
/ɛl/	elevator	['ɛlə,veɪrə]
/jɛ/	yellow	['jɛləw]

<sup>1</sup> There may be glottalization of [t] in this and several other words in the list. This will only be transcribed if the [tʔ] is in the transition of interest, as there is variation in whether final stops are glottalized or not.

/wə/  
watch [wətʃ]<sup>2</sup>

#### Affricate-vowel and vowel-affricate transitions

/tʃæ/  
chapel [tʃæpəl]

/æʃ/  
latches [lætʃɪz]

/dʒʌ/  
jump [dʒʌmp]

/ʌdʒ/  
judge [dʒʌdʒ]

#### Nasal-stop transitions (stop-nasal not used because of limited distribution)

/nt/  
bent [bent<sup>2</sup>] Within syllable  
sentiment [ˈsentəmənt] Crosses syllable boundary, within foot  
reinterpret [ˌrɪnɪˈtɜːprət] Crosses foot boundary

/nd/  
band [bænd] Within syllable  
wander [ˈwɑːndə] Crosses syllable boundary, within foot  
reconditioned [ˌrɪkənˈdɪʃənd] Crosses foot boundary

#### Stop-fricative and fricative-stop transitions

/ks/  
axe [æks] Within syllable  
hacksaw [ˈhæksə] Crosses syllable boundary, within foot  
unacceptable [ˌʌnəkˈseptəbəl] Crosses foot boundary

/ts/<sup>4</sup>  
cats [kæts] Within syllable  
Betsy [ˈbetʃi] Crosses syllable boundary, within foot

/st/<sup>5</sup>  
stop [stap] Word initial  
based [beɪst] Word final  
pastime [ˈpæstəɪm] Word medial

<sup>2</sup> Although the author distinguishes /ə/ and /ɔ/, neither the speaker for the stimulus words nor the majority of the subjects do. Therefore, the distinction is shown as neutralized in this transcription.

<sup>3</sup> The /t/ in this word is often produced as a flap or deleted entirely, but in the careful speech situation of the recording, the speaker for the experiment did produce a stop.

<sup>4</sup> No good example of this sequence crossing a foot boundary could be found, although there may be one that my methods of searching the dictionary did not locate.

<sup>5</sup> For /s/-stop transitions, position relative to syllable and foot boundaries cannot be varied, since the /s/-stop cluster is usually considered part of the following syllable. Word initial, medial, and final position can be varied, however.

/sk/	skate	[skeɪt]	Word initial
	mask	[mæsk]	Word final
	discount	['dɪskaʊnt]	Word medial

#### Stop-approximant and approximant-stop transitions

/tɹ/	train	[t <sup>h</sup> reɪn]	First syllable, aspirated
	string	[strɪŋ]	First syllable, unaspirated
	Detroit	[dɪt <sup>h</sup> roʊt]	Second syllable, aspirated
/kɹ/	crops	[k <sup>h</sup> raps]	First syllable, aspirated
	scrap	[skræp]	First syllable, unaspirated
	acrobat	['ækroʊbæt]	Second syllable, unaspirated
/dɹ/	drop	[drɒp]	
/gɹ/	groan	[ɡroʊn]	
/pl/	plain	[p <sup>h</sup> leɪn]	First syllable, aspirated
	split	[splɪt]	First syllable, unaspirated
/tw/	twelve	[t <sup>h</sup> welv]	
/ɹt/	court	[kɔrt <sup>ɹ</sup> ]	
/rk/	cork	[cɔrk]	
/lp/	help	[help]	

#### Nasal-fricative and fricative-nasal transitions

/nz/	fans	[fænz]	
/ns/	dance	[dæns]	Within syllable
	fancy	['fænsɪ]	Crosses syllable boundary, within foot
	unconcealed	[,ʌnkən'sɪld]	Crosses foot boundary
/sn/	snow	[snoʊ]	
/zn/	Disney	['dɪzni]	

Approximant-nasal transitions (nasal-approximant not used because of limited distribution)

/r̩n/      farm      [fɑr̩m]

/r̩/      corn      [kɔr̩n]

/l̩m/      film      [fɪl̩m]

Nasal-affricate transition (affricate-nasal not used)

/ntʃ/      ranch      [ræntʃ]

Fricative-approximant and approximant-fricative transitions

/fl/      flash      [flæʃ]

/fr/      fragile      [ˈfrædʒl]

/sl/      sleep      [slɪp]      Word initial  
Iceland      [ˈaːslænd]      Word medial

/sw/      swan      [swan]

/f/      golf      [gɒlf]

/rf/      wharf      [wɔrf]<sup>6</sup>

/ls/      false      [fals]      Within syllable  
calcium      [ˈkælsiəm]      Crosses syllable boundary, within foot

Approximant-affricate transitions

/ltʃ/      cultural      [ˈkʌltʃə]

/rdʒ/      marginal      [ˈmɑrdʒən]

Stop-stop transitions

/pt/      optical      [ˈɒptɪkəl]

/kt/      pact      [pækt]

<sup>6</sup> Although this dialect does not distinguish /a/ and /ɔ/, [ɔ] does surface as the allophone of /o<sup>w</sup>/ before /r/.

## Fricative-fricative transitions

/fs/	coughs	[kafs]
/vz/	nerves	[nəvz]

## Nasal-nasal transition

/mn/	amnesty	[ˈæmnəsti]
------	---------	------------

## Approximant-approximant transition

/rɪ/	garlic	[ˈgɑrlɪk]
------	--------	-----------

## Vowel-vowel transitions

/aˈa/	biopsy	[ˈbɑːpsɪ]	First vowel stressed
	biography	[bɑːˈɡrɑːfi]	Second vowel stressed
/aˈoʷ/	biotech	[ˈbɑːoʷˌtɛk]	First vowel stressed
/iə/	eon	[ˈiən]	First vowel stressed
/aˈæ/	diagonal	[daːˈæɡənəl]	Second vowel stressed
/iæ/	react	[riˈækt]	Second vowel stressed

## Transitions involving diphthongs, [ə], or syllabic consonants

/tə/	tiger	[tˈɪgə]
/aɪ/	bite	[baɪt̚]
/de/	data	[ˈdeɪtə]
/eɪ/	fade	[feɪd]
/aʷt̚/	doubt	[daʷt̚]
/oʊb/	soybean	[soʊˈbiːn]
/toʷ/	toad	[toʷd]
/oʷt̚s/	oats	[oʷt̚s]
/kə/	courage	[ˈkəʊədʒ]

/ɜ:k/	circle	['sɜ:kɪ]
/tʌ/	button	['bʌtʌn]
/t/	beetle	['bi:tɫ]
/p/	apple	['æpɫ]

Of course, there are often several possible ways to transcribe the words above, or several possible ways to pronounce them. In most cases, different pronunciations will have no effect on the experiment. The realization of any segment after the two segments of interest does not matter, as no subject will ever hear the segments after the last gate. Thus, whether the third and fourth syllables of a word such as "diagonal" ([da<sup>i</sup>ˌæɡənɪ]) are actually produced with reduced vowels or syllabic consonants does not matter. If there are discrepancies between the phonetic transcription given here and the pronunciation of the speaker for the experiment in segments before the segments of interest, this will not matter as long as the speaker's pronunciation is one that subjects recognize.

One should note that while not all possible combinations of place, manner, voicing, or vowel quality of both segments could be used, I made every effort to include transitions representative of as wide a variety of these factors as possible. For example, in the fricative-vowel and vowel-fricative transitions, both strong (voiceless sibilant) and weak (voiced non-sibilant) fricatives, as well as voiced sibilant and voiceless non-sibilant fricatives, appear. In the VV transitions, it is not often possible to have the same VV sequence with stress on the first V in one word and on the second V in another because of reduction of unstressed vowels. One such case ("biopsy" ['bæ<sup>i</sup>ˌɒpsi], "biography" [ba<sup>i</sup>ˌɒɡrəfi]) was included, and the remaining VV transition words were chosen to represent a wide variety of vowel quality combinations with some variation of stress placement. Except where phonotactically impossible, the same segments used for a CV



transition also appear in a corresponding VC transition. (The transitions /o<sup>b</sup>/ in "soybean," and /a<sup>w</sup>/ in "doubt" however, were included without corresponding /bo<sup>j</sup>/ or /ta<sup>w</sup>/ transitions through an error.) Even though the transitions used represent far from all the possible combinations of all factors, the variety of transitions appearing in these stimulus words represent a great step forward from the extremely restricted sequences of segments used in past work on dynamic cues in speech perception. Therefore, this experiment allows for a test of the theory in much wider, and more realistic, environments. If one is truly interested in the timing of perception of English segments, one cannot continue to restrict the tests to stop-vowel and vowel-stop sequences.

#### 2.1.1.4. Other potential effects

In addition to the factors discussed above, which are varied in the word list, there are several potential influences which are controlled through the choice of words, or which cannot be controlled. The most important thing to control for in the choice of stimulus words is the change in cohort size from the first of the two segments of interest to the second. Using a simple version of the cohort model, the cohort consists of all the words in a listener's English lexicon which begin with the string of segments the listener has heard up to a given point. For example, the cohort for the string /wo<sup>w</sup>r/ [wɔːr], which begins the stimulus word "wharf," includes words such as "war, ward, worn, whorl, warble, warhead, warm" and morphologically related forms such as "wars, warring, wards, warmly" etc., along with "wharf" and its homophones "Whorf" (the linguist) and "Worf"

.

---

(a character in a popular television series)<sup>7</sup>. When a listener hears enough of the word "wharf" to perceive the /f/, however, the cohort is greatly reduced, probably to "wharf, Whorf, Worf, warfare" and any morphologically complex forms beginning with these morphemes, such as "wharfside."

The cohort size itself (as opposed to the change in cohort size) of any stimulus word at the point of the transition of interest is not important. In stimuli for which the transition of interest consists of the first two segments of the word, the cohort size, especially before the second segment is perceived, will be very large. For example, the cohort of the stimulus "saddle" (/særl/) at a point before the vowel /æ/ is perceived will consist of all words in the English lexicon beginning with /s/. However, it is important that there be a relatively large change in the cohort size of the stimulus between the stage of the first segment of interest and the stage of the second segment of interest. If the cohort of the stimulus is the same at the first and second segments of interest, listeners will have as much lexical information about what the word could be at the first segment as at the second, and thus exactly the same answers will be possible whether they have perceived the second segment yet or not. This makes it impossible to identify a point at which they recognize the second segment. In the word "migrant," for example (which was not used in the experiment), the cohort both for /mæ<sup>1</sup>g/ and /mæ<sup>1</sup>gr/ is the same, consisting of "migrate, migrated, migrating, migraine, migraines, migrant, migrants, migratory" (based on a search of the CMU dictionary) since there are no English words (in the CMU dictionary) which begin with /mæ<sup>1</sup>g/ followed by anything other than /r/. In constructing the word list, I chose words which had as large a change in cohort size during the transition of interest as possible.

The change in cohort size from the first segment of interest to the second should be as large as possible, but like the raw cohort size, will vary depending partly on how far

---

<sup>7</sup> Neutralization of /w/ and /v/ is assumed.

from the beginning of the word these segments fall. In cases such as "saddle" (/særl/) or "snow" (/snoʷ/), the change in cohort size is quite large, since the cohort at the first segment of interest consists of all English words beginning with /s/, and the second segment rules out a large proportion of the cohort. When the transition in question falls later in the word, the cohort may change from three or four words to just one. This is especially true of the stimulus words chosen to split CC transitions across foot boundaries or to have VC transitions in the second syllable, as both of these conditions require that several segments precede the transition in question. In many of these cases, such as the stimulus words "unacceptable" ([ʌnək'septəbl̩]) and "reconditioned" ([ˌrɪkən'dɪʃənd]), the only way to find an appropriate word with any change in cohort size at all is to use words beginning with two prefixes (or strings which derive historically from prefixes). "Unacceptable" at the point of the /k/ has a cohort including "unaccountable, unaccountably, unaccompanied, unaccredited, unaccustomed," as well as "unacceptable," but once the /s/ is included, the cohort drops to just "unacceptable" and forms derived from it ("unacceptably") (CMU 1995)<sup>8</sup>.

This leads to the question of controlling morphological complexity. While it is very likely that lexical access of words with affixes might proceed somewhat differently than access of morphologically simple words, words with some affixes or with strings which derive historically from Latin prefixes were included in the experiment. I made no effort to avoid morphologically complex words. Listeners must perform lexical access for morphologically complex words as well as simple ones in the perception of real speech, and in many cases, use of complex words was necessary in order to manipulate factors

---

<sup>8</sup> "Unexciting" is probably also a member of the /ʌnəks/ cohort. Some words familiar enough for listeners to recognize them are probably missing from the searchable dictionary, and variation in the phonemic transcription of reduced vowels (as schwa or as unstressed full vowels) can also lead to failure to include a word in a particular cohort. One can never be sure of the exact cohort a particular listener will have because of differences in individual's lexicons, so problems with the dictionary only add to this uncertainty slightly.

such as boundary placement (as mentioned above) or to allow for the combination of certain manners of articulation (word final /nz/ usually occurs where /z/ is a suffix, as in "fans" /fænz/). However, a large proportion of the stimulus words are monomorphemic. If a disproportionate number of them were complex, or perhaps began with many of the same prefixes (such as "un-, non-, com-"), listeners might begin to pay more attention to morphological patterns than they do in normal speech perception. This was one reason for not attempting to place every transition in both the first and second syllables.

While I did not control morphological complexity, only content words, not function words, were used as stimulus words. This division is in keeping with psycholinguistic evidence that content words and function words are treated differently in lexical access (Cutler and Norris 1988, Grosjean and Gee 1987). This does not mean, however, that subjects could not respond with a function word.

I did not control word frequency in this experiment. It does not matter whether the words used are of high or low frequency, as long as they are of sufficient frequency for listeners to think of them readily. Ideally, one would like all members of a stimulus' cohort to be of approximately equal frequency, so that subjects' responses would be equally distributed throughout the cohort, instead of most subjects concentrating their answers on the most frequent members of the cohort. However, it is not possible to find (many) words in English for which all or most members of the cohort are of approximately the same frequency. Therefore, no attempt was made to control this point. In some cases, the number of different responses given by the subjects may be artificially low if one or two members of the cohort have much greater frequency than others. This possibility should be considered in interpreting the results for individual words, but cannot be avoided. As for overall frequency, the words used for the experiment ranged from very frequent to rather infrequent. In some cases, such as "unconcealed," ([ʌnkən'sild]), the word may have been of too low frequency: some subjects found it difficult to think of an appropriate

answer to the stimulus [<sub>r</sub>ʌnkən's]. This problem was relatively rare, however, and usually only occurred with the stimuli which were designed to split CC transitions across foot boundaries.

A further problem occurs when the stimulus up through the first segment of interest is itself a whole word, and particularly one of high frequency. Taking "fans" ([fænz]) as an example, when the stimulus ends at some point during the /n/, the cohort theoretically includes "fan, fans, fanned, fantasy, fantasize, fancy, fancied, fanfare, Fannie, Faneuil" etc. Subjects may respond almost exclusively with "fan," though, since it forms a word of relatively high frequency by itself. Whenever the stimulus itself forms a whole word, subjects will probably be less likely to supply a longer form beginning with that string than to answer simply with that string itself (Grosjean 1980, Tyler 1984). For CV, VC, and VV transitions, the stimulus at the stage of the first segment of interest often does not form a word by itself, but for CC transitions, it usually does. This is nearly unavoidable within the English lexicon. One must consider whether a shorter part of the stimulus forms a whole word (an embedded word) when interpreting the results for individual words.

Potential effects from cohort size, change in cohort size, word frequency, morphological boundaries, and stimuli which form whole words all relate to the same problem: the process of lexical access is inherently nonlinear. All of these problems could be avoided by using nonsense syllables with a closed class response as stimuli instead of using real words with an open response. However, while that methodology might give more easily interpretable results for the timing of segment perception, it would have no bearing on the question of timing of lexical access or spoken word recognition. If one wishes to investigate the use of dynamic acoustic cues in spoken word recognition as well as in segment perception, one must contend with these nonlinearities. Ways to recognize and address these problems in the interpretation of results will be discussed in Chapter 4.

## 2.1.2. Production of the stimuli

### 2.1.2.1. Recording of the data

The 128 words above were randomized, and one male speaker of North American English read the entire list twice. The speaker is from Colorado, and does not distinguish [a] and [ɔ]. (Since the majority of the listeners are young Californians, who also do not distinguish these vowels, this merger was not considered a problem.) The speaker is a linguist, but did not know the purpose of the experiment. In reading the list, the speaker read each word as a separate utterance, and did not use list intonation. The speaker was instructed to read at a slightly fast pace, but not unnaturally fast. This did not lead to exceptionally fast speech. The recording was done in a double-walled sound booth, with all recording equipment except a microphone outside the booth, in order to minimize background noise. The recording was done with a DAT recorder. The speaker and the experimenter could communicate using headphones during the recording, and the experimenter monitored the speaker's pronunciation of the stimulus words.

The data was digitized on the Kay Computerized Speech Lab system at a sampling rate of 16,000 Hz. Each word was trimmed to leave 400 ms of silence before the word. This data was transferred to the ESPS/XWaves system for further analysis and processing.

### 2.1.2.2. Gating

For each of the two tokens of each word, I chose two points, one to use as the endpoint of the first gate, and one to use as the endpoint of the last gate. In general, since the goal was to gate through a diphone of the word, the first point was at approximately the middle of the first segment of interest, and the second was at approximately the middle of the second segment of interest. For monophthongal vowels, nasals, and fricatives, a point approximately halfway through the segment was sufficient. For voiceless stops, affricates, and diphthongs, however, an arbitrary but consistent choice of the halfway point of the segment would be undesirable: depending on the duration of aspiration for a voiceless

stop, for example, some tokens might have the halfway point before the burst of the stop while others would have it after the burst. Arbitrary imposition of the halfway point as the first or last gating point would thus make comparison across words impossible, as some stop-vowel tokens would include pre-burst gates and others would not.

Therefore, for medial stops (both voiceless and voiced) and affricates in CV transitions, the initial gating point was placed shortly before the beginning of the burst. For these segments in VC transitions, the final gating point was placed after the burst, during the aspiration or affrication noise if any was present. For diphthongs, if the diphthong was the first of the two segments of interest, the initial gating point was placed in the middle of the steady state of the initial vowel quality of the diphthong. If the diphthong was the second segment of interest, the final gating point was placed in the middle of the steady state of the latter vowel quality of the diphthong. Of course, diphthongs do not always have steady state portions for both (or either) of their two parts, but it was not usually difficult to identify a point which appeared to be representative of the initial or final vowel quality of the diphthong. Another segment type which required special treatment was the flap: if a flap was the first segment of interest, the initial gating point was placed at the beginning of the flap, which corresponds to placement before the release for stops. If the flap was the second segment of interest, the final gating point was placed just after the "release" of the flap. For many sonorants, particularly for postvocalic /r/, it was difficult to identify any part of the sonorant separate from the vowel. If postvocalic /r/ was the first segment of interest, the initial gating point was placed late in the preceding vowel.

For word-initial and final segments in the transition of interest, I followed a different procedure. If the first segment of interest was word-initial, the initial gating point was placed 20 ms after the beginning of the segment instead of halfway through it. If the second segment of interest was word-final, the final gating point was placed at the end of the segment. This was because the word list included no silence-to-segment or segment-to-

silence transitions as the transition of interest, so if the initial and final gating points had been placed in the midpoint of the segment for these cases, no word would have had gates covering the first half of the word-initial segment or the last half of the word-final segment.

Once the initial and final gating points for all tokens had been determined, I selected one of the two repetitions of each word for use in the experiment. If there was a problem with one of the tokens of a word, such as a production error or a problem in the recording, the token without such a problem was used. For word final stops, if one token had an unreleased stop, the token with a release was used, and the final gating point was placed after the release. When the speaker had produced one token less clearly than the other (for example, "acrobat" as [ækəbæt] instead of [ækɹəbæt], or "vacuum" with no clear [v]), the more clear token was used. However, the majority of the words had two equally good tokens. In these cases, the token with a shorter difference between the initial and final gating points, that is, the one that had been read faster, was used. Since the number of listeners necessary for the experiment depends on the number of gates, this may have slightly decreased the necessary number of subjects. The speaker did not read any of the words at an unnaturally fast pace, so the choice of the shorter token should not introduce any undue effects.

The selected token of each word was gated at 20 ms intervals, with the first gate leaving the word intact from the 400 ms of silence before the word up to the initial gating point, then cutting to a square wave. The speech signal was ramped down and the square wave ramped up over a 10 ms window beginning at the gating point. After the gating point, the square wave continued for 500 ms. The second gate left the signal intact from the 400 ms of silence before the word up to 20 ms after the initial gating point, then cut to a square wave in the same fashion. Gates at each 20 ms interval were created until the final gating point was reached. See Figure 2.1 for an example of the resulting stimuli.

If the final gating point did not happen to be a multiple of 20 ms after the initial gating point, the following procedure was followed to determine the final gate: if the



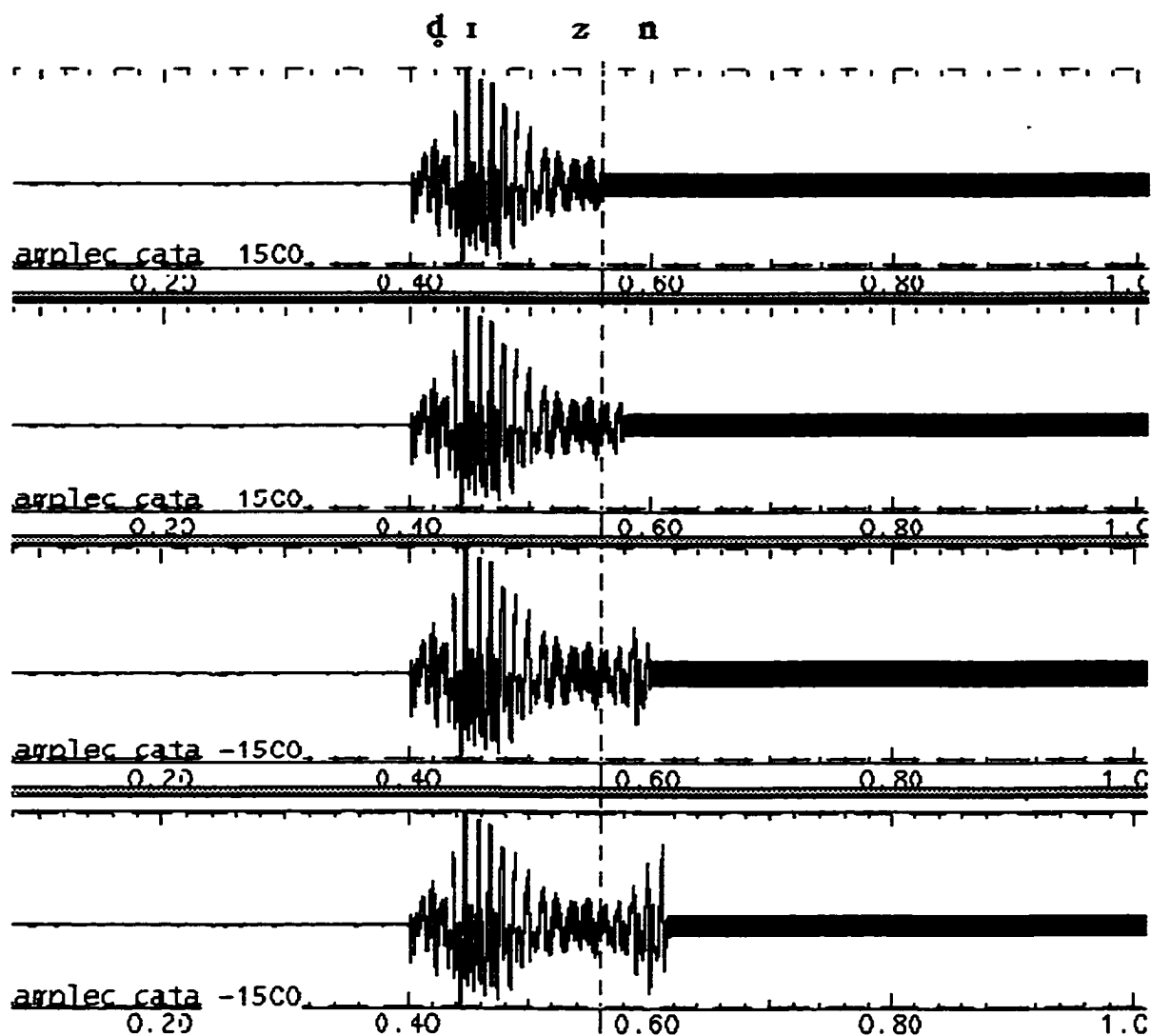


Figure 2.1. Gated stimuli for the word "Disney" /dɪzni/, in which the transition /zn/ is of interest. The shortest gate allows the listener to hear from before the beginning of the word to part way through the /z/, the first segment of the transition of interest. The next transition includes 20 ms more, the next 20 ms more, and the last extends to part way through the /n/.

predetermined final gating point was 10 ms or more after the last regular 20 ms interval gate to fall before it, I added an additional gate at exactly the predetermined final gating point. Thus, in these cases, the penultimate and final gate are separated by less than 20 ms. If the predetermined final gating point was 5 ms or less after the last regular gate to fall before it, no additional gate was added. In these cases, the actual final gate ended slightly before the predetermined final gating point. (In one case, through an error, the final gate was placed at the predetermined final gating point, and the last regular gate, which should have fallen 4 ms before the predetermined final gating point, was omitted. Thus, in this one case, the last and next to last gates are separated by 24 ms instead of by 20 ms.) Finally, if the predetermined final gating point fell 6-9 ms after the last regular gating point before it, the individual token was examined to determine whether any potential cues fell between the last regular gating point and the predetermined final gating point. If the second segment of interest was a relatively long, steady state segment such as a monophthongal vowel, /s/, etc., this 6-9 ms interval near the middle of the segment was judged not to require an additional gate, and the last regular gate was used as the final gate. However, when the second segment of interest was a flap or a voiceless unaspirated stop, for example, the omission of 6-9 ms could make a large difference, so an additional gate at the predetermined final gating point was added.

The choice of a square wave to cut to instead of white noise or silence was based on a pilot experiment. In a pilot version of this experiment, all of the procedures up to this point were the same, but the stimuli were gated to white noise for 200 ms instead of to a square wave for 500 ms. Some subjects tended to perceive the white noise as a speech sound, namely as /s/ or /f/. The addition of /s/ or /f/ after the stimulus segments would force the subject to use an entirely different cohort of words in choosing their responses. For example, in a late gate of the word "elevator" ([<sup>1</sup>eləveɪrə]), instead of the cohort including all words beginning with /el/, such as "elk, elf, elves, elevator, Elton, eligible"

etc., the cohort would be restricted to "else, Elsa." In many cases, the addition of a voiceless fricative might remove all words from the cohort and make it difficult for the subject to give any answer. Similarly, cutting a signal suddenly to silence may introduce the spurious percept of a /p/ or some other segment (Öhman 1966, Pols and Schouten 1978). This would also alter the cohort.

For the final version of the experiment, a square wave was chosen as a sound which is very unlikely to be perceived as a speech sound. The square wave used had a fundamental frequency of 500 Hz, and sounded like a beep. The speaker for the experiment had a rather low voice, and the high fundamental frequency of the square wave, far out of the speaker's pitch range, was intended to make the square wave even less likely to be perceived as a speech sound. The use of a square wave seems to have been successful in not introducing the spurious percept of any speech sound. Because the speech signal was ramped down while the square wave was ramped up (over a 10 ms interval), there was no click at the gating point.

### 2.1.3 Subjects and procedures during the experiment

The maximum number of gates required for any stimulus word was 14. (There were four such words.) In order to avoid potential hysteresis effects (Ohala and Ohala 1995), or other unknown effects of hearing the same word at multiple gating stages (discussed above in Section 1.4.5), the individual presentation method was used: no subject heard the same word at more than one gating stage. Therefore, 14 separate groups of subjects were necessary. To make data for different stimuli comparable, particularly when evaluating the data based on the number of different responses given by the entire group of subjects to hear a given stimulus, the same number of subjects must hear each stimulus. Thus, the number of subjects had to be a multiple of 14.

The stimuli were divided among the 14 groups as follows: for the 14-gate words, one gating stage was assigned to each of the 14 groups in random order. For words with

fewer gates, gating stages were randomly assigned to groups such that each gating stage of a given word was assigned to a different group from each other stage of that word (that is, no group received more than one stage of the same word), and each group received the same total number of stimuli (each from a different word). The total number of stimuli for the entire experiment resulting from the gating process described in the previous section was 868, which is coincidentally a multiple of 14 ( $62 \times 14 = 868$ ). Therefore, each group received 62 stimuli, representing 62 different words. Because of the random assignment of gating stages to groups, no particular group heard only initial gating stages or only final gating stages, for example. If each group had heard stimuli only of a particular stage, one group might have heard a large proportion of the stimuli which were gated out at the hypothesized critical point (the point of maximal spectral change). If that group happened to include a few subjects with special characteristics (unusually attentive or inattentive subjects, for example), this could have affected the results for the entire experiment. Therefore, the random distribution of gating stages to groups was considered preferable. In addition, although a given group of subjects all heard the same 62 stimuli, a different random order of the 62 stimuli was used for each subject.

Almost 200 subjects were recruited from undergraduate level linguistics classes and from undergraduate and graduate student members of the linguistics department. The proportion of subjects who were themselves linguists was not large (less than 20 subjects who were graduate students in linguistics beyond their first semester or undergraduate linguistics majors). The majority of the subjects were recruited from Linguistics 55, Languages of America, because students in this course (introductory level) were offered extra credit in the course as a reward for participating in the experiment. Approximately 110 students from this course served as subjects. Students in Linguistics 100, who have had somewhat more linguistics training, were also offered extra credit. All volunteers were allowed to participate in the experiment, in order to make the extra credit available to all students, but the data of some subjects was excluded for reasons which will be discussed

below. All subjects received candy as a reward for participation, regardless of whether they also received extra credit points or not.

I tested each subject individually in a double walled sound booth. A computer monitor inside the booth displayed a button labeled "next" in addition to buttons for use in beginning and ending the experiment (Figure 2.2). Subjects were given an answer sheet with blanks numbered 1 through 62, and were assigned to one of the 14 groups. I instructed the subjects that they would hear parts of 62 words, and that in each case, they would hear the beginning of a word followed by a beep. They were instructed that when they had heard a part of a word like this, they should "write down what you think the word might be (what word you heard or what word you think you might have heard the beginning of) in the appropriate blank on the answer sheet." They were specifically told that they might hear something which was a whole word, in which case they could write either that word or a longer word beginning with it, but that they would usually hear something which was not a whole word by itself, and that they should respond with an entire word. They were asked to give the response they thought of even if they were not sure how to spell it, and to give a whole word response even if the part of a word they heard was so short that they were not sure what they heard.

I then orally produced and discussed several examples. I pointed out that there is no single correct answer, and gave the example /br-/ , which subjects were told could be "brown, breath, broken, braked" etc. I then pointed out that responses could also be longer words, exemplified by /disp-/ and the possible responses "disposed, dispense, dispel." Next, subjects were warned that they should not be concerned with whether their responses were spelled like they sound, with the example /nɑ<sup>1</sup>-/ for the responses "night, knife." Finally, subjects were told that proper names and place names were also acceptable responses, as exemplified by /o<sup>w</sup>k-/ and the response "Oklahoma" along with "oak, O.K."

I then instructed subjects in the use of the Supercard™ program which was used to present

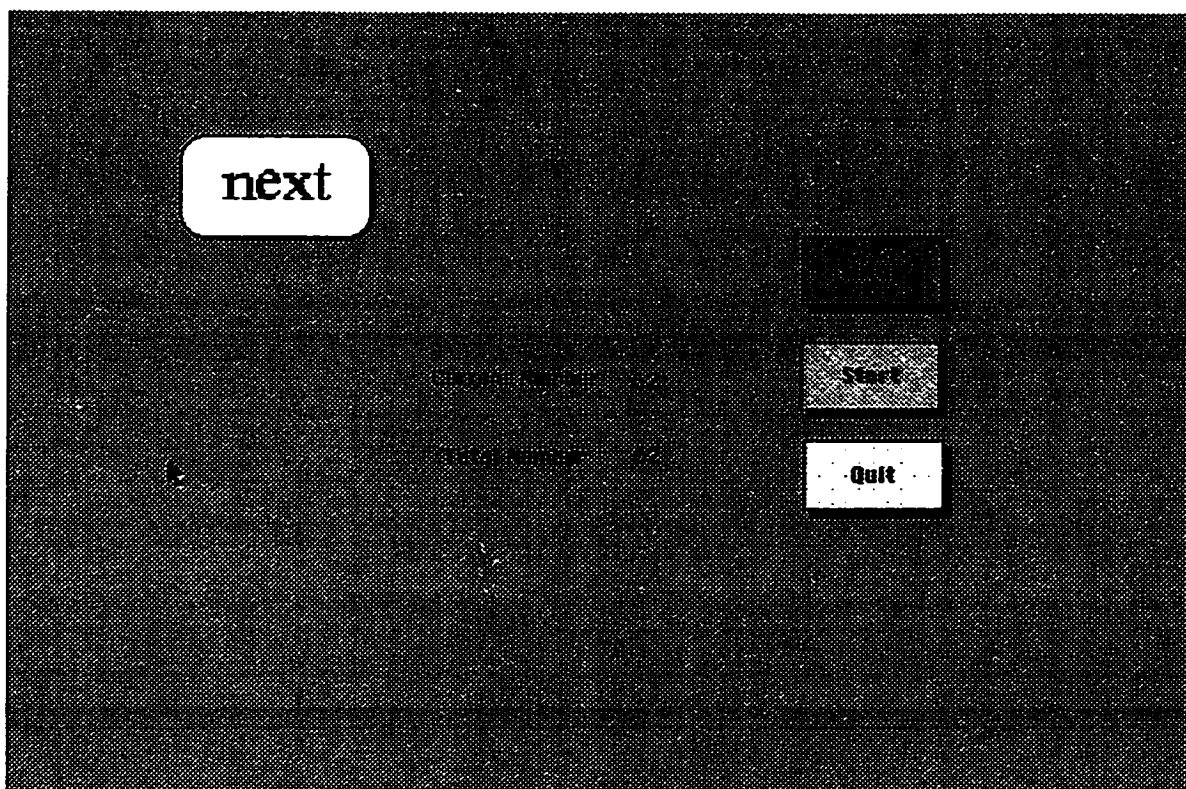


Figure 2.2. The screen subjects for the English experiment used. Subjects controlled the timing of presentation of stimuli by clicking on the "Next" button. Responses were written on paper, not entered through this screen.

the stimuli, and reminded them to be sure that they were filling in the line with the correct number for each response and to give whole word answers for every stimulus. They were told the test was on English words.

The subjects then took a seven item practice test while I was present in the sound booth. Stimuli for the practice test were created in the same way as the stimuli for the real test, except that the speaker was the experimenter, and all subjects heard the practice stimuli gated to the same point. The first four practice stimuli were the same examples discussed in the paragraph above. "Brown" [braʷn] was gated out at a point early in the voicing, to avoid including a perceptible portion of the vowel. Partially because this was the first word of the practice test, and subjects were not yet used to hearing the appended square wave, many subjects failed to perceive any sound other than the beep of the square wave for this stimulus. In such cases, they were reminded that some stimuli were quite short, and told what this stimulus had been, but reminded that in the real test, they should always give some whole word response, no matter how unsure they were. "Dispose" was gated out after the burst of the /p/, but some subjects gave responses with a different place of articulation for the stop. They were not corrected in such cases. "Knife" was gated out at a point late in the diphthong, and "Oklahoma" was gated out after the burst of the /k/. Subjects rarely had difficulty with these words.

Three more practice stimuli, which subjects had not heard before, followed: "black" ([b<sup>l</sup>æk]), "remodel" ([ri'ma<sup>r</sup>l]), and "boxed" ([b<sup>l</sup>akst]). Each was gated out at a point far enough into the second of the segments shown here in bold that the experimenter judged that segment to be perceptible. Subjects rarely had any difficulty with "black," producing responses beginning with /b/. For "remodel," the /a/, because it had been cut off, was not long enough for some subjects to perceive it as /a/, and they perceived it as /ʌ/ instead (Lang and Ohala 1996). Some then said they could not think of any word beginning with /rimʌ/, in which case they were told that in the real test, they should give a

whole word response even if they felt it was not completely correct. Other subjects responded to this test stimulus with words with other vowels, such as "remove." They were not corrected. Almost all subjects responded to "boxed" with "box," as the stimulus itself formed a common, morphologically simple word. They were told that "box" as well as "boxed, boxer, boxing" etc. were equally acceptable. I then left the sound booth and closed the doors. Subjects were instructed to wait until the doors were closed before beginning the test.

After entering an assigned name ("Subject A," etc.) which I had given them, the subjects clicked an "OK" button, and the computer played the first stimulus over headphones. After writing their answer, subjects clicked the "next" button, which caused the next stimulus to be played. Thus, subjects controlled the rate of stimulus presentation. They could not, however, repeat a stimulus. There was no interaction with the experimenter during the test unless there was a problem with the computer presenting the stimuli, which was rare.

At the end of the test, subjects gave the experimenter their answer sheet. The experimenter checked the responses to make sure all were legible, and asked the subjects for their pronunciation of any homographic heterophones ("lead" [lid, led]), words with more than one possible pronunciation ("economics" [ikənamiks, ekənamiks]), borrowings which can be pronounced with varying degrees of Anglicization ("Diablo" [diablow, dæjæblo]), or extremely low frequency words. Low frequency words were checked in case the subject was not familiar with the spoken word (words learned only orthographically) and used a variant pronunciation. The experimenter also confirmed subjects' intended answers for any words which were misspelled, as well as for technical terms or proper names not known to the experimenter. This process had to be done quickly before the subject left, and in some cases, the experimenter failed to notice



homographs or words with more than one possible pronunciation. The treatment of such cases will be discussed below, with the results.

At the end of the experiment, I also briefly interviewed subjects about their language background. They were asked whether English was their native language, and whether it was the first language they learned as small children. If it was not, they were asked what their first language had been, from what age they had started learning English, whether that had been in a classroom situation or in an English speaking country, and which language they were most comfortable with now. All subjects were asked whether they had learned any other languages (besides English and their first language, if not English) as children. Based on this information, subjects were classified as native English speakers if English was their only language or if they said that English was now their dominant language and I could not detect any non-native accent. They were classified as balanced bilinguals if they said that they were now equally comfortable in both English and their other language and I could detect no non-native accent in English. If subjects said that they were better at their other language than at English, or if I detected a non-native accent, subjects were classified as non-native speakers. Information about dialect, when substantially different from California English (British, Boston, etc.) was also recorded. Because of the multicultural composition of California, and of the Berkeley campus in particular, the majority of the subjects had spoken at least one language other than English as children. However, a large proportion were now balanced bilinguals, English dominant bilinguals, or English dominant with only passive knowledge of the other language. Detailed information on subjects' language backgrounds is included in Appendix A.

The data of some subjects was excluded. All of the subjects who were classified as non-native English speakers were excluded, but balanced bilinguals and English dominant bilinguals were not excluded. Native speakers of English who spoke a dialect different from California English were not excluded, since all subjects were living in California, and were assumed to be able to perceive the speech of this area (and of the Coloradan speaker

for the experiment) adequately. However, one British speaker stated that she has difficulty understanding many words in Californians' speech, so her data was excluded. In addition, a few subjects who said they were native speakers of Singapore English, and one who said English as learned in Ethiopia was his dominant language, were excluded. A few subjects stated that they had known hearing losses, and one that she sometimes had speech perception difficulties due to a stroke. All of these subjects were excluded. A few subjects were excluded because they failed to follow directions on the experiment, and did not give whole word answers. For example, some wrote "sss" or "fff" instead of a word beginning with /s/ or /f/. Others simply left some spaces blank despite the instructions to give whole word responses to all stimuli. Subjects who gave incomplete word responses or no response at all for three or more stimuli were excluded. (The treatment of the few remaining missing responses will be discussed in the results section.) Finally, since only multiples of 14 subjects could be used, and there were at least 11 remaining subjects in each group, but 12 in a few groups, from the groups with 12 subjects, the bilingual subject who had acquired English at the latest age was excluded. This left 154 subjects, 11 in each of the 14 groups, whose data will be considered below.

## 2.2. The Japanese experiment

### 2.2.1. Differences from the English methods

As much as possible, I followed the same principles in designing the Japanese experiment as were for English. Words were chosen to represent a particular two segment sequence in the same way, although the phonemic inventory and phonotactic constraints of Japanese, of course, required modifications to the list of transitions investigated. As in the English experiment, words were chosen to represent CV, VC, VV, and CC transitions (although the possible CC transitions are highly limited). Several CV and VC transitions were placed in both first and second syllables. CC transitions in Japanese are limited to the moraic nasal followed by another consonant, geminate obstruents, and some consonants

followed by /y/<sup>9</sup>. NC clusters and geminates were placed only at the boundary of the first and second syllables, but Cy clusters could be placed in either the first or second syllable. Devoicing of vowels can also produce at least phonetic clusters of voiceless obstruents, and these transitions were also used in the experiment, always in the first syllable.

The location of the pitch accent is not expected to affect segment perception as much as stress is expected to affect English segment perception. Fujimura et al. (1978) showed that while both English and Japanese listeners make more use of CV than of VC transitions, English listeners pay more attention to VC transitions when the vowel is stressed than when it is not, while Japanese listeners do not alter the relative use of CV and VC transitions when accent placement is varied. However, I did manipulate accent placement in several CV and VC Japanese stimuli for the current experiment. The pitch accent manipulation was that a transition could appear either before any accent in the word (that is, before the accent of an accented word or in a word with no lexical pitch accent), or in the mora after the accented mora. In the Tokyo Japanese pitch accent system, there is a non-distinctive rise in pitch during the first mora, after which pitch remains relatively high, falling slightly, until approximately the end of the accented mora. The mora after the accented one has a sharp fall in pitch. If there is no accent in the word, pitch falls gradually from the second mora to the end of the word (Pierrehumbert and Beckman 1988). Therefore, a division parallel to English into transitions in the accented mora as opposed to transitions in any mora other than the accented one would be meaningless. Instead, I chose to manipulate accent placement by putting transitions either before the pitch fall for the accent (pre-accentual) or during it (the mora after the accented one). A third such category would be transitions in moras after the pitch fall for the accent has ended (two or more moras after the accented one), but this category was not used because of the difficulty of finding words which are still ambiguous at such a late point in the word.

---

<sup>9</sup> Japanese will be phonemically transcribed using the traditional Japanese system, in which /y/ is a palatal glide. However, because there are important allophonic alternations, phonetic transcriptions will be given as well, and these will be in IPA.

As with the English experiment, the segments forming a CV transition were also used for a corresponding VC transition. However, in the case of the transitions [aɸ] and [oç], the corresponding fricatives were not used in CV transitions because they are conditioned by the following vowel, so that correct perception of the fricative determines the following vowel, and gating would not show when the following vowel was perceived. Similarly, I included the transition /aw/, but not /wa/, because /w/ can only appear before the vowel /a/ due to historical changes.

For purposes of choosing the stimulus words, geminates, C-glide sequences, and all VV sequences (including long vowels and possible diphthongs) were treated as consisting of two segments, not one. This is not a theoretical position, but merely a practical decision in construction of stimuli: if such segments are considered as single segments, a transition such as /too/ is extremely long and requires a large number of gates. By using one word for the /to/ transition and another for the perception of /oo/, the number of gates, and thus of subjects, is reduced. This choice involves only the assumption that the perception of the vowel quality of /o/ is the same in /to/ and in /too/, or that the perception of the place, manner and voicing of /k/ is the same in /ak/ and in /akk/. Although the experiment was not designed to test this, the results do support this assumption, as the quality of a vowel or the place, manner, and voicing of a consonant appear to be perceived well before the length of a long segment is perceived.

### 2.2.2. The word list

Based on these criteria, 76 Japanese words, shown in (2), were chosen. The machine searchable dictionary JDIC (Breen 1996) was used to generate a list of words with appropriate segmental structure for each of the target transitions. For each transition, the words of this list were then evaluated for change in cohort size between the first and second segment of interest, considering both segmental information and accent placement (for which I used the NHK (1985) accent dictionary, since the JDIC dictionary does not mark

pitch accent). The word list was checked by a native speaker of Japanese who is also a linguist to make sure all words are commonly known. As with the English word list above, the two segment sequence to be gated is shown in bold.

## (2) Japanese word list

## Stop-vowel and vowel-stop transitions

<i>/to/</i>			
<i>/todana/</i>	<b>[todana]</b>	'closet'	First syllable, pre-accent
<i>/tatoe'ru/</i>	<b>[tatoeru]</b>	'compare to'	Second syllable, pre-accent
<i>/ka'to/</i>	<b>[kato]</b>	'crossing'	Second syllable, falling
<i>/ka/</i>			
<i>/kakari'iN/</i>	<b>[kakariin]</b>	'official' (n.)	First syllable, pre-accent
<i>/hakama'/</i>	<b>[hakama]</b>	'pleated skirt'	Second syllable, pre-accent
<i>/sya'kai/</i>	<b>[jakai]</b>	'society'	Second syllable, falling
<i>/da/</i>			
<i>/dama'ru/</i>	<b>[damaru]</b>	'be quiet'	First syllable, pre-accent
<i>/midare'ru/</i>	<b>[midareru]</b>	'be disheveled'	Second syllable, pre-accent
<i>/ku'da/</i>	<b>[kuda]</b>	'pipe'	Second syllable, falling
<i>/o/</i>			
<i>/hotoke'/</i>	<b>[hotoke]</b>	'Buddha'	First syllable, pre-accent
<i>/himoto'/</i>	<b>[çimoto]</b>	'origin of a fire'	Second syllable, pre-accent
<i>/ak/</i>			
<i>/hakobu/</i>	<b>[hakobu]</b>	'carry'	First syllable, pre-accent
<i>/hatake/</i>	<b>[hatake]</b>	'field'	Second syllable, pre-accent
<i>/ha'yaku/</i>	<b>[hajaku]</b>	'quickly'	Second syllable, falling
<i>/ad/</i>			
<i>/kadai/</i>	<b>[kadai]</b>	'theme'	First syllable, pre-accent
<i>/hanada'yori/</i>	<b>[hanadajori]</b>	'flower greetings'	Second syllable, pre-accent
<i>/ka'nada/</i>	<b>[kanada]</b>	'Canada'	Second syllable, falling

## Nasal-vowel and vowel-nasal transitions

<i>/me/</i>			
<i>/megumi/</i>	<b>[megumi]</b>	'blessing'	First syllable, pre-accent
<i>/tomeru/</i>	<b>[tomeru]</b>	'stop'	Second syllable, pre-accent
<i>/ne/</i>			
<i>/nemui/</i>	<b>[nemui]</b>	'sleepy'	
<i>/em/</i>			
<i>/kemuri/</i>	<b>[kemuri]</b>	'smoke'	First syllable, pre-accent
<i>/tabemo'no/</i>	<b>[tabemono]</b>	'food'	Second syllable, pre-accent
<i>/en/</i>			
<i>/teni'motu/</i>	<b>[tenimotsu]</b>	'carry-on luggage'	

## Fricative-vowel and vowel-fricative transitions

<i>/so/</i>			
<i>/soda'tu/</i>	<b>[sodatsu]</b>	'grow up'	
<i>/za/</i>			
<i>/zabu'toN/</i>	<b>[zabutou]</b>	'cushion'	

/sya/			
	/syabe'ru/	[ʃabe'ru]	'chat'
/ho/			
	/hokeN/	[hokeŋ]	'insurance'
/os/			
	/zyosee/	[d͡ʒosee]	'woman'
/az/			
	/kaza'ri/	[kazari]	'decoration'
/asy/			
	/basyo/	[baʃo]	'place'
/oh/			
	/gohoo/	[gohoo]	'mistaken announcement'
/ah/ ([ϕ])			
	/wahuku/	[waϕuku]	'Japanese clothing'
/oh/ ([ç])			
	/dohyoo/	[doçjoo]	'sumo wrestling ring'

#### Sonorant-vowel and vowel-sonorant transitions

/ra/			
	/harada'tu/	[haradatsu]	'get angry'
/yu/			
	/yubi'/	[jubi]	'finger'
/ar/			
	/kara'i/	[karai]	'hot (spicy)'
/uy/			
	/huyoo/	[ϕujoo]	'unnecessary'
/aw/			
	/mawari/	[mawari]	'surroundings'

#### Affricate-vowel and vowel-affricate transitions

/tya/			
	/tyazuke/	[t͡ʃazuke]	'rice and tea soup'
/zyo/			
	/zyokyo'ozyu/[d͡ʒokjood͡ʒu]		'assistant professor'
/at/ ([t͡ʃ])			
	/mati'/	[mat͡ʃi]	'city'
/oz/ ([d͡ʒ])			
	/tozi'ru/	[tod͡ʒiru]	'shut'

#### Nasal-stop transitions

/Nt/			
	/haNtai/	[hantai]	'opposite'
/Nd/			
	/kaNdoo/	[kandoo]	'impression'
/Nk/			
	/teNkiN/	[teŋkiŋ]	'transfer (at work)'

## Nasal-fricative transitions

/Nz/	/kaNzeN/	[kanzeŋ]	'perfect'
/Ns/	/seNsoo/	[sensoo]	'war'

## Nasal-sonorant transitions

/Nr/	/keNritu/	[kenritsu]	'run by the prefecture'
/Ny/	/koNyaku/	[koŋjaku]	'engagement'

## Nasal-affricate transition

/Nty/	/kiNtyoo/	[kintʃoo]	'nervousness'
-------	-----------	-----------	---------------

## Transitions with devoiced or deleted vowels

/suk/	/sukuna'i/	[skunai] <sup>10</sup>	'few'
/sik/	/sikaku/	[ʃ <sup>h</sup> kaku]	'qualification'
/kit/	/kitamuki/	[k <sup>h</sup> ʔtamuki]	'North face'
/kut/	/kokutetu/	[kokoʔtetsu]	'national railroad'

## Consonant-glide transitions

/ky/	/kyaku/	[kjaku]	'guest'	First syllable
	/dakyoo/	[dakjoo]	'compromise'	Second syllable
/hy/	/hyoo/	[ɕjoo]	'list'	
/ry/	/ryokaN/	[rjokaŋ]	'Japanese style inn'	

## Geminates

/tt/	/mottaina'i/	[mottainai]	'waste'
/kk/	/sakka/	[sakka]	'author'
/ss/	/sassoku/	[sassoku]	'immediately'
/ssy/	/hassya/	[haʃʃa]	'departure of a train'
/Nm/	/teNmetu/	[temmetsu]	'flashing'

<sup>10</sup> It is not the purpose of this dissertation to argue that the vowels which have traditionally been called devoiced are actually deleted. This transcription simply reflects the fact that in the productions of these words used for the experiment, there is no evidence of a phonetic devoiced vowel, only of the two surrounding consonants.

/Nn/	/a <sup>N</sup> naizyo/	[annaid̥z̥o]	'information office'
Long vowels			
/oo/	/tootyaku/	[toot̥jaku]	'arrival'
/ee/	/keego/	[keego]	'honorific'
/uu/	/syuukaN/	[juukaŋ]	'custom'
Vowel-vowel transitions			
/ao/	/haori/	[haori]	'type of Japanese clothing'
/ia/	/siatu/	[siatsu]	'acupressure'
/æ/	/kaeri'miti/	[kaerimit̥i]	'the way home'
/ai/	/taiko/	[taiko]	'drum'
/oi/	/koibito/	[koibito]	'boyfriend/girlfriend'
Mora nasal-vowel and vowel-mora nasal transitions			
/Ni/	/teNiN/	[teɕiŋ] <sup>11</sup>	'store employee'
/iN/	/hiN/	[hiŋ]	'goods'
/aN/ ([n])	/maNne'Nhitu/	[mannepçitsu]	'fountain pen'
/eN/ ([m])	/seNmoN/	[semmoŋ]	'specialization'

This word list represents a very wide variety of the possible segment transitions of Japanese. Although Furui's work was a gating study of Japanese, he used only CV and CyV transitions, which were the entire utterance, and therefore always initial. These were also just syllables, not real words. In this experiment, the inclusion of a wide variety of transition types, particularly VC, CC, and VV transitions, the manipulation of position in the word, and the use of real word stimuli provide much more representative and realistic environments for studying segment perception. Thus, even though Furui's study was also

<sup>11</sup> Prevocalic mora nasals are often realized as nasalized vowels in Japanese. The speaker's production of this word had nasalization through most of the word.



on use of dynamic cues in Japanese, the current study is quite different, and is likely to have some differences in the results because of this.

As with the English experiment, word frequency, cohort size, and lexical neighborhood density were not controlled (except to make sure that a native speaker considered all the words familiar). However, words were chosen to have a large change in cohort size from the first to the second of the segments in the transition of interest, and the location of pitch accent was considered in determining cohorts, since Cutler and Otake (to appear) show that Japanese listeners do use accent information in lexical access.

### 2.2.3. Production of stimuli

A male native speaker of Japanese from the Tokyo area was the speaker for the experiment. He read the word list twice, in random order, from a list written in Japanese orthography (mixture of Chinese characters and the two syllabaries). Before making the recording, I checked his pronunciations of the words to make sure he used the same accent placement as had been determined from the accent dictionary (NHK 1985), as there is some individual or dialectal variation in lexical accent placement even among Tokyo dialect speakers. The two productions of the word list were recorded using a DAT recorder in a sound treated booth. (Unlike the English recording, the recording equipment and the experimenter were also present inside the recording booth, but the signal to noise ratio of the recording was still quite good.)

The Japanese stimulus words were digitized, trimmed, and gated in the same way as the English words. However, the initial gating point for word-initial stops was placed immediately after the burst instead of an arbitrary 20 ms after the onset of the word. For geminate stops, the end of the first gate was placed approximately one fourth of the way through the stop closure and the end of the final gate was placed just after the burst. For other geminates (including the moraic nasal followed by another nasal) and long vowels, the end of the first gate was placed one fourth of the way through the geminate, and the end

of the final gate three quarters of the way through the geminate. For moraic nasal to consonant transitions, the end of the first gate was approximately halfway through the moraic nasal. Otherwise, the same procedures were followed as are described above for English.

The longest transitions required ten 20 ms gates, so ten separate groups of subjects were necessary. The gated stimuli were divided among the ten conditions in the same way as for the English experiment, with the different gates of a word always assigned to different conditions. The gating process produced 449 stimuli, so one additional gate was added for the word /keNritu/ 'prefectural' (before what would otherwise have been the shortest stimulus) in order to have an equal number of stimuli (45) in each condition.

#### 2.2.4. Subjects and procedures

A total of approximately 130 subjects were recruited at three universities in Japan. Approximately 85 undergraduate students in English courses at Keio University in Yokohama, approximately 30 undergraduate students at Aichi Shukutoku University near Nagoya, and approximately 15 undergraduate and graduate students at Sophia University in Tokyo participated in the experiment. The students at Keio University participated during class time. Subjects at Aichi Shukutoku University were paid a small amount of money for their participation, but subjects at the other two universities were not paid because the professors arranging their participation felt it would be inappropriate (particularly for the Keio students, because they participated during class time). All subjects received a small souvenir (a Berkeley pen or pencil) and some chocolate, whether they were paid or not.

All subjects were native speakers of Japanese, and only a few had had experience with any language other than Japanese during childhood. Data from one subject was excluded because he had spent most of his childhood in the U.S., but none of the other subjects had spent more than a few years outside of Japan at any age. Obviously, the

Japanese subjects were far more homogeneous in their language background than the English speaking subjects were. The majority of the Japanese subjects were from the Tokyo/Yokohama and Nagoya areas, but some subjects did come from other parts of Japan. No subjects were excluded because of dialect background: dialect differences in Japanese affect primarily the pitch accent system, which is not a major subject of investigation in this experiment. Furthermore, all of the subjects have extensive exposure to the standard dialect through television, and most were living in the Tokyo area at the time of the experiment. Information on subjects' language and dialect backgrounds is included in Appendix A.

The experiment was run in the language labs of Keio and Aichi Shukutoku universities, and in a sound treated room at Sophia University. Subjects were run in groups of as few as three or as many as fourteen subjects. Stimuli, which had been recorded on tapes, were presented over headphones, using either the audio equipment of the universities' language labs or walkman tape players and headphones which I supplied. (In the latter case, higher quality headphones than those which came with the walkman tape player were substituted.) The environment in which the stimuli were presented was not as quiet or as consistent for the Japanese experiment as for the English: the language labs (where most of the subjects participated in the experiment) were not much more sound treated than a normal classroom, the audio equipment used varied, and running subjects in groups also increased the background noise. However, the environment was still relatively quiet, and these changes were necessary in order to run large numbers of subjects in Japan.

Subjects were given an answer sheet with 45 numbered blanks, and an instruction sheet, which had the instructions they would hear and four sample answers written out, in Japanese. The instructions, in Japanese, were played over the room speakers for all subjects to hear. This instruction tape had been recorded by the same Japanese native speaker who produced the stimulus words, and the example stimuli and the warning sound used to signal the beginning of a new stimulus were spliced into the instruction tape. The

tape instructed subjects that they would hear the beginning parts of 45 words, and that after each stimulus, they should write a whole word of which the sound they heard might have been the beginning part in the appropriate line of the answer sheet.

Subjects were told that there was more than one possible answer, and an example stimulus /zyootai/ [d͡ʒootai] 'situation,' gated at a point late in the /oo/ (but before the transition to the following /t/) was given, with /zyootai, zyoohiN, zyooo, zyookeN/ suggested as possible answers. A second sample stimulus, /seekatu/ [seekatsu] 'lifestyle,' gated just after the burst of the /k/, was given, with /seekai, seekaku, seekoo, seekatu/ suggested as possible answers. (This second example was used in order to give an example in which the stimulus, as gated, forms a CVC sequence which is not permissible by itself in Japanese.) Subjects were told that foreign borrowings and proper names were also acceptable, which was exemplified with a sample stimulus which had been recorded as /tossa/ [tossa] 'quick,' gated early in the /o/, for which /toosuto/ 'toast' and /tookyoo/ 'Tokyo' were suggested as possible answers<sup>12</sup>. Subjects were told that some stimuli might form whole words by themselves, and that in such cases, they could answer with either that word itself or a longer word beginning with it. The last example stimulus (recorded as /nattoo/ [nattoo] 'fermented soybeans,' and gated during the /a/) had the suggested answers /na/ 'name,' /na/ 'vegetable,' /natu/ 'summer,' /nakama/ 'friends,' and /nattoo/.

This method of presenting sample stimuli has disadvantages as compared to the practice test used for the English experiment. If subjects perceived different segments in the practice stimulus from those used for suggested answers, they might find the task confusing. Furthermore, no extremely short sample stimulus could be presented, which would have demonstrated to subjects the need to give some answer even when unsure, because any suggested answers chosen for an extremely short stimulus would probably

---

<sup>12</sup> Borrowing is a concept very well known to educated speakers of Japanese.

diverge from subjects' perceptions. However, because subjects were run in groups, the supervised individual practice test method used for English was not possible.

Subjects were also instructed in the manner in which the stimuli would be presented, which is discussed below. They were asked to write their responses in the usual Japanese orthography (Chinese characters for words normally written in Chinese characters, syllabary otherwise), and to write the pronunciation above the Chinese characters using the syllabary<sup>13</sup>. Subjects were also told to write in the syllabary if they could not remember the Chinese characters for an answer. The instruction sheet showed one possible answer for each of the four examples written in the suggested manner. Subjects were given a chance to ask questions. I then reminded them that they should always give some Japanese word as an answer, even if the stimulus was so short that they were very unsure of their answer.

The stimuli had been recorded onto tapes with a brief warning sound approximately one second before each stimulus. This warning sound was one of the standard warning sounds of a Macintosh computer, and was highly distinct from both speech and the square wave used for gating. Each stimulus (after the 500 ms square wave) was followed by an eight second pause, in which subjects were to write their answers. Before the warning sounds for the first, eleventh, twenty-first, etc. stimuli, the number 1, 11, 21, 31, or 41 (in Japanese) was inserted, spoken by the same speaker who recorded the stimuli, to help subjects avoid losing track of the correct number on the answer sheet.

After the experiment, subjects were asked to fill in a brief dialect questionnaire. (As the questionnaire was printed on the reverse side of the answer sheet, some subjects began filling it out before the experiment, but they were discouraged from doing so.) Subjects were asked whether they were native speakers of Japanese, what part of the country they were from, until what age they had lived there, and from what age they had lived in the

---

<sup>13</sup> This addition of a syllabary to characters, called "hurigana," is commonly used to show readers how to pronounce names or characters they might not know.

Tokyo or Nagoya area. If they had lived in any other area of the country or outside of Japan for a period of more than a year, they were asked to write these locations and the ages at which they had lived there. Because of the relative homogeneity of the Japanese subjects, it was not necessary to interview them individually about their language backgrounds or current fluency.

More Japanese subjects than English speaking subjects failed to give responses to all stimuli. This difference is probably based in a cultural difference: the Japanese subjects were, on the average, far more worried about performing well or correctly in the experiment than the English subjects were. Data from Japanese subjects who gave no response at all for more than three stimuli were excluded. Since each condition must have an equal number of subjects, and most conditions had twelve subjects, subjects in excess of twelve who had failed to respond to stimuli were excluded. Of the twelve subjects in each of the ten conditions who remained, one subject failed to respond to three stimuli, eight failed to respond to two stimuli, and fourteen failed to respond to one stimulus. This is a much larger number of stimuli with no responses than in the English experiment, but still comprises only six tenths of a percent of the Japanese stimuli, concentrated in the extremely short stimuli. The treatment of such stimuli will be discussed in the results section.

### 3. Degree of spectral change

In order to evaluate the use of dynamic cues in speech perception in a wide variety of environments, one must be able to determine whether a particular portion of a speech signal is changing or not, and how much. Linguists can state in a general fashion that some speech sounds or parts of speech sounds involve a change, such as the transitions of formants into or out of a consonant or the rapid amplitude change at the burst of a stop or affricate. Sounds such as diphthongs are inherently changing, regardless of environment. However, it is difficult to determine a priori criteria that would state whether formant transitions, the burst of an affricate, or the change from burst noise to aspiration noise in an aspirated stop constitute a greater degree of change, for example. Therefore, it is desirable to have a measure of degree of spectral change which can be applied to any speech signal in an objective way. I know of only one such measure which has been published, that developed by Furui (1986).

In this chapter, I will explain the measure  $D$ , discuss how I located the points of maximal spectral change using  $D$ , and explain the problems with this measure of degree of spectral change. Since this is a relatively unknown quantity to measure in phonetic work, and most readers will not be familiar with the way  $D$  responds to various acoustic events, I will also use this chapter to present enough examples of the response of  $D$  to various speech signals for the reader to gain a sense of how this measure of degree of spectral change works. Furui (1986) included only one example of the response of  $D$  to a syllable, which is not enough for future researchers to understand the characteristics of this measure. In this chapter, I hope to provide enough examples and discussion of  $D$  to make it more available for future research.

### 3.1. Furui's measure of degree of spectral change D

#### 3.1.1. Calculation of the measure D

The measure D, defined by Furui, is based on the amount of change in cepstral coefficients over a time window. Cepstral coefficients are the inverse Fourier transformation of the log-spectrum, in a way, the spectrum of the spectrum. Since they are calculated from the spectrum of a signal, they represent values for a short window (Furui uses 30 ms) surrounding a point in time, and are not a time-varying measure themselves. For a given point in time, more than one cepstral coefficient exists, and they are referred to as the first, second, third, etc. order cepstral coefficients. The various order cepstral coefficients for a point in time provide information about various components of the signal at that time, such as the overall amplitude of the signal (the 0th order cepstral coefficient), how far apart from each other the formant frequencies are, and the spacing of the harmonics of the spectrum (which itself reflects the fundamental frequency).

Furui's measure D involves finding the values of a particular order cepstral coefficient at each of several consecutive points in time (frames) in a signal. The slope of these cepstral values over time is found, and the average slopes over time for the first through tenth cepstral coefficients (containing information about formants and fundamental frequency) is then calculated. This gives a measure of the average degree of change for the parts of the spectrum over a certain length of time.

The equations Furui (1986:1020) supplies for this measure are shown in (1)<sup>1</sup> and (2). The quantity  $a_i$  in equation (1) is a necessary intermediate step for calculating the degree of spectral transition D.

---

<sup>1</sup> This equation, as printed in Furui's (1986) article, actually says to sum the cepstral coefficient times  $n$  only from frame  $n=n_0$  to frame  $n_0$ , that is,  $\sum_{n=n_0}^{n_0}$  instead of  $\sum_{n=-n_0}^{n_0}$ . However, this is clearly a typographical error, as there is no reason to sum over just one value of  $n$ , namely  $n_0$ . The equation should be as written here, with the addition of a negative sign before the  $n_0$  in the numerator of the equation, parallel to the denominator of the equation, so that the cepstral coefficients are summed over all the values of  $n$ .



$$(1) a_i = \frac{\sum_{n=-n_o}^{n_o} C_i(n) \cdot n}{\sum_{n=-n_o}^{n_o} n^2}$$

$C_i(n)$  is the  $i$ -th order cepstral coefficient at the  $n$ -th frame (where  $i$  can vary between 1 and some maximum, which Furui sets to 10). Furui provides a figure of a hypothetical example which is helpful in understanding the function of this equation. A simplified version of this figure is shown in Figure 3.1 (Furui 1986:1021).

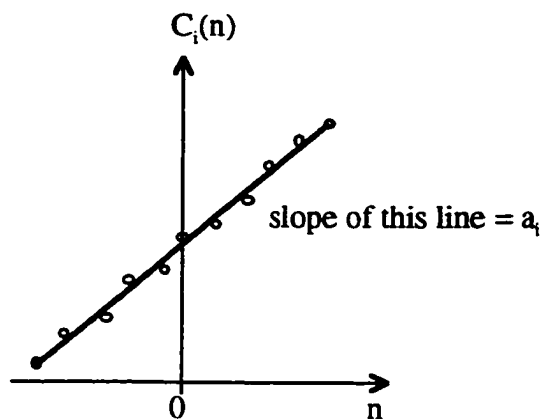


Figure 3.1. Open dots are the values of the  $i$ -th cepstral coefficient at each frame number  $n$ .  $a_i$  is the slope of the regression line fitted to those values.

This is a graph of the values for the  $i$ -th cepstral coefficient (any particular one) over time, or frame number. (Because this is done for each order cepstral coefficient, this could be the cepstral coefficient representing information about pitch, or about the spacing of the formants, for example.) Frame number is represented as  $n$ . Furui uses a window of 50 ms from which to calculate  $a_i$ , with each frame  $n$  being separated by 5 ms. Thus, there are eleven values of  $n$  for each window over which  $a_i$  is calculated, from 25 ms before the center of the window to 25 ms after the center of the window. The result of the equation in (1),  $a_i$ , is the slope of the regression line fitted to the  $i$ -th order cepstral coefficient over time, for example, the slope of the line fitted to the points in Figure 3.1.

The equation in (2) defines the measure D (Furui 1986:1020).

$$(2) D(t) = \frac{\sum_{i=1}^p a_i^2}{p}$$

This equation squares the slope of the line for each order cepstral coefficient, and then averages the squared slopes.  $p$  is the highest order of cepstral coefficient used, which Furui set to 10. Squaring the slopes means that negative and positive slopes are treated equally, since squaring them makes them all positive. This is necessary because the slope of the  $i$ -th order cepstral coefficient over time is a measure of degree of change over time for that order cepstral coefficient, and whether the change involves an increase or a decrease in the actual values of the cepstral coefficient is irrelevant. Averaging the summed squared cepstral coefficient slopes gives a measure of average change during the 50 ms time window for all the components of the spectrum, since the various order cepstral coefficients each represent a particular component of the spectrum. Note that only the cepstral coefficient slopes from first order to tenth order are included in the averaging. The 0th order cepstral coefficient's slope is not averaged into the measure  $D$ . This is because the 0th order cepstral coefficient represents the overall amplitude of the signal (not changes in amplitude), and overall recording level or speech level is of no interest for calculating degree of spectral change.

The result of the equations for the measure of degree of spectral change  $D$  is a numerical value for each 5 ms of a signal. This measure has no units, as it is an average slope. However, for this experiment, only the location of peaks of the measure  $D$  ( $D_{\max}$  points) is important, not actual values of the measure. That is, it is only used as a relative measure, so numerical values are rarely of interest. In the figures showing the measure  $D$  below, the numerical scale for the measure  $D$  is included, but this is only important for comparing the values of  $D$  at various points.

In sum, these equations provide an objective measure of degree of spectral change, which is based on the average degree of change in individual cepstral values during a 50 ms

window, represented as the slope of a line fitted to those values. These equations were implemented in our lab, setting all parameters to the same values Furui used ( $p=10$ ,  $n_0=5$ , window length for  $a_i=50$  ms, frame advance of  $n=5$ ms, window length for calculation of cepstral coefficients=30 ms). In future research, it might be useful to calculate  $D$  for some group of speech signals with some of these parameters varied, especially the window lengths. However, in the current research, I wished to replicate Furui's methods as closely as possible for the acoustic analysis of degree of spectral change.

Using these equations, I calculated the degree of spectral change ( $D$ ) for all of the recorded words from which stimuli had been generated.  $D$  was calculated from the original recorded productions of the word, not from the gated stimuli. The change from the speech signal to the square wave where the signal is gated out (discussed in section 2.1.2.2. above) would result in a large value of the measure  $D$ . Since listeners knew that the square wave was not part of the word to be perceived, the high value of  $D$  associated with the onset of the square wave should not be considered.

### 3.1.2. Method of locating the $D_{\max}$ point or points

For each stimulus word, I located the maximum point or points of the measure  $D$  (the point of maximal spectral change) within the gated area and recorded the time of these points relative to the beginning point of the sound file. Since all files have approximately 400 ms of silence before the onset of the word, this is easily converted into the time from the beginning of the word.

#### 3.1.2.1. Cases with only one peak of $D$ in the gated area

Figure 3.2 shows an example of a stimulus word, "chapel" (for the  $/t/\text{j}\text{æ}/$  transition) which has only one point of maximal spectral change, or  $D_{\max}$  point, within the gated area, namely at the onset of voicing. One should note that there are other peaks of  $D$  outside the gated area, but these are associated with transitions between segments other than the two

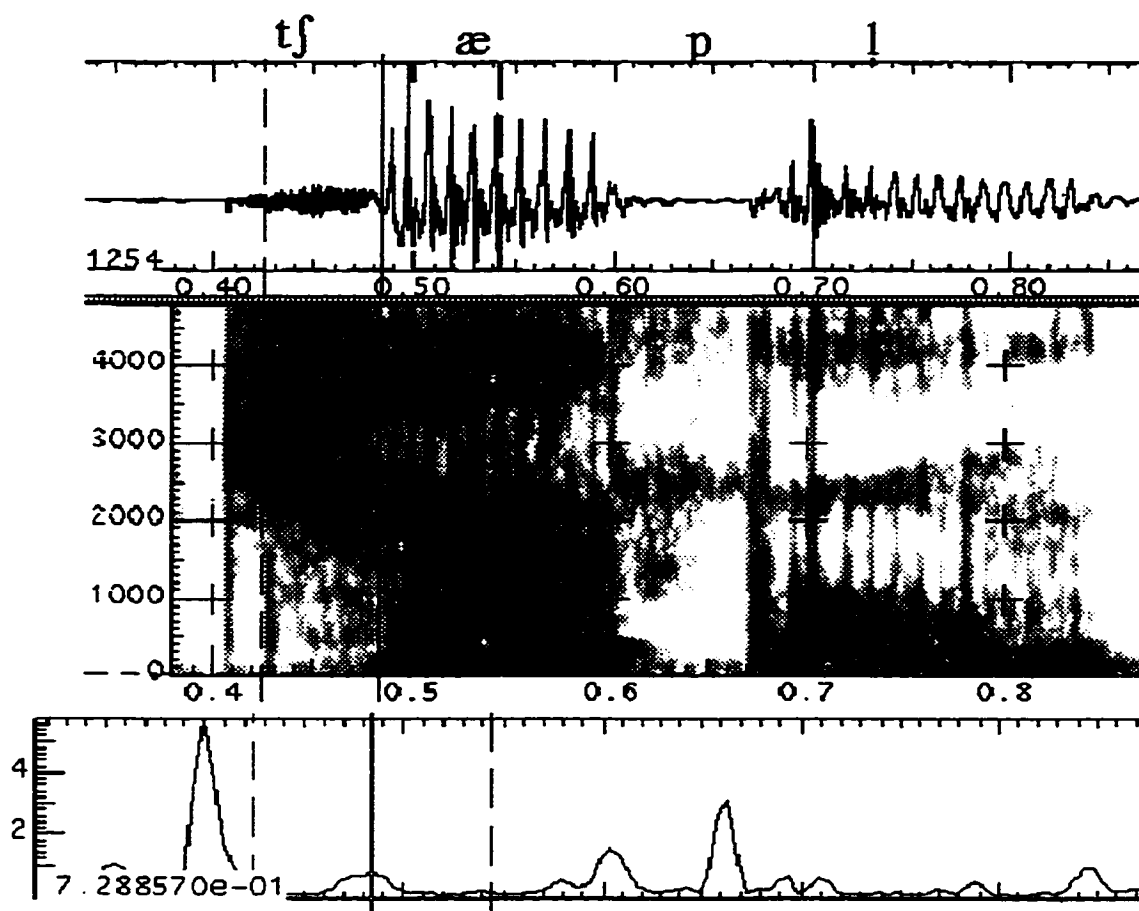


Figure 3.2. Waveform, spectrogram, and measure D for the word "chapel" /tʃæ/. Vertical dashed lines (through all three windows) show the beginning and end of the gated area. Vertical solid line shows the location of the maximum of D within the gated area. Other peaks of the measure D, outside the gated area, are associated with the burst of the affricate (before the gated area), and with the end of the first vowel and the burst of the /p/ (after the gated area), for example.

segment sequence of interest, such as the peaks at the closure for the /p/, the release of the /p/, and the onset of the word initial  $\widehat{[tʃ]}$ . Since these other peaks of D are outside the gated area, so that no listeners hear a signal which ends near that  $D_{\max}$  point, they are irrelevant to the current experiment, even if they have a larger value of D than the  $D_{\max}$  point within the gated area.

### 3.1.2.2. Cases with more than one acoustic change expected in the gated area

One might think that only the point with the greatest D value within the gated area should be considered, since the main purpose is to find out if perception improves most quickly at the point of maximal spectral change. However, there are numerous cases in which there is more than one local maximum of the measure D within the gated area, and measuring only the greatest maximum would omit important information. The gated area (the area between the endpoint of the shortest and the longest gate, covering the two segment transition of interest) often contains more than one change in the signal, and more than one important acoustic feature. For example, in a voiceless stop to vowel sequence, there is one peak of D at the beginning of the burst, another at the onset of voicing, and possibly another at the change from the burst noise quality to the aspiration noise quality, although this third point may be too close in time to the peak at the burst to appear as a separate peak. In a word-medial stop-vowel transition, the shortest gate ends just before the beginning of the burst (Section 2.1.2.2), so all of these points would be included within the gated area. Figure 3.3 shows the word /hakama/ [hakama] 'Japanese style pleated skirt' (for the transition /ka/), which has a large peak of the measure D at the burst of the /k/ and another smaller peak at the onset of voicing of the following /a/.<sup>2</sup>

---

<sup>2</sup> In looking at the figures which illustrate the evaluation of the measure D, it is important to note the scale for the window displaying D for each figure, as the scale is adjusted for each figure to best show the peaks in the measure D for that word. As some peaks are greater than 10, while others are smaller than 0.5, this adjustment of the scale was necessary. The scale is always included in the figure, at the left of the window displaying the measure D.

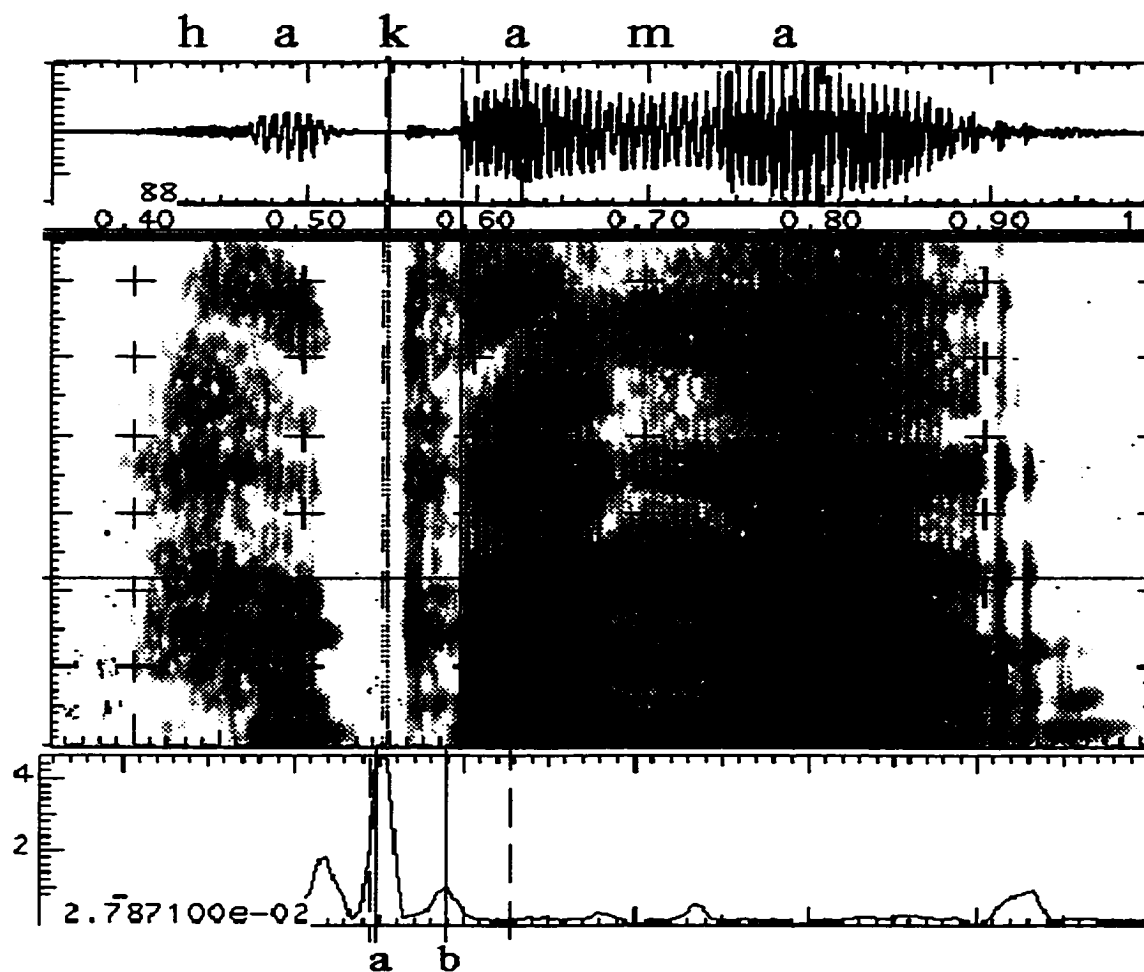


Figure 3.3. Waveform, spectrogram, and D for the word /hakama/ 'Japanese style pleated skirt,' in which the transition of interest is [ka]. Dashed lines show the boundaries of the gated area, and solid lines show the location of the  $D_{\max}$  points. The peak labeled "a," just after the beginning of the gated area, corresponds to the release of the /k/, and the second, smaller peak ("b") corresponds to the onset of voicing of the /a/.

The measure  $D$  will always show a higher value for the beginning of a voiceless stop burst than for the onset of voicing: a sudden change from silence to noise is probably the greatest and fastest change possible in a speech signal. However, considering only the point with the absolute highest value of  $D$  and ignoring the local maximum of  $D$  at the onset of voicing would miss important information. The purpose of comparing salient points in the perceptual data to the  $D_{\max}$  point is to determine whether listeners perceive segments based on dynamic cues. If listeners perceive a vowel after a stop based on the changing information around the onset of voicing, this would be lost by considering only the largest peak of  $D$ , the one at the onset of the burst, as the point of most spectral change (the  $D_{\max}$  point). In general, the part of the signal which provides useful dynamic cues to the listener might, because of the way in which  $D$  is calculated, be only a local maximum of  $D$ , with the absolute maximum of  $D$  located elsewhere.<sup>3</sup> Therefore, both local maxima of  $D$  (the one at the burst and the one at onset of voicing) were counted as  $D_{\max}$  points in such cases, where there is more than one clear acoustic event in the signal. In the same way, in vowel-stop or sonorant-stop transitions, both the peak at the closure of the stop and the peak at the burst were counted as  $D_{\max}$  points.

This procedure of counting both peaks of  $D$  as  $D_{\max}$  points when more than one acoustic change is expected within the gated area, as for transitions involving medial stops, was followed regardless of the relative size of the two peaks. In some cases involving voiceless stops, the peak of  $D$  for the burst is so large as to make the peak associated with the onset of voicing appear negligible. This situation is shown in the word "skull" /kʌ/<sup>4</sup> in Figure 3.4. Since there is a peak (labeled "b"), albeit a very small one, exactly at the onset

---

<sup>3</sup> As a convention, I will refer to any local maximum in the measure  $D$  as a peak of  $D$ , but I will reserve the term " $D_{\max}$  point" for peaks which were counted as maxima for purposes of comparing them to the perceptual data. That is, " $D_{\max}$  point" will mean a peak of  $D$  which was actually recorded for use in the calculations reported in Chapter 4, and not a peak of  $D$  which was excluded for the reasons described in this chapter. Only the  $D_{\max}$  points are reported in Table 3.2 below.

<sup>4</sup> Whenever I mention any word used in the experiment, the two segment transition of interest is either shown in bold, or written after the word, as here, if I felt it unnecessary to transcribe the entire word.

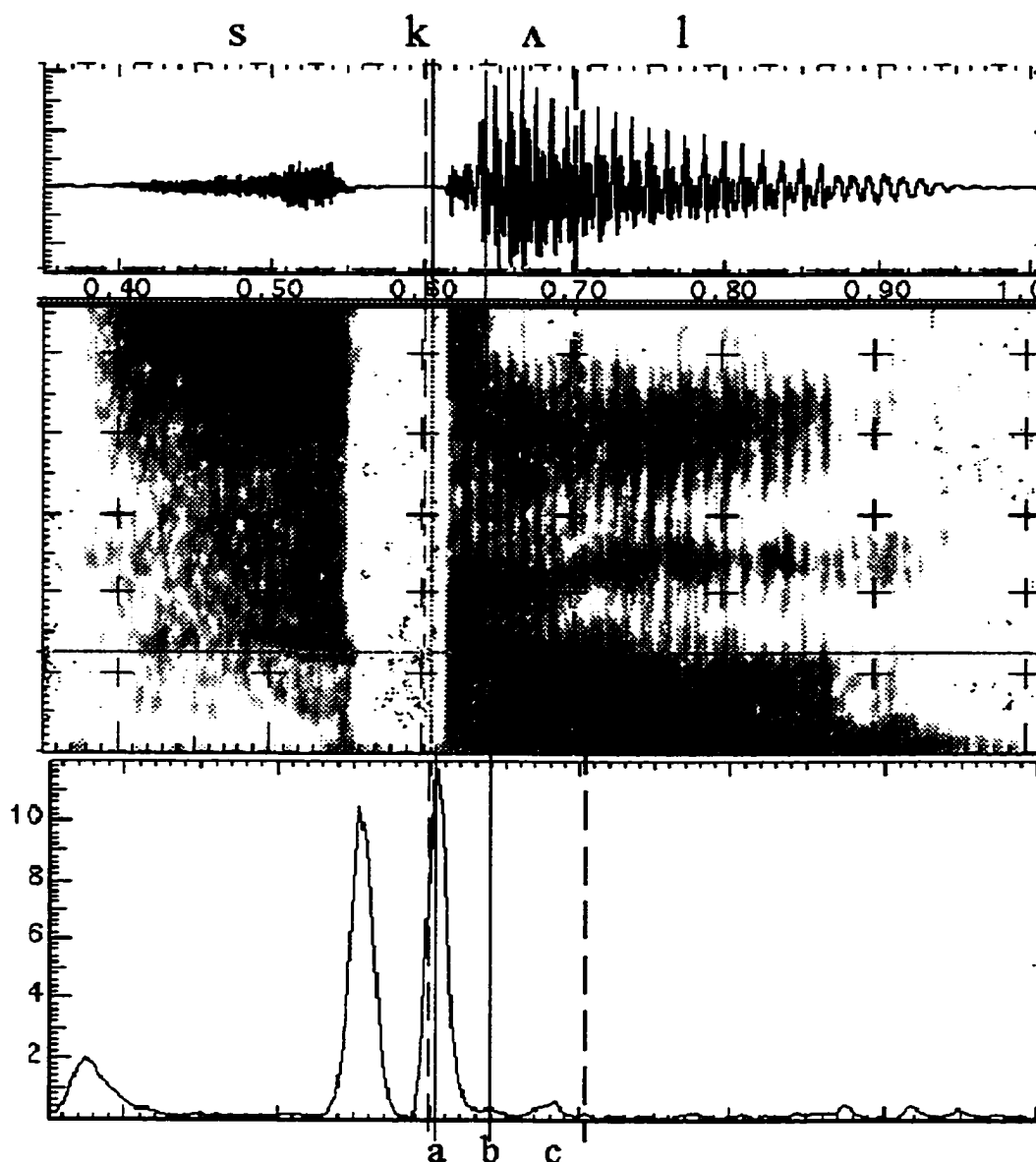


Figure 3.4. Waveform, spectrogram, and  $D$  for the word "skull" /kʌl/. Dashed lines show the borders of the gating area. Two  $D_{\max}$  points were counted, one large one at the burst of /k/ (labeled "a") and one very small one at the onset of voicing of the vowel (labeled "b"). The peak near the end of the gated area ("c"), slightly larger than the one for the onset of voicing, was not counted as a  $D_{\max}$  point, as it is associated with the weakening of the third, fourth, and fifth formants, which probably reflects the transition to the next segment, /l/. (Although it is difficult to see in this figure that the point labeled "b" is a peak, it has a higher value of  $D$  than the points surrounding it, and it is also exactly at the onset of voicing.)



of voicing, both the large peak for the burst ("a") and the small one for the onset of voicing were counted as  $D_{\max}$  points.

For some types of transitions, there are two local maxima of  $D$  within the gated area, each associated with a different acoustic event, but neither peak is consistently larger than the other (unlike the peaks for bursts vs. onset of voicing). For example, there are usually two peaks of  $D$  for a flap, one at the beginning of the flap and one at the end, even though these are quite close to each other in time. Since the gated area was chosen to include most of the segment for such very short segments as flaps, both local maxima of  $D$  usually fall within the gated area. However, either the peak for the onset of the flap or the peak for the release (or end) of the flap can have the larger value of  $D$ . In Figure 3.5, the word "citizen" [ɪr] has a larger value of  $D$  at the release of the flap ("b" in the figure) than at the onset of the flap ("a"). In Figure 3.6, the word "committee" [ɪr] has the larger peak at the onset of the flap ("a," with the peak at release labeled "b"), however. In such cases, considering only the point with the largest value of  $D$  would not only omit important information, but also introduce a source of inconsistency into the data. Therefore, as with burst and onset of voicing, the peaks for both the beginning and end of the flap were counted as  $D_{\max}$  points. When there was more than one  $D_{\max}$  point for a transition, I also noted which  $D_{\max}$  was the largest, but the local maxima associated with each acoustic event were recorded regardless of their relative size.

### 3.1.2.3. Peaks associated with changes other than the transition of interest

In some cases, the gated area contains local maxima of  $D$  which are associated with an acoustic event other than the change from the first to the second segment of interest. This is especially likely for word initial or word final transitions, since the gated area for these extends to the beginning or end of the word, respectively, as discussed above in section 2.1.2.2. For word final transitions, the amplitude of the final segment often trails

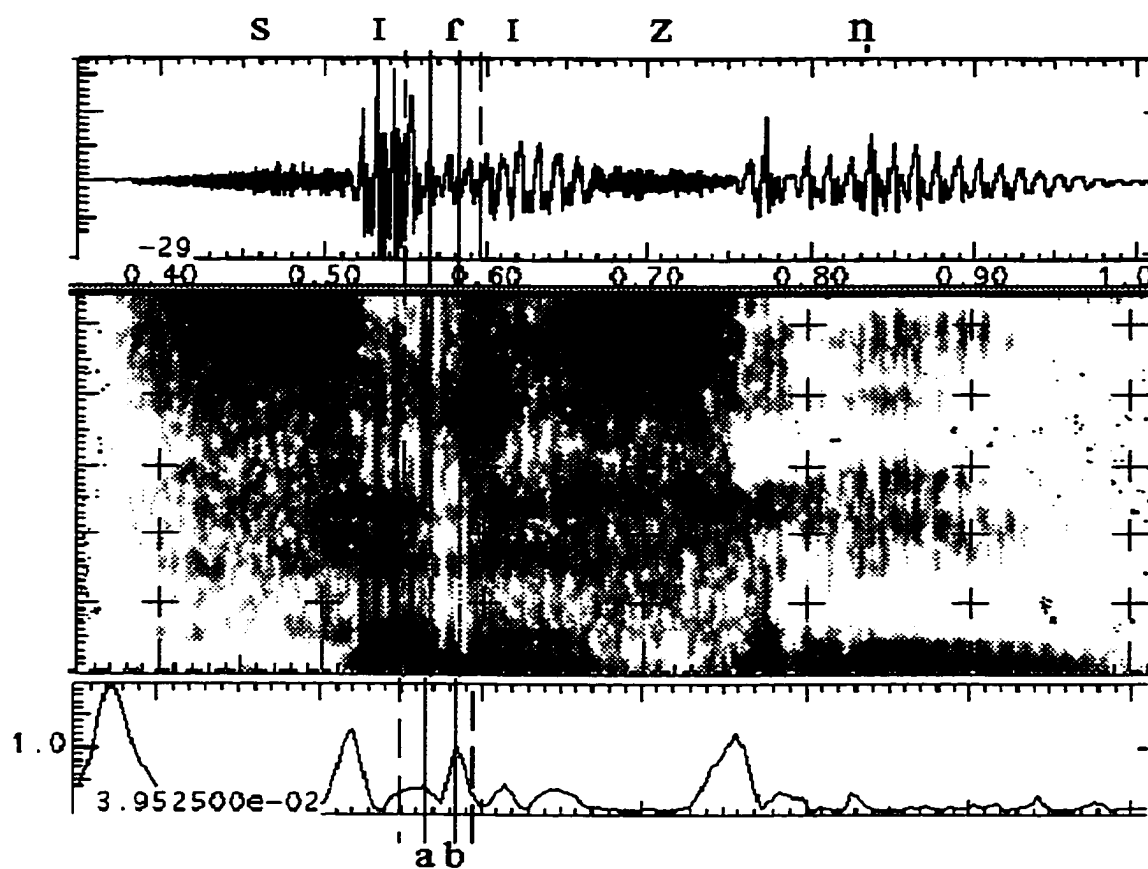


Figure 3.5. Waveform, spectrogram, and D for the word "citizen" [ɪr], in which there are two peaks of D within the gated area, the larger at the "release" of the flap ("b") and the smaller at the onset of the flap ("a"). (Despite the brevity of flaps, they do typically have a very short slightly noisy release, which results in a peak of D separate from the one at the onset of the flap.)

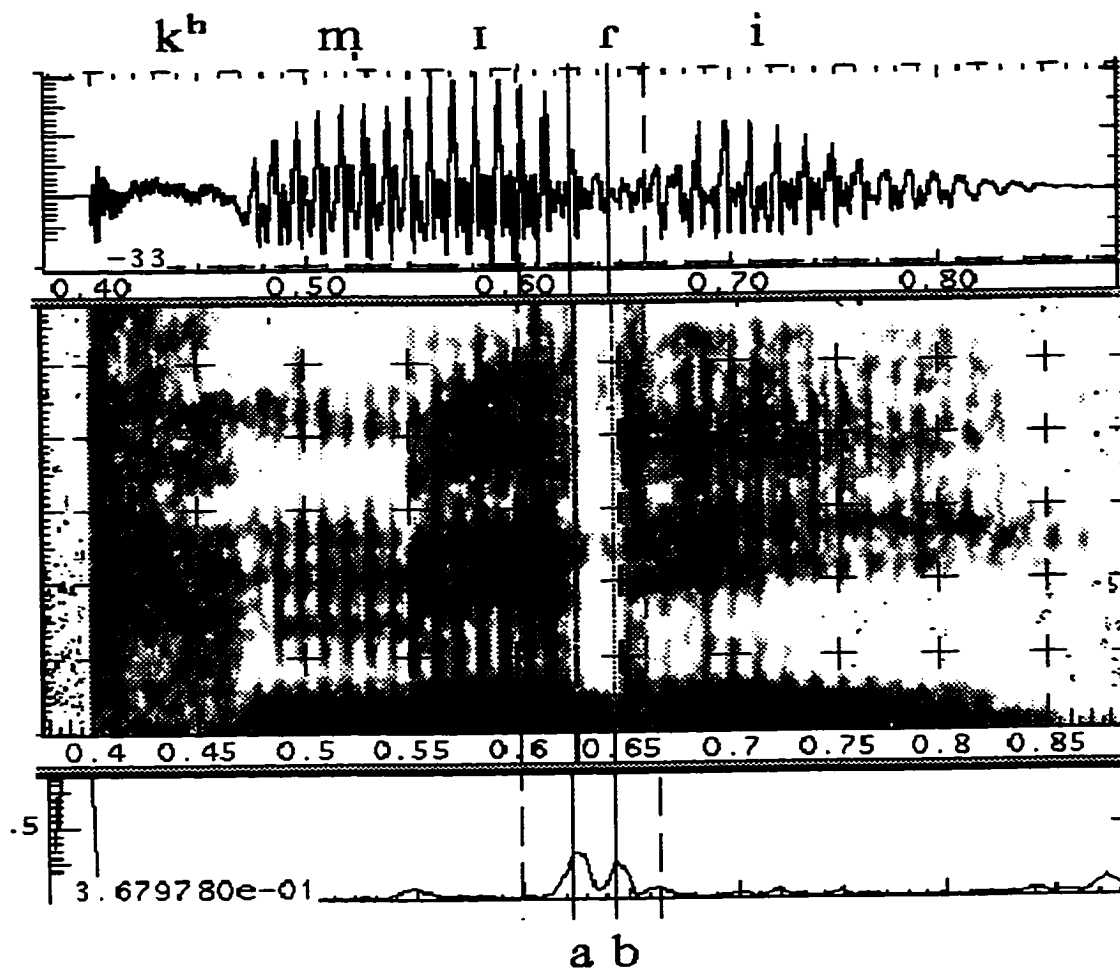


Figure 3.6. Waveform, spectrogram, and D for the word "committee" [ɪr], in which the peak of D for the onset of the flap ("a") is larger than the one for the release of the flap ("b"), the reverse of the previous figure.

off gradually, so the gated area may include a peak of D associated with the offset of the higher formants, for example, or with a reduction in amplitude of a final fricative. Figure 3.7 demonstrates this with the word "fair" /eɪr/, for which the largest peak within the gated area (labeled "e"<sup>5</sup>) is associated with the cessation of the fourth and fifth formants near the end of the word. Two more peaks of D, of similar height, follow, one associated with the cessation of the second and third formants (labeled "f"), and the last associated with the complete end of voicing ("g"). However, the transition of interest in this word is from the vowel to /r/, and it is clear that these peaks are associated with a transition other than the one of interest, namely the one from /r/ to silence. For purposes of comparing the area in which listeners make the most progress toward recognizing the /r/ (the second segment of the transition of interest) to the point of maximal change, one would not wish to compare the perceptual data to the peaks of D associated with the end of the word.

Even for medial transitions, in some cases, a local maximum of D associated with a transition other than the one of interest may appear within the gated area. Figure 3.8 demonstrates this with the word /sakka/ 'author' /kk/. Here, the largest  $D_{\max}$  point ("a") is associated with the release of the geminate /kk/, but a smaller peak at the end of the gated area ("b") is associated with the onset of voicing for the vowel.<sup>6</sup> The sequence of interest is the geminate consonant /kk/, but the onset of voicing of the following vowel is associated with the following transition, namely /ka/. The peak at onset of voicing of the vowel might be expected to be near where listeners make progress toward perceiving the /a/, but the question for this word is where listeners perceive the length of the geminate /kk/. Similarly in Figure 3.4 above, the small peak near the end of the gated area (labeled "c"), larger than the one for the onset of voicing, appears to be associated with the

---

<sup>5</sup> In labeling the peaks of D consecutively, I omitted the letter "d" as a peak label in order to avoid confusion with the measure D itself.

<sup>6</sup> The very highest point of this later peak is actually 5 ms after the end of the gated area, but even if it had been within the gated area, it would be associated with the following transition.

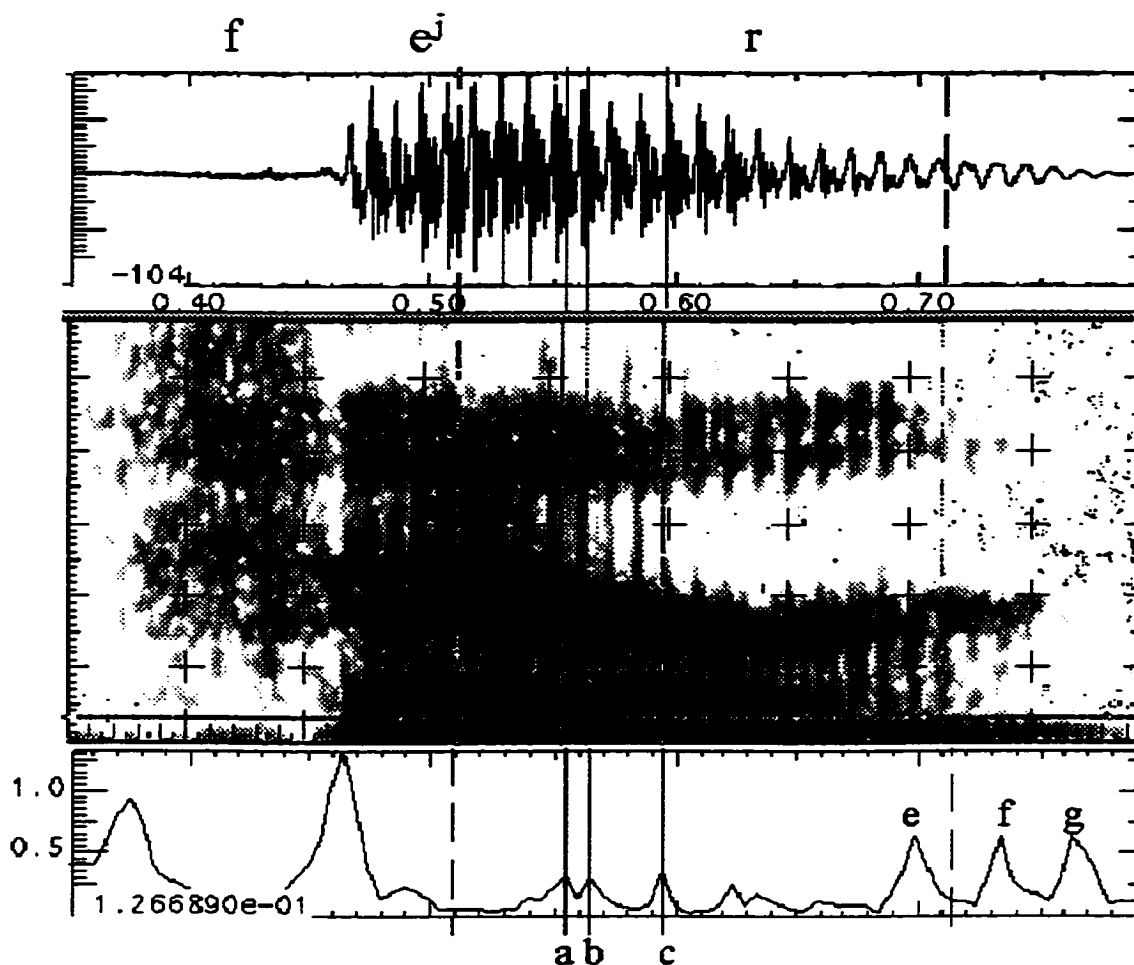


Figure 3.7. Waveform, spectrogram, and  $D$  for the word "fair" /eɪr/. As usual, the dashed lines mark the borders of the gated area, and the solid lines mark the locations chosen as  $D_{\max}$  points. Note also the three relatively large peaks of  $D$  (values greater than 0.5, thus not large, but larger than the small peaks chosen as  $D_{\max}$  points) around the end of the gated area. The first of these three peaks ("e") occurs where the fourth and fifth formants end. The second is at the end of the second and third formants ("f"). The third is at the end of all voicing ("g"). These points are not counted as  $D_{\max}$  points for the transition of interest, even though the first is the largest peak of  $D$  in the gated area, because they are associated with the change to the end of the word, not the change to the /r/.

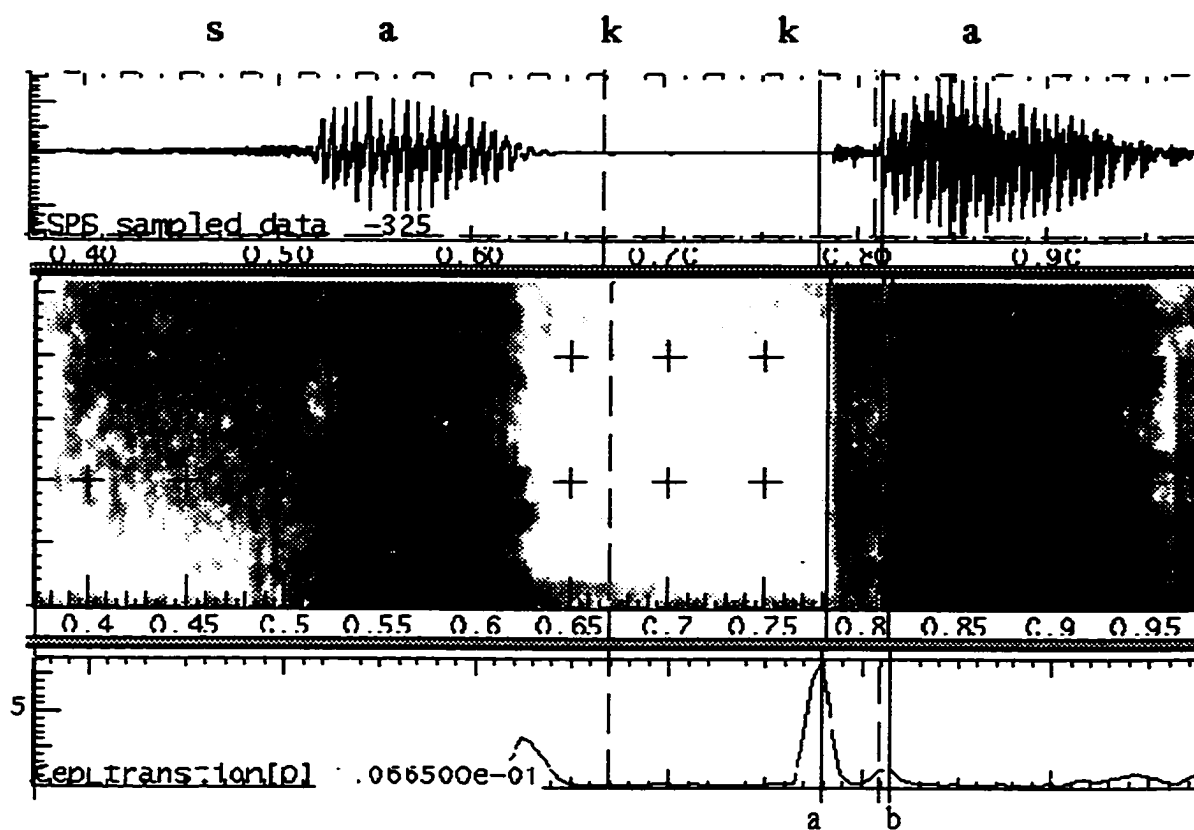


Figure 3.8. Waveform, spectrogram, and D for the word /sakka/ 'author' [kk]. Only the large peak of D associated with the /k/ burst ("a") is counted as a  $D_{\max}$  point. The smaller peak associated with the onset of voicing ("b") is not counted, because it is for the transition to the next segment after the two segments of interest. (The highest point of that peak is actually 5 ms after the end of the gated area, but would not be counted as a  $D_{\max}$  point even if it fell within the gated area.)

weakening of the third and higher formants, which probably reflects the transition to the /l/, a transition other than the one being investigated.

Such local maxima of D, which are associated with a transition other than the one of interest, were not included in the calculations. In general, it is desirable to minimize the number of local peaks of D which are counted as  $D_{\max}$  points for each transition of interest. This is because the primary method of analysis of the results for the experiment is to compare the location of the  $D_{\max}$  point(s) to the location of the area which is most important for perception (discussed in Chapter 4) to determine whether that area occurs near the  $D_{\max}$  point. The larger the number of distinct time points counted as maxima of D, the greater the probability that a particular area, the area most important for perception, will occur at one of the  $D_{\max}$  points by chance. Thus, in order to provide a fair test of the hypothesis that listeners make disproportionate use of areas of the signal with great spectral change, one must balance the need to locate any areas which might have an important amount of spectral change with the need to identify as few points as possible as having maximal spectral change. The issue of the area of most importance for perception falling at a maximum of D by chance will be discussed in section 4.2.2. below. In the attempt to achieve this balance, I chose to count more than one local maximum of D when each was associated with a separate acoustic change of the transition of interest (burst and onset of voicing in a stop vowel transition, for example), but not to include local maxima of D associated with other transitions (as in "fair" and /sakka/ above).

#### 3.1.2.4. Changes in the signal which are not always expected

An additional complication is introduced by cases in which there is more than one local maximum of D, and it is clear what change in the signal caused each peak, but unlike the burst/onset voicing cases, one could not have predicted in advance that these changes in the signal would exist. For example, the cessation of voicing often causes more than one peak in the measure D, particularly if the end of the voicing is creaky. One would normally

expect the cessation of voicing to constitute one change in the signal, but the very long periods of creaky voicing often result in a separate peak of the measure D for each period, as shown for the word "ranch" /ntʃ/ in Figure 3.9 (peaks "a, b, c"). In such cases, only the largest peak of D associated with the end of voicing and within the gated area (in this case "b") is counted, because the existence of separate peaks for each period of creaky voice is probably an artifact of the window length used for calculating D.

One would also expect the transition from a voiceless fricative to a vowel (or nasal) to constitute just one acoustic change in the signal. However, at the end of a fricative, there is often a short period of silence or lessened amplitude of frication just before voicing begins. This is because one must have high oral air pressure in order to produce a fricative such as /s/ (to force air through the close constriction of the fricative), but must have relatively low oral air pressure in order for voicing to begin (in order to have air flow through the vocal cords). (Ohala 1981a, 1983 discusses the same mechanism in VOT of stops.) The measure D frequently reflects two separate peaks in the transition from a voiceless fricative to a following voiced segment because of the separation in time of the reduction in frication and the onset of voicing. Figure 3.10 demonstrates this with the word "snow" /sn/. When the final part of a fricative assimilates to the quality of the following vowel, creating a change in the quality of frication noise during the fricative, two separate peaks of D may also appear. Figure 3.11 shows the word /syaberu/ 'to chat' [ʃa], which has peaks in the measure D both at a point where the frication noise becomes similar to the following /a/ ("a") and at the actual onset of voicing ("b"). In such cases, both peaks of D were counted as  $D_{\max}$  points if the smaller peak was at least 60% as large as the largest peak and it was clear from the waveform or spectrogram what acoustic change the smaller peak was associated with. In the case of the two separate peaks for the change from a fricative to a vowel, the peaks are often of very similar size and are separated in time by several tens of milliseconds, as in "snow" in Figure 3.10. Counting only the largest peak



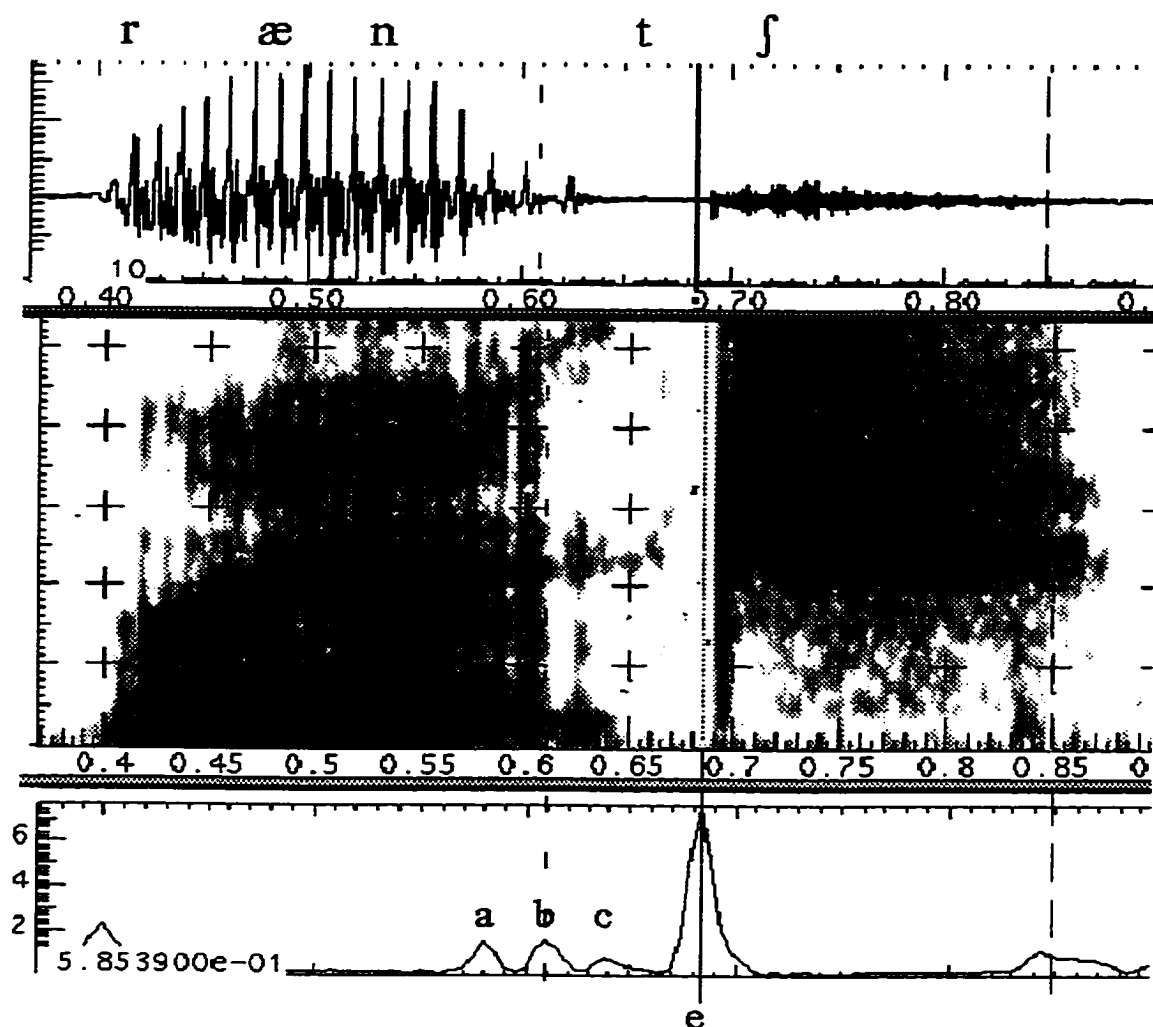


Figure 3.9. Waveform, spectrogram, and D for the word "ranch" /ntʃ/. The dashed lines show the borders of the gated area, the solid line shows one  $D_{\max}$  point ("e"), and another  $D_{\max}$  point ("b") is located exactly at the beginning of the gated area (the left dashed line). Note that there are three separate peaks ("a, b, c") associated with the end of voicing of the /n/ because the final three periods of the /n/ are creaky. Only the largest of these peaks within the gated area was counted as a  $D_{\max}$  point, and that is the one exactly at the beginning of the gated area.

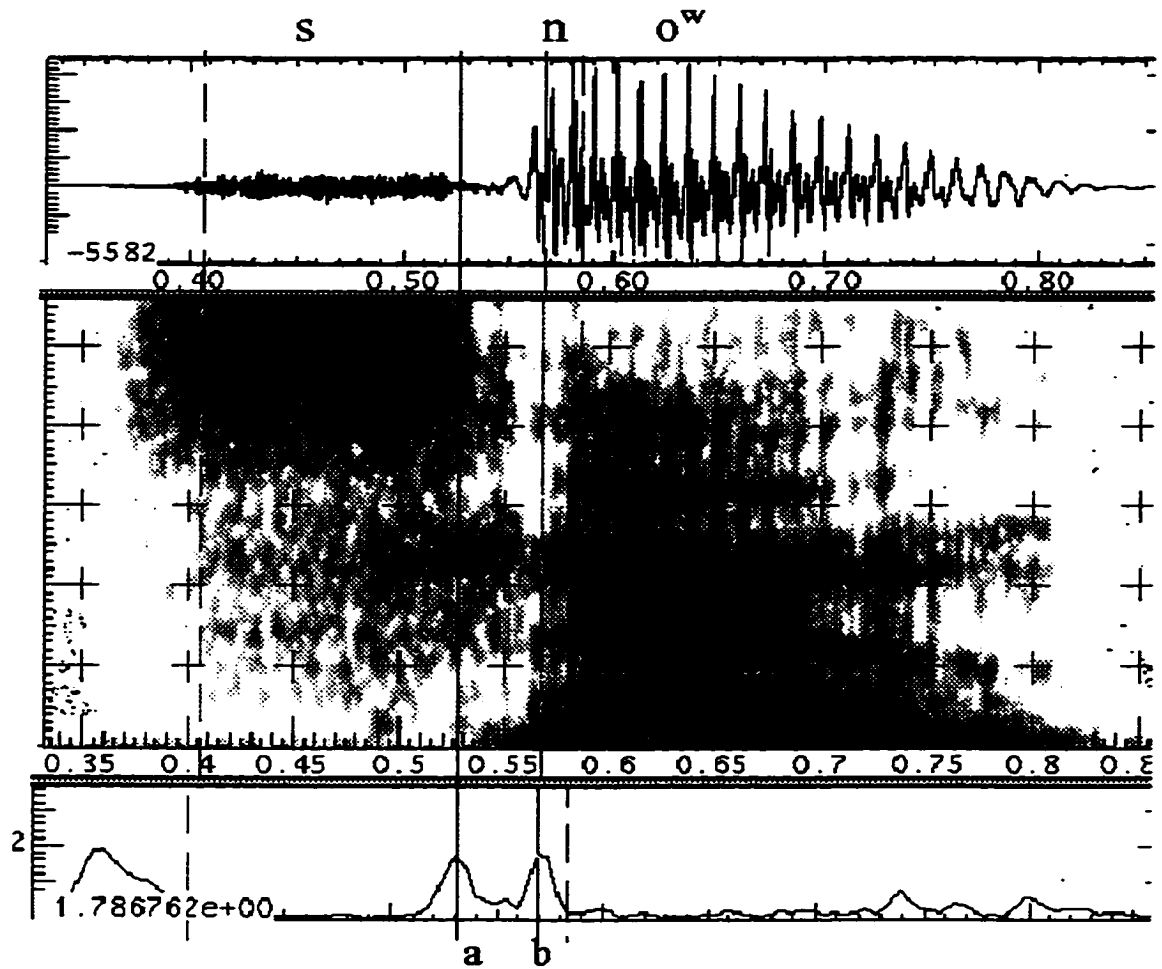


Figure 3.10. Waveform, spectrogram, and D for the word "snow" /sn/. There are two separate peaks in the measure D for the change from /s/ to /n/, the first ("a") at a point where the amplitude of frication noise decreases sharply (visible in the waveform and the higher frequencies of the spectrogram more clearly than in the lower frequencies). The second is at the onset of voicing ("b"). Both peaks are counted as Dmax points because the smaller is more than 60% as large as the larger, and it is clear what acoustic changes both are associated with.

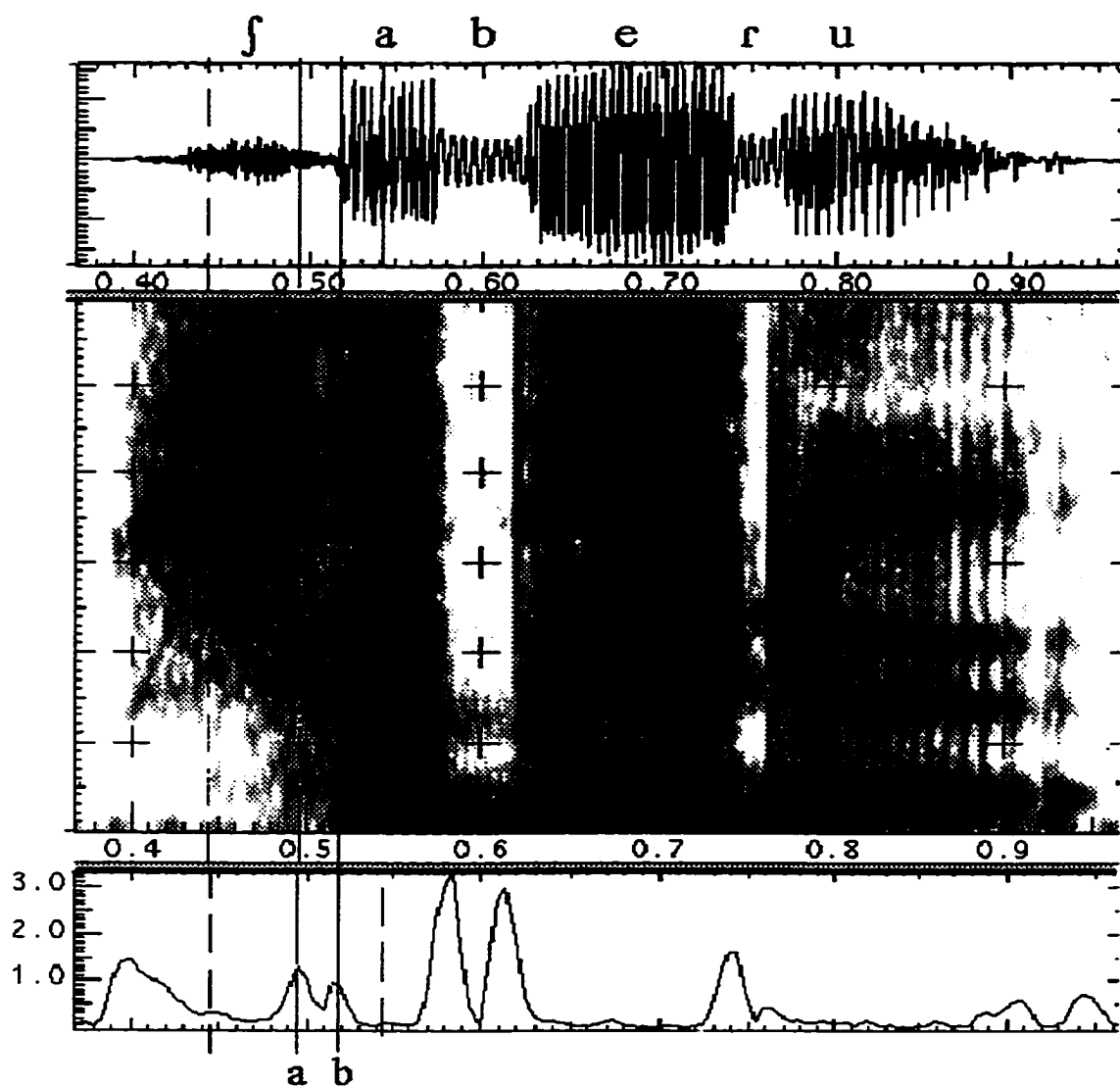


Figure 3.11. Waveform, spectrogram, and D for the word /syaberu/ 'to chat' [ʃa]. There are two peaks of D associated with the change from [ʃ] to [a]. The first ("a"), and larger, is at the point where the fricative loses much of its high frequency energy and gains energy around the frequencies of the formants of the following vowel. The second ("b") is at the onset of voicing of the vowel. Both peaks are counted as  $D_{\max}$  points, as in the previous figure.

as a  $D_{\max}$  point in such cases might lose important information, as one cannot predetermine which change in the signal will be most important for perception.

### 3.1.2.5. Highest value of D not a peak

In some cases, the point within the gated area which has the greatest value for the measure D is not a peak of the measure D, but rather on the skirts of a peak of D which falls outside the gated area. This is likely to happen when the gated area begins shortly after or ends shortly before a stop, if the gated area itself does not have any events which cause a large value of D. Figure 3.12 demonstrates this with the word "caboose" /kəʊ/, in which the largest value of D occurs at or near the end of the gated area (at the point labeled "b"), but this is just part of the rise for the following /b/. This rise begins within the gated area because the unstressed vowel of the gated area is so short. In such cases, the point with the largest value of D within the gated area is not counted as a  $D_{\max}$  point, since it does not represent the point with the most spectral change—the point with maximal spectral change is the actual peak of D, which falls outside the gated area. That is, if listeners are using the part of the signal with maximal spectral change disproportionately for perception, the point just inside the gated area which happens to have the highest D value within the gated area would not be a particularly important point for perception. For such transitions, the peak or peaks within the gated area which were not part of the rise of D toward a peak outside the gated area were used as  $D_{\max}$  points, as shown in Figure 3.12 (the peak at "a").

### 3.1.2.6. Summary of criteria for locating $D_{\max}$ points

Table 3.1 summarizes the criteria for identifying a point as a  $D_{\max}$  for the situations which have been discussed in this section. These are all cases in which one can determine by comparison with the spectrogram or waveform that a peak is associated with a clear change in the acoustic signal. Cases in which the peaks of D are not clearly related to any acoustic change will be discussed in section 3.2.1 below.

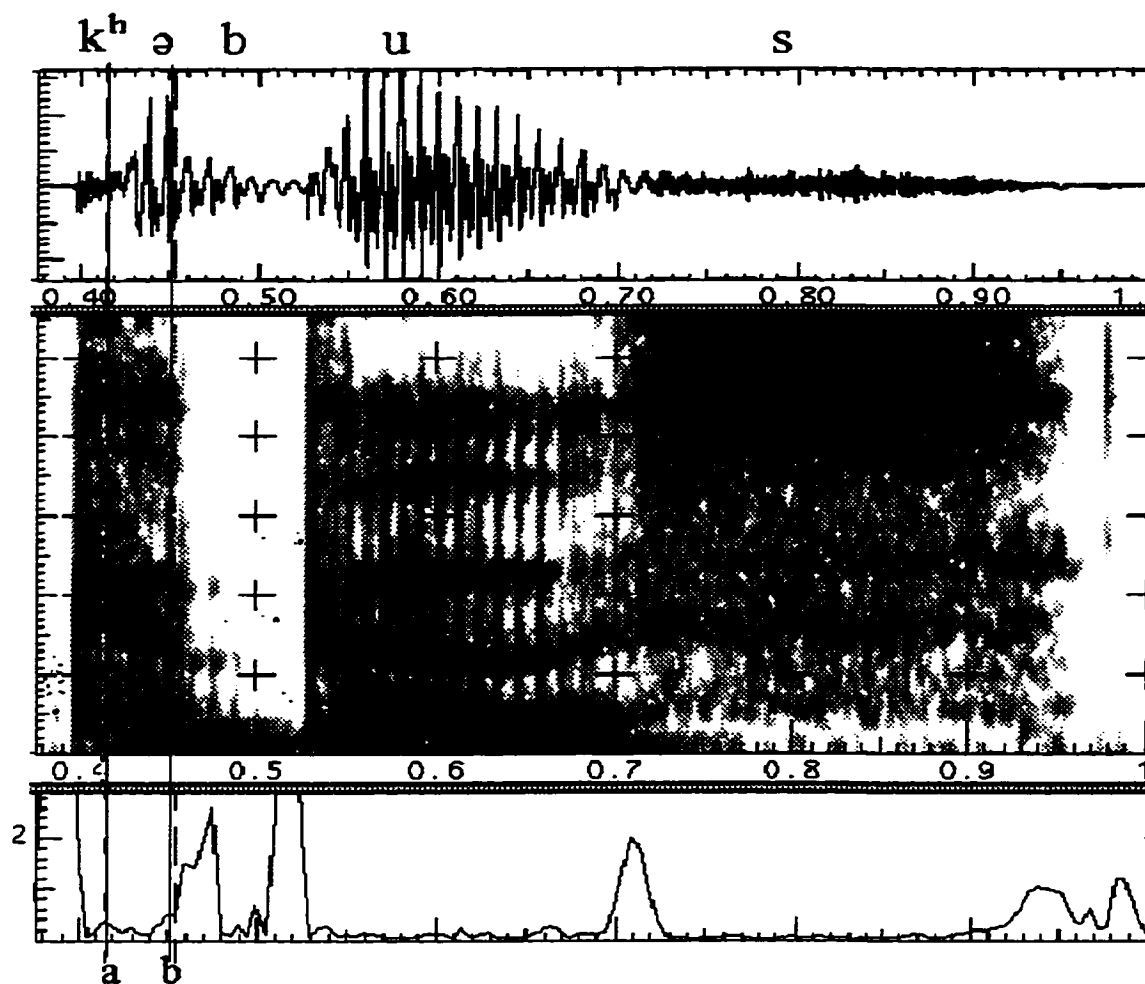


Figure 3.12. Waveform, spectrogram, and D for the word "caboose" /kə/. The point with the largest value of D within the gated area is at or just before the end of the gated area ("b"), but this is not a real peak of D, it is only the beginning of the rise of D toward the peak for the /b/ closure, which falls outside the gated area. Therefore, the point counted as the  $D_{\max}$  point is the one just after the beginning of the gated area ("a"). This is the highest value of D within the gated area which is actually a peak of D, and is associated with the onset of voicing of the /ə/.

Table 3.1. Criteria for counting peaks of the measure D as  $D_{\max}$  points. Each type of transition to which the criterion applies is shown in italics. A description of where peaks of D fall in such cases is given in parentheses.

Situation	Types of transitions (Locations of peaks)	Decision
Only one acoustic change within gated area	<i>Word initial stop-vowel sequence</i> (peak at onset voicing)	Count the single peak as $D_{\max}$
Multiple acoustic changes expected, peaks associated with each	<i>Medial stop-vowel</i> (burst and onset voicing) <i>Medial vowel-stop</i> (closure and burst) <i>Flaps</i> (onset and release of flap)	Count both peaks as $D_{\max}$ points, regardless of relative size
Multiple acoustic changes present but not predicted, peaks associated with each	<i>Fricative-vowel</i> (change in quality of frication noise and onset voicing) <i>Stop-vowel</i> (if there is a separate peak for change from burst noise to aspiration noise)	Count both peaks as $D_{\max}$ points if the smaller is at least 60% as large as the larger
Multiple peaks as an artifact of window length	<i>Cessation of voicing with creaky voice</i> (peaks at each period of creaky voice)	Count only the largest peak for cessation of voicing
Peak associated with an acoustic change other than the transition of interest	<i>Transition to word final segment</i> (peak for end of word), <i>some others as discussed in the text above</i>	Do not count peaks associated with transitions other than the one of interest
Greatest value of D within gated area is not a peak (edge of rise for peak outside gated area)	<i>Very short vowels followed by stops, some other cases</i>	Count only peaks within the gated area, not higher values of D which are not peaks

### 3.2. Problems with the measure D

#### 3.2.1. Changes the measure D does not reflect well

##### 3.2.1.1. Insensitivity of D to linguistically relevant gradual changes

In many cases, there is no clear maximum of D, even though there is acoustic change in the signal. This is the case when the change between segments is gradual, and

especially if it is a change only in the frequencies of the formants (not in amplitude), as in a vowel-vowel transition, vowel formant transitions into or out of a consonant, or sometimes the transition from a vowel to a liquid. Figures 3.13 ("biotech" /a<sup>j</sup>o<sup>w</sup>/), 3.14 (/hatake/ 'field' /ak/), and 3.15 ("elevator" /ɛl/) show examples of the lack of any high values of the measure D where there are clear and rather large changes in the formants. There are extremely few cases in the data in which there is any sign of a peak in the measure D for a change in formants, but Figure 3.16 shows one such rare case, the word /koN<sup>y</sup>aku/ [koŋjaku] 'engagement.' The palatalization of the nasal (because of the following palatal glide) results in very large and sudden changes in formants, which may be reflected by the two small peaks in D (labeled "a" and "b") which occur shortly before and after the nasal, during the periods of the most change in the formants.

The measure D reacts strongly to sudden changes in amplitude, such as the onset of the burst of a voiceless stop or the cessation of formants at the closure of a postvocalic voiceless stop, as was shown in several figures in Section 3.1.2. The highest values of the measure D are associated with the burst of a word initial voiceless stop, as shown in Figure 3.17 (point "a"), in the word "custom" /kʌ/. The measure D also reacts well to sudden changes in the distribution of energy, such as the change from a vowel to a fricative. Figure 3.18 demonstrates this with the word "shell" /ʃɛ/, although the peak of D for the change from frication to vowel is far smaller than the peak of D for the burst in figure 3.17.

Transitions from vowels to and from nasals sometimes show a sudden spectral discontinuity between the vowel and the nasal on a spectrogram, and this sometimes, but not always, corresponds to a small but clear peak in the measure D. Figure 3.19 shows the word /seNmoN/ [semmoŋ] 'specialization,' which has a small peak in the measure D at the onset of the nasal. However, in Figure 3.20, the word /maNneNhitu/ [mannençtsu] 'fountain pen' has no peak of D at the onset of the nasal, which appears as a rather sudden

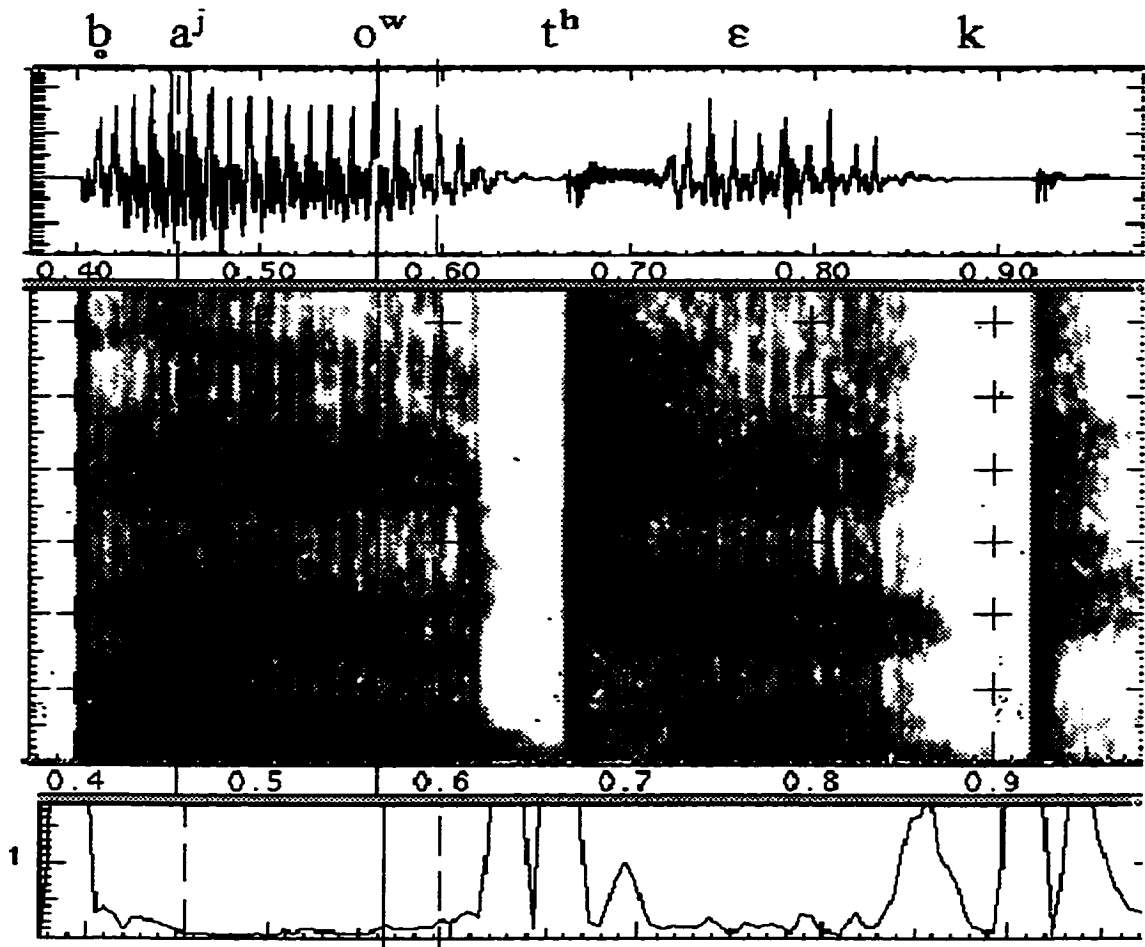


Figure 3.13. Waveform, spectrogram, and *D* for the word "biotech" /aʰoʷ/. Despite considerable movement of the formants during the two diphthongs, *D* remains very low throughout the gated area. The solid line shows the point which was counted as the *D*<sub>max</sub> point, but it is quite small (less than 0.3), and not much higher than the surrounding values of *D*. It is not clear what acoustic change this point is associated with, so its location may not be meaningful.



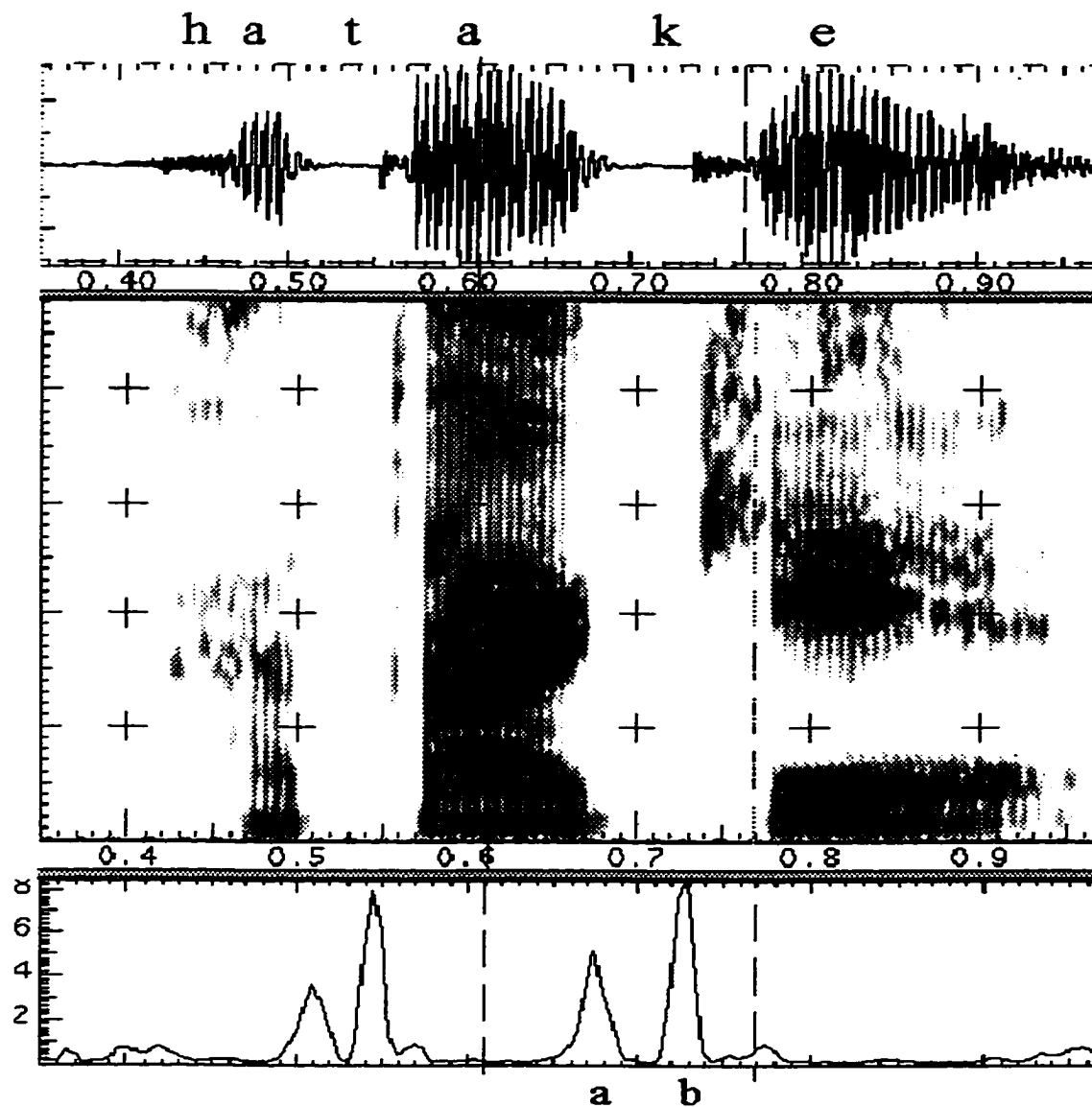


Figure 3.14. Waveform, spectrogram, and D for the word /hatake/ 'field' /ak/. Note the clear formant transitions (converging F2 and F3) during the /a/ going into the /k/. D, however, only shows a peak at the end of voicing ("a"), and another at the burst of the /k/ ("b"), with very little response for the change in formants, if any.

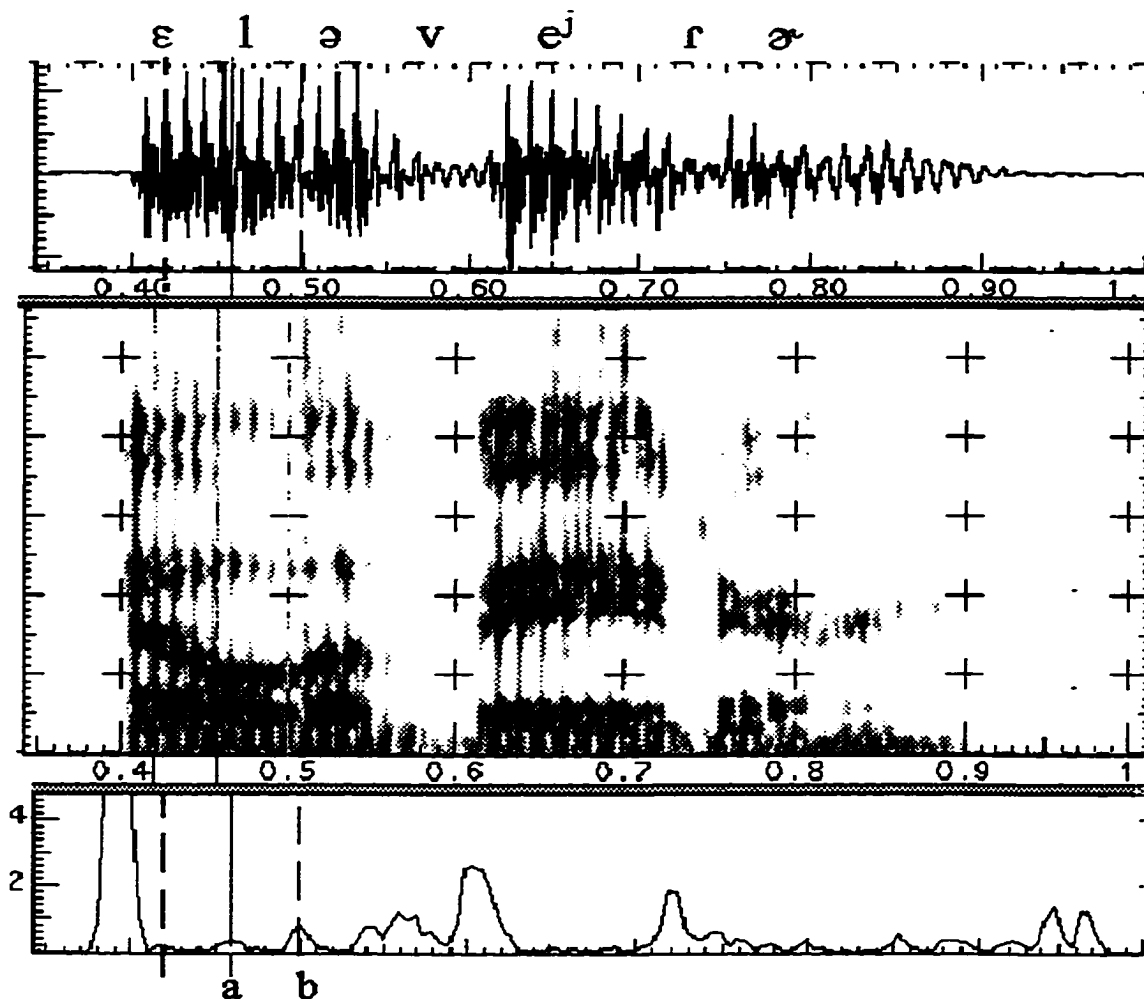


Figure 3.15. Waveform, spectrogram, and  $D$  for the word "elevator" /ɛl/. The small peak of  $D$  which is counted as the  $D_{\max}$  point ("a") occurs during a large change in  $F_2$ , but probably only reflects the decrease in amplitude of  $F_4$  and  $F_5$ . To the extent that the change from the vowel to the /l/ is marked by a decrease in amplitude of higher formants, this point will be at the correct location, but to the extent that change in formant frequencies (especially  $F_2$ ) provides important cues, the  $D_{\max}$  point may be inaccurate. (The apparent peak of  $D$  at the end of the gated area ("b") is actually 1 ms after the end of the gated area, and therefore is not counted as the  $D_{\max}$  point.)

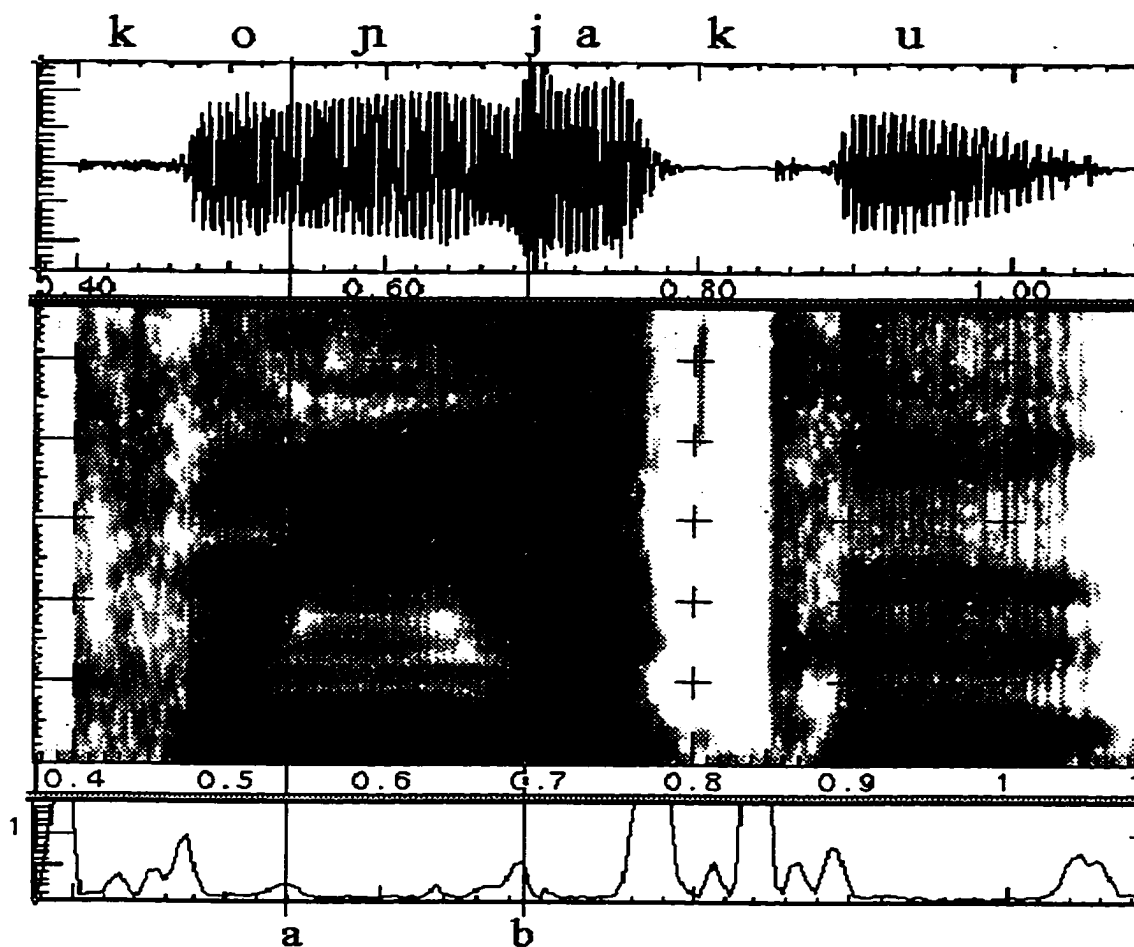


Figure 3.16. Waveform, spectrogram, and D for the word /koNyaku/ [konjaku] 'engagement.' A rare case in which peaks of D may reflect changes in formant frequencies. Two small peaks of D (solid vertical lines labeled "a" and "b") occur during the formant transitions into and out of the palatal nasal. Both occur at least a few periods away from the spectral change between the nasal and surrounding vocalic segments, where one might expect the  $D_{\max}$  for a VN or NV transition to fall, and both fall during large, quick movements of F2. There is no clear change in amplitude at any frequency at these points which could explain their occurrence. (Borders of the gated area are not shown in this figure because the response of D to this particular signal, not the method of locating  $D_{\max}$  points, is at issue.)

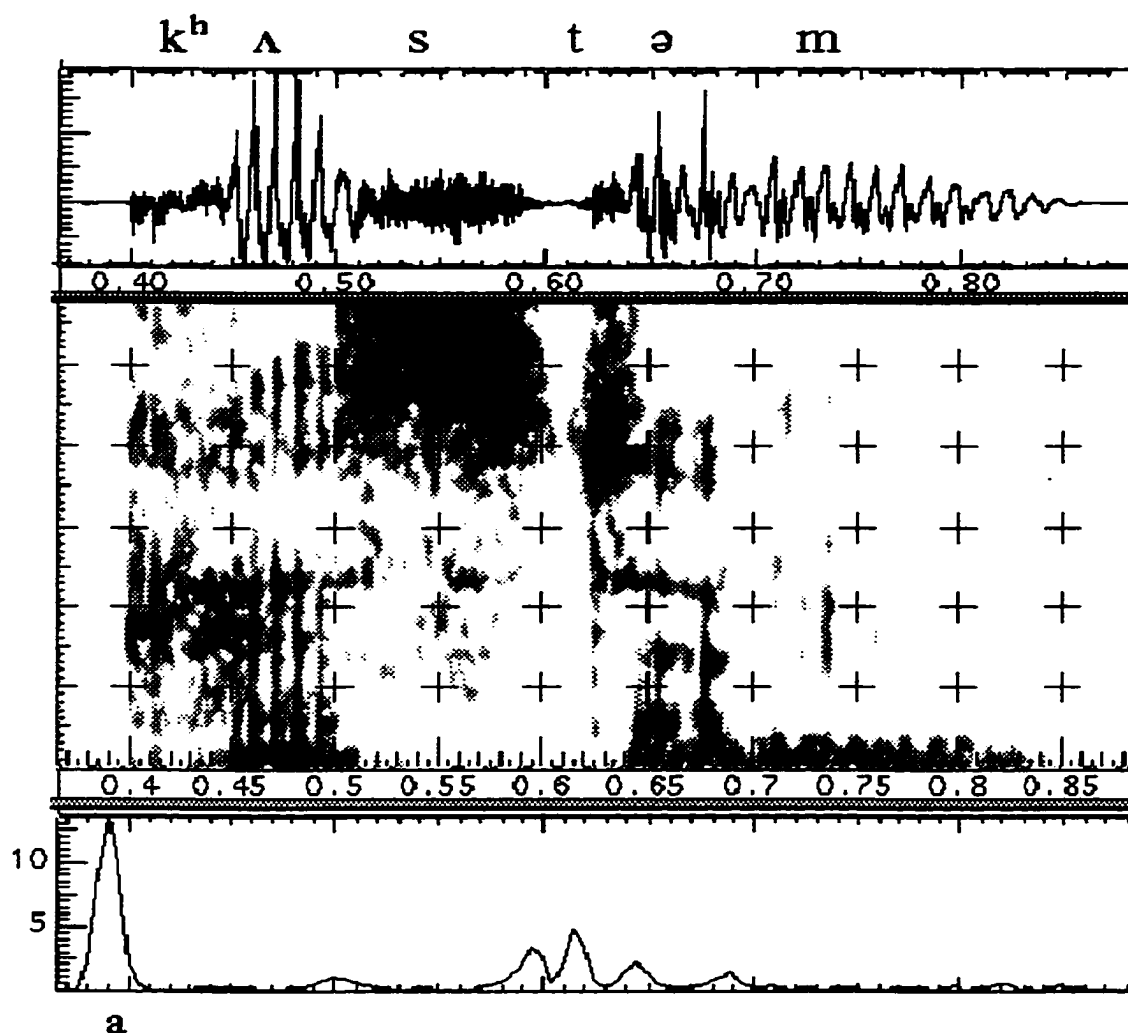


Figure 3.17. Waveform, spectrogram, and  $D$  for the word "custom" /kʌ/. There is a very large peak of  $D$  (labeled "a"), with a value greater than 10, for the beginning of the burst of the /k/. This is comparable to the highest values of  $D$  obtained in the entire experiment. (Boundaries of the gated area and  $D_{\max}$  points are not shown, as the peak at issue is outside the gated area.)

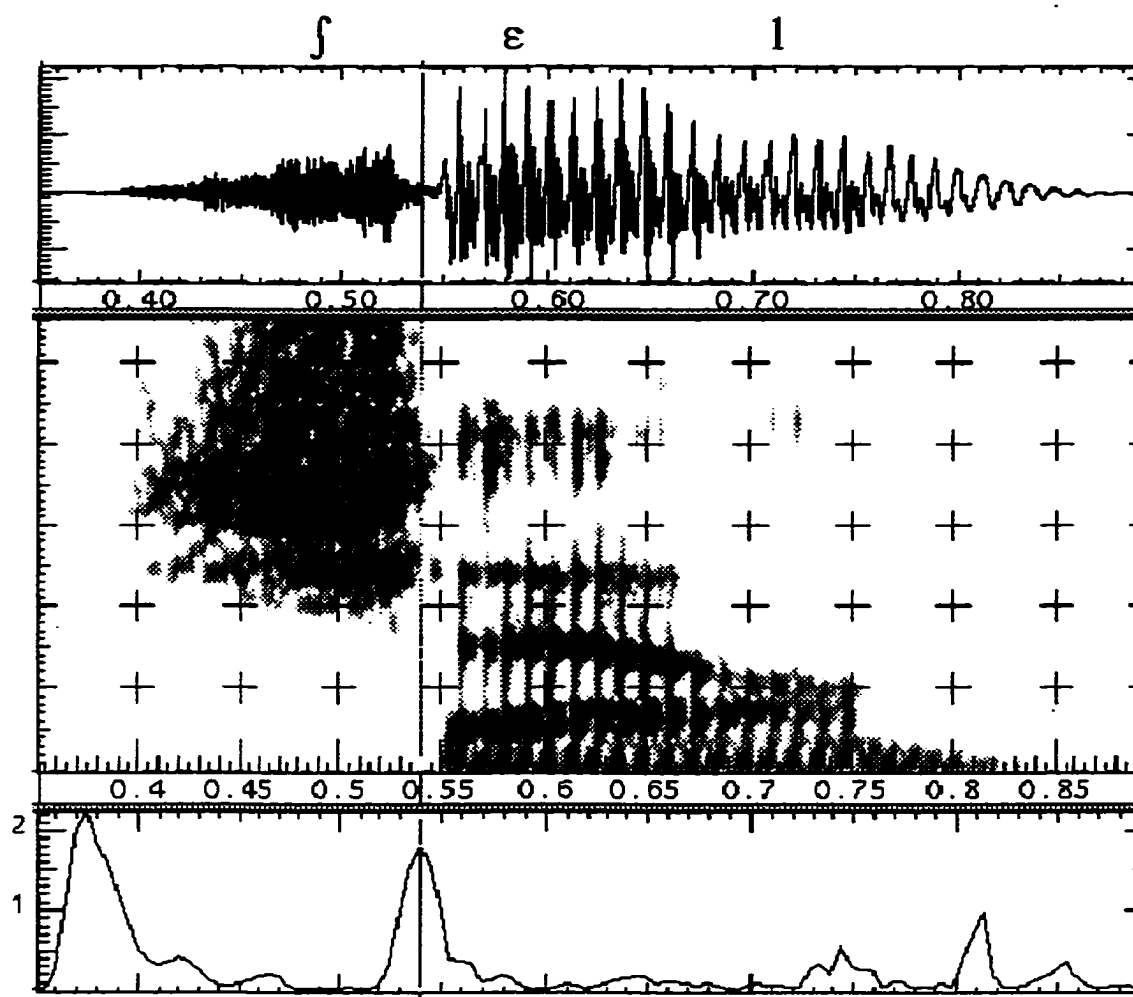


Figure 3.18. Waveform, spectrogram, and D for the word "shell" /ʃɛ/. A clear peak with a value of approximately 2 occurs at the end of frication and near the onset of voicing. Note that while this is a clear peak relative to the values of D for the surrounding signal, it is much smaller than the peak for the stop burst in the previous figure.

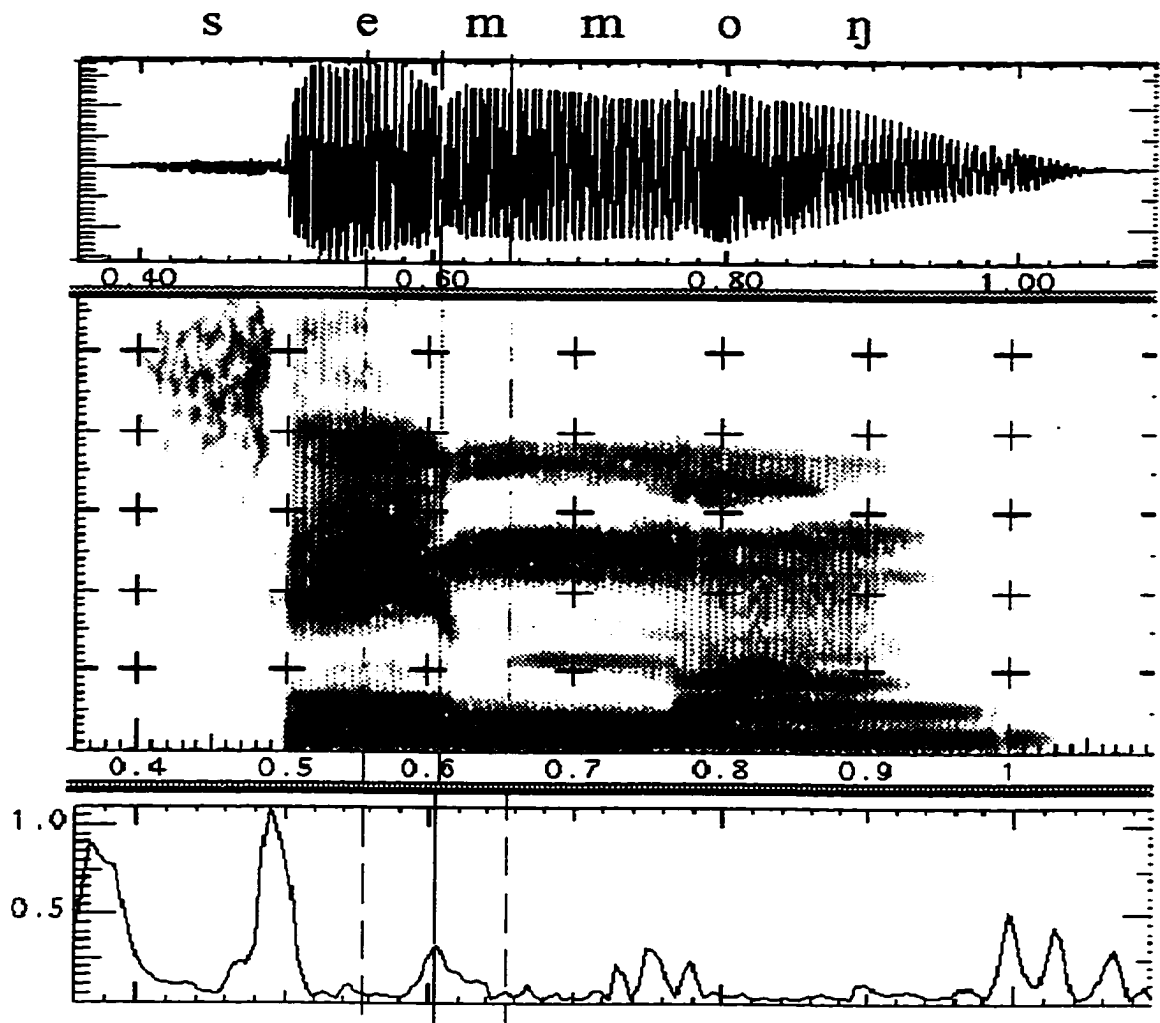


Figure 3.19. Waveform, spectrogram, and D for the word /seNmoN/ [semmon] 'specialization.' A clear, though small, peak of D occurs at the sudden spectral change from the vowel to the nasal. (Boundaries of the gated area and  $D_{\max}$  point are shown with dashed lines, as before.)

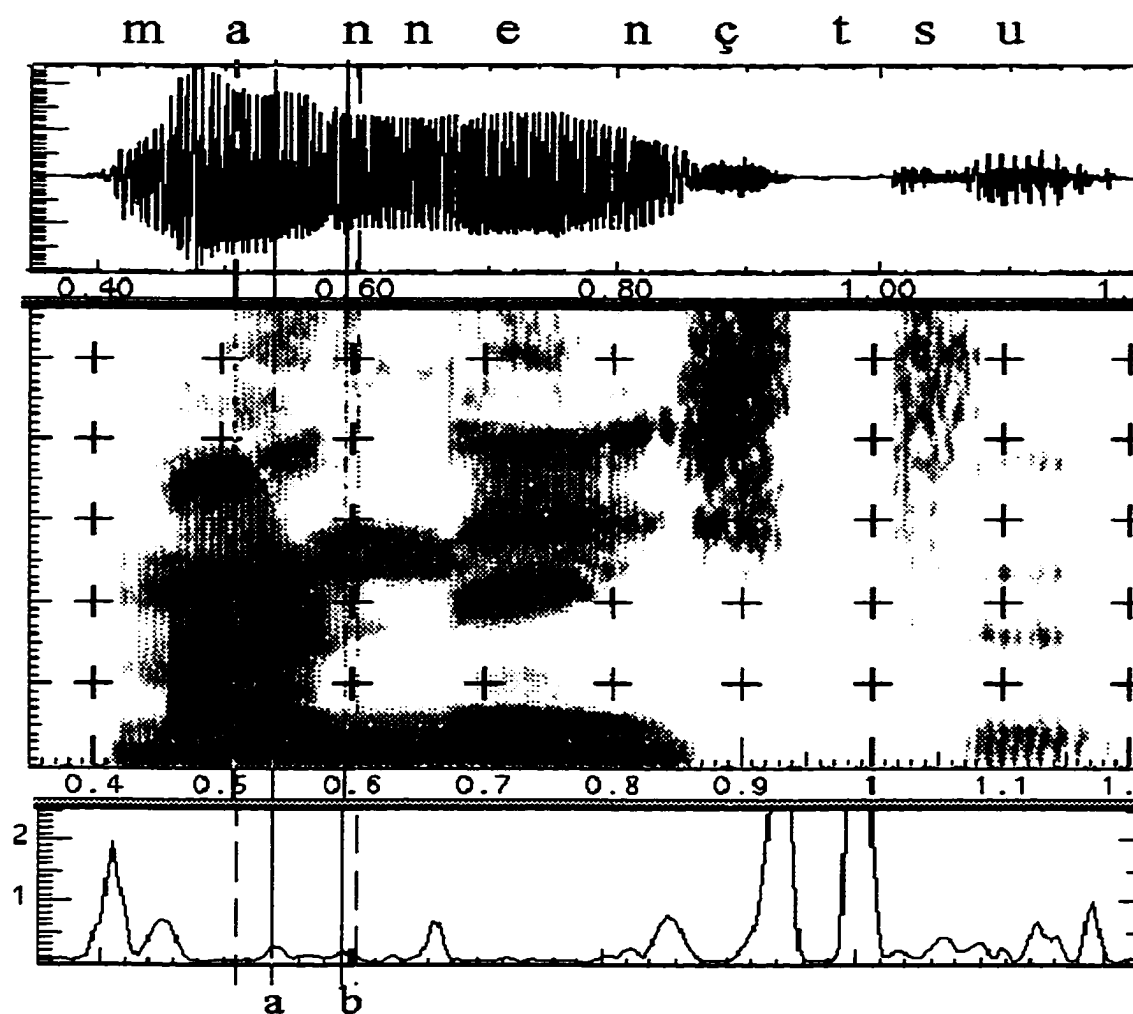


Figure 3.20. Waveform, spectrogram, and D for the word /maNneNhitu/ [mannençtsu] 'fountain pen.' There are two small peaks of D within the gated area ("a" and "b"), but they are not much larger than the values of D for several other time points within the gated area, unlike in the previous figure. Neither one occurs at the onset of the nasal, although there is a sudden change in the spectrum between the vowel and the nasal.

change in the spectrogram. There are two very small peaks of D within the gated area, but one falls several periods before the onset of the nasal ("a") and the other falls a few periods after it ("b"). Thus, the response of the measure D to the onset of nasals is inconsistent.

### 3.2.1.2. Sensitivity of D to linguistically irrelevant changes

In some cases, the  $D_{\max}$  point appears to reflect a small change in amplitude at some frequency, which may be linguistically irrelevant. In Figure 3.21, "trail" /reɪ/, there are two very small peaks of D (both less than 0.5) within the gated area. The first ("a") occurs exactly at the point at which the fourth and fifth formants increase in amplitude. The second ("b") occurs at the onset of high frequency noise above the fifth formant. If these are the acoustic changes to which the measure D is responding, they (especially the latter change) are not changes to which we would wish it to be sensitive. In Figure 3.22, "groan" /gr/, there are two small peaks of D within the gated area. The first ("a") occurs at the onset of voicing, but the second ("b"), which is slightly larger, occurs two periods later, for no discernible reason. Not even a change in amplitude at the high frequencies is visible here.<sup>7</sup> Both of these cases demonstrate that when the linguistically important changes in a signal do not result in an elevation of D because they are too slow, the location of the  $D_{\max}$  point within the gated area is likely to be arbitrary relative to the location of perceptual cues for the transition of interest. Thus, the measure D is sometimes insensitive to linguistically important changes (formant transitions), but is also sensitive to linguistically insignificant changes.

---

<sup>7</sup> There is a third peak of D within the gated area ("c"), to the right of the two discussed. This one is even smaller than the other two, but not by much. I chose not to count this peak as a  $D_{\max}$  point in order to minimize the number of spurious  $D_{\max}$  points included, because it appears to reflect the one very loud period late in the gated area. For some words which do not have any clear maxima of D, the choice of which small peaks to count as  $D_{\max}$  points may sometimes be slightly arbitrary, even though I attempted to apply consistent criteria throughout. These are words for which the measure D is not an adequate measure of degree of spectral change anyway.



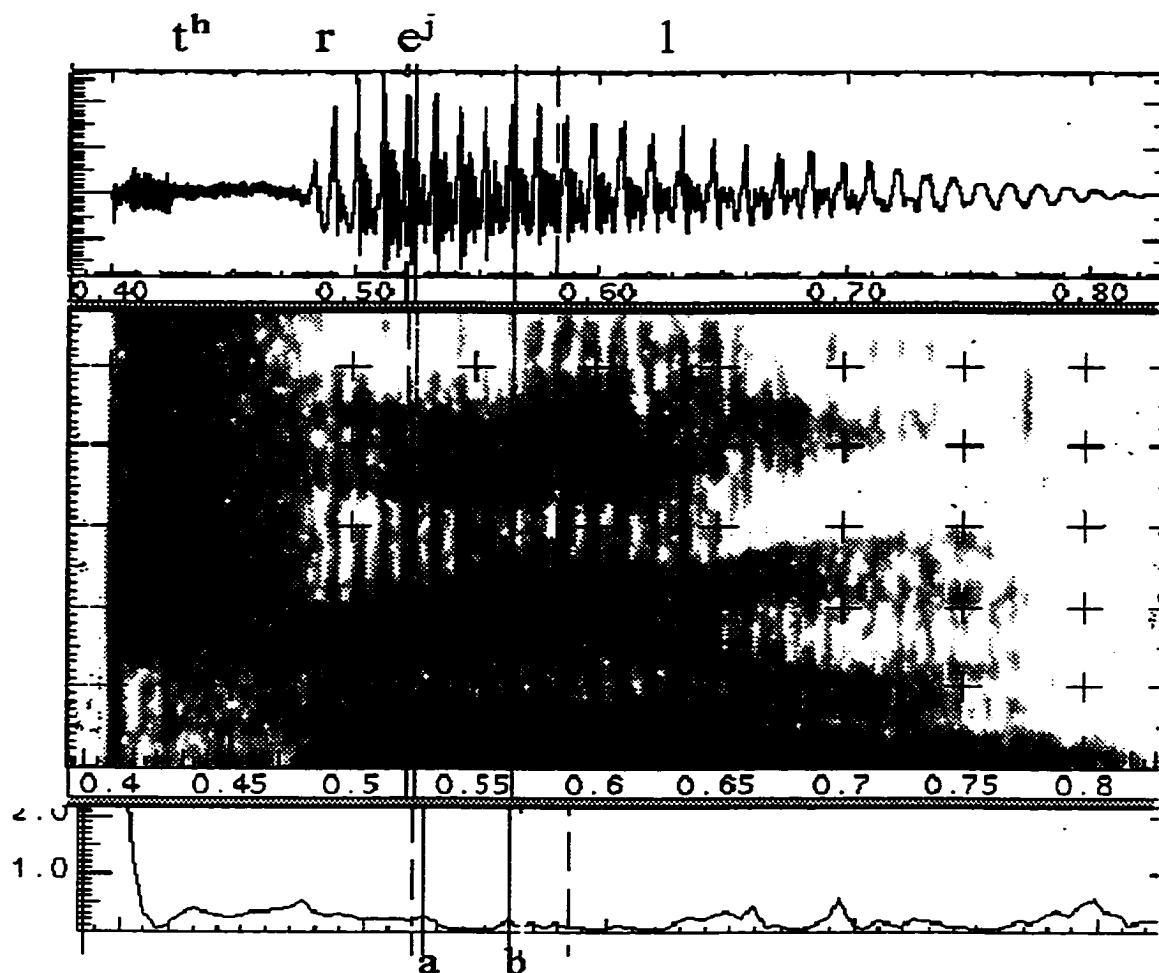


Figure 3.21. Waveform, spectrogram, and D for the word "trail" /reɪl/. There are two very small peaks of D within the gated area, although they are not much larger than surrounding values of D. The first ("a") occurs at the point where F4 and F5 gain amplitude and become "solid." The second ("b") occurs just at the onset of high frequency noise above F5. These are probably linguistically and perceptually insignificant changes, but D fails to show elevated values for the movement of F2 and F3 during this time, showing peaks only for these small changes in high frequency amplitude.

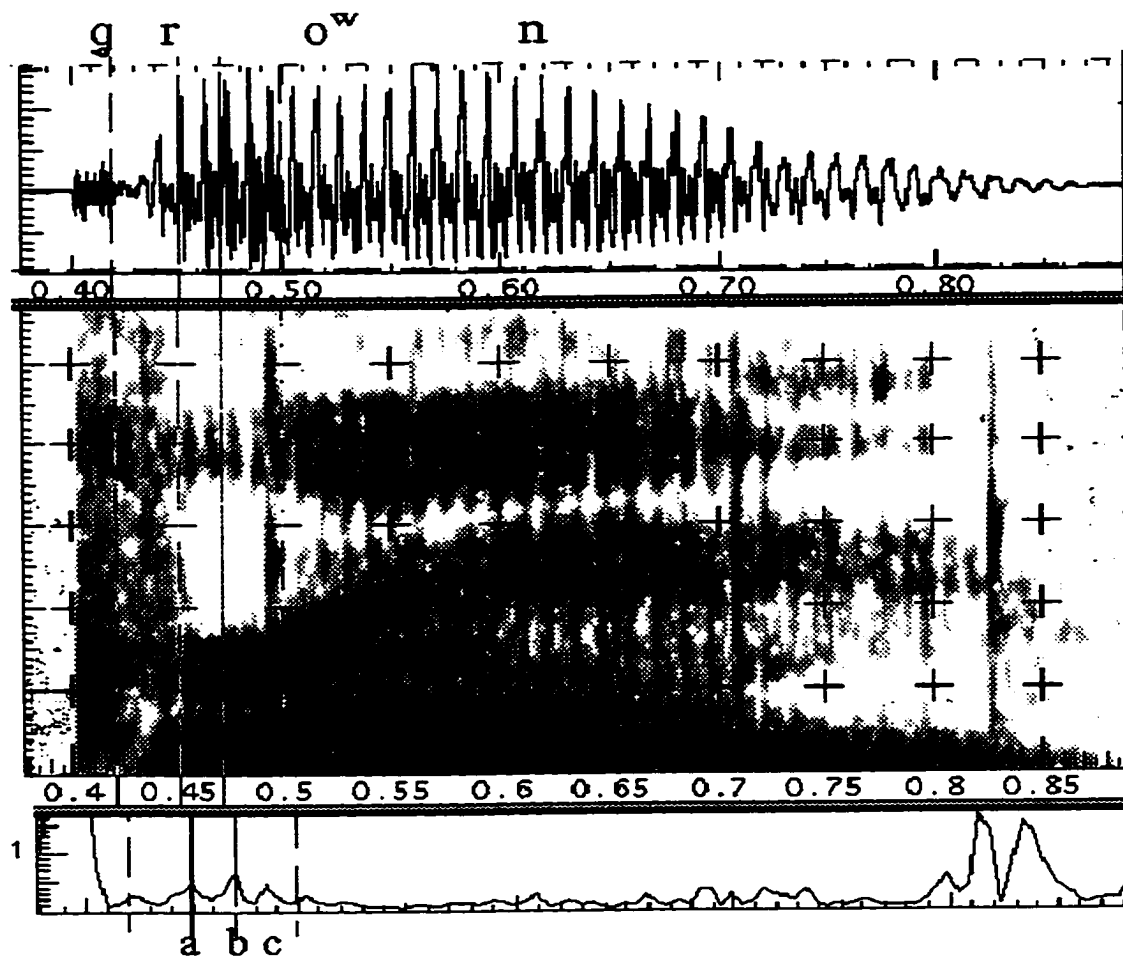


Figure 3.22. Waveform, spectrogram, and D for the word "groan" /gr/. There are three small peaks of D during the gated area. The first is at the onset of voicing ("a"), but the second ("b," the largest) is not at any clear change in the signal. The third ("c"), the smallest, falls at one abnormally loud period.

### 3.2.1.3. Importance of sensitivity to amplitude and frequency changes

The fact that the measure  $D$  is highly sensitive to changes in amplitude is in itself desirable, and not a failing of the measure. Traditional work in phonetics has often emphasized frequency characteristics of the signal rather than changes in amplitude. (Compare the number of studies reporting measurements of formant frequencies to the number of studies reporting rise time or ratios of amplitudes.) This is probably because formant frequencies are easier to measure than amplitudes, since overall amplitude varies with recording level and speech level. The auditory system is probably sensitive to sudden changes in the amplitude of a signal (Greenberg 1994, 1996a<sup>8</sup>), so it is appropriate that a measure of spectral change respond to such changes. This is one of the reasons why it is important to have an objective measure of spectral change, so that traditional biases about what to measure do not exclude important types of changes from consideration. However, the insensitivity of the measure  $D$  to more gradual changes in formant frequencies is not desirable, as this sort of change is very often linguistically significant. When the linguistically relevant changes are exclusively changes in formants, and therefore not reflected by  $D$ , the only points in the signal which do have elevated values of  $D$  tend to reflect extremely small changes in amplitude. Thus, the insensitivity of  $D$  to formant changes renders it excessively sensitive to changes in amplitude.

### 3.2.1.4. Criteria for selecting $D_{\max}$ when $D$ does not reflect changes well

For vowel-stop sequences with clear formant transitions,  $D$  shows a peak for the stop closure, but none for the formant transitions. In these cases, a  $D_{\max}$  point is only recorded for the closure, so the  $D_{\max}$  results simply record one linguistically relevant change

---

<sup>8</sup> Greenberg's reviews of work on response of the auditory system to signals does not include a signal which remains the same in frequency while increasing suddenly in amplitude. However, Greenberg's description of the function of the onset chopper units of the posteroventral cochlear nucleus (1984:4212) implies that at least these units are sensitive to differences in amplitude over a wide amplitude range. In addition, the phenomenon of adaptation makes it clear that the auditory system is highly sensitive at least to changes in amplitude from silence to sound.

and omit the other. However, transitions such as vowel-vowel and vowel-liquid transitions have no changes which D does respond well to in the gated area, and therefore have no clear peak of the measure D associated with any acoustic event. In such cases, I recorded the location of the point with the greatest value of D within the gated area, regardless of whether I could identify the acoustic change which caused it or not. Such small peaks sometimes reflect small changes in amplitude of high frequency noise, as in Figures 3.21 and 3.22, but it is not always possible to find a change in the signal which can account for these small peaks.

There are often several local peaks of D within the gated area, all quite small and of similar size. In such cases, recording only the peak with the largest value of D would introduce a great deal of arbitrary variation into the results. Therefore, I recorded the locations of all peaks of D which were of approximately the same size. (I did not define an exact ratio to use as a criterion for "approximately the same size," and simply located the maxima by eye.) However, in many of the cases with extremely low values of D throughout the gated area, the very small peaks of D may only be the high points of the variation in the baseline D response, and the choice of which such small peaks to count as maxima of D may therefore be meaningless. It might be better to postpone evaluation of such cases until an alternative measure of degree of spectral change, which does reflect slower but linguistically relevant changes, can be developed.

### 3.2.2. Temporal inaccuracy

Although one can manipulate the length of the time window over which one calculates the spectrum of a signal, the spectrum must be calculated over some window, not at a single point in time. Similarly, the measure of spectral change must inherently represent change over some time window. Furui (1986) used a 30 ms window for calculating spectra, and a 50 ms window for calculating the slopes of cepstral coefficient values, which are the basis of the measure D. He believes that the 50 ms window for

calculation of the slopes of the cepstral coefficients is long enough to capture the changes between segments (pointing out that 50 ms is approximately one third the duration of an average Japanese syllable). He also says it is short enough that fitting a straight line to the values of the cepstral coefficients over time is legitimate (1986:1021): if a much longer window were used, the values of the cepstral coefficients might rise with the onset of one segment and then fall with the onset of the next, which would render the slope of the linear regression line meaningless as a measure of degree of change. This is demonstrated in Figure 3.23, an extension of Figure 3.1 to a longer window, using hypothetical data.

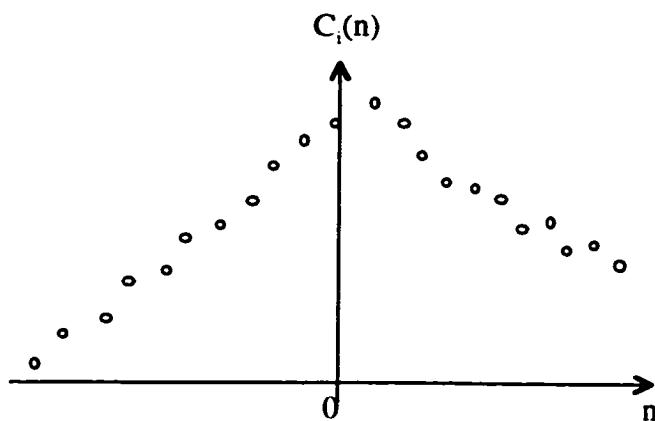


Figure 3.23. A hypothetical case in which too long a window was used for calculating the slope of the regression line of the cepstral coefficient values over time. The values increase until the point after  $n=0$  (an arbitrary choice), but then decrease afterward. To represent this data accurately, two different regression lines would be necessary, one for the rising data and one for the falling data. If just one regression line were fitted, its slope would not accurately reflect the degree of change in the data.

However, the window lengths Furui chose (and which I also used for this experiment) are not without difficulties. The window lengths may be partly responsible for the failure of the measure  $D$  to react to slower changes in the signal, such as formant transitions. More clearly, both the 30 ms window length for calculating cepstral coefficients and the 50 ms window for calculating  $D$  contribute to some inaccuracies in the location of some peaks of  $D$ . This is most noticeable in the location of the  $D_{\max}$  point associated with the bursts of stops: the peak of the measure  $D$  usually occurs 5-15 ms before the onset of the burst, as determined from the waveform or spectrogram. Figure

3.24 demonstrates this problem with the word "academic," in which the peaks in the measure D for both the first /k/ ("a") and the (partially devoiced) /d/ ("b") occur 8 ms before the bursts<sup>9</sup>. This temporal slippage may indicate that the window is too long for rapidly changing segments such as voiceless stop bursts and their following aspiration noise, for reasons similar to the hypothetical example detailed in Figure 3.23. This temporal inaccuracy is unfortunate, but I made no attempt to "correct" the locations of the  $D_{\max}$  points, and simply recorded them at the time of the peak in the measure D, with a note that they occurred early. This phenomenon is almost completely restricted to bursts of stops.

### 3.2.3. Alternative measures of spectral change<sup>10</sup>

#### 3.2.3.1. Possible alternative measures not yet implemented

As discussed above, there are numerous problems with D as a measure of degree of spectral change. This measure worked well for Furui's (1986) data, but he used only Japanese CV and CyV syllables, so the variety of transitions with which the measure was faced was smaller, and most of the types of transitions which have only slower acoustic changes were not included. Therefore, it would be useful to develop an alternative way to measure degree of spectral change. I will, for the most part, leave this for future research, but I will offer some suggestions for possible approaches to the problem. First, manipulating the window lengths for the measure D (for calculating either the cepstral coefficients or  $a_i$ ) might help with the problem of insensitivity to slower changes. However, making the window too long introduces problems for quickly changing segments, as discussed above. I suspect that this sort of manipulation will not improve the results for more than a subset of the transition types, and would probably introduce some additional problems. Since the current window lengths already produce separate peaks of

---

<sup>9</sup> The peak for the final /k/ is also slightly early.

<sup>10</sup> I would like to thank Takayuki Arai, Khalil Iskarous, Herb Clark, Stefan Frisch, and Keith Alcock for suggestions on alternative measures.

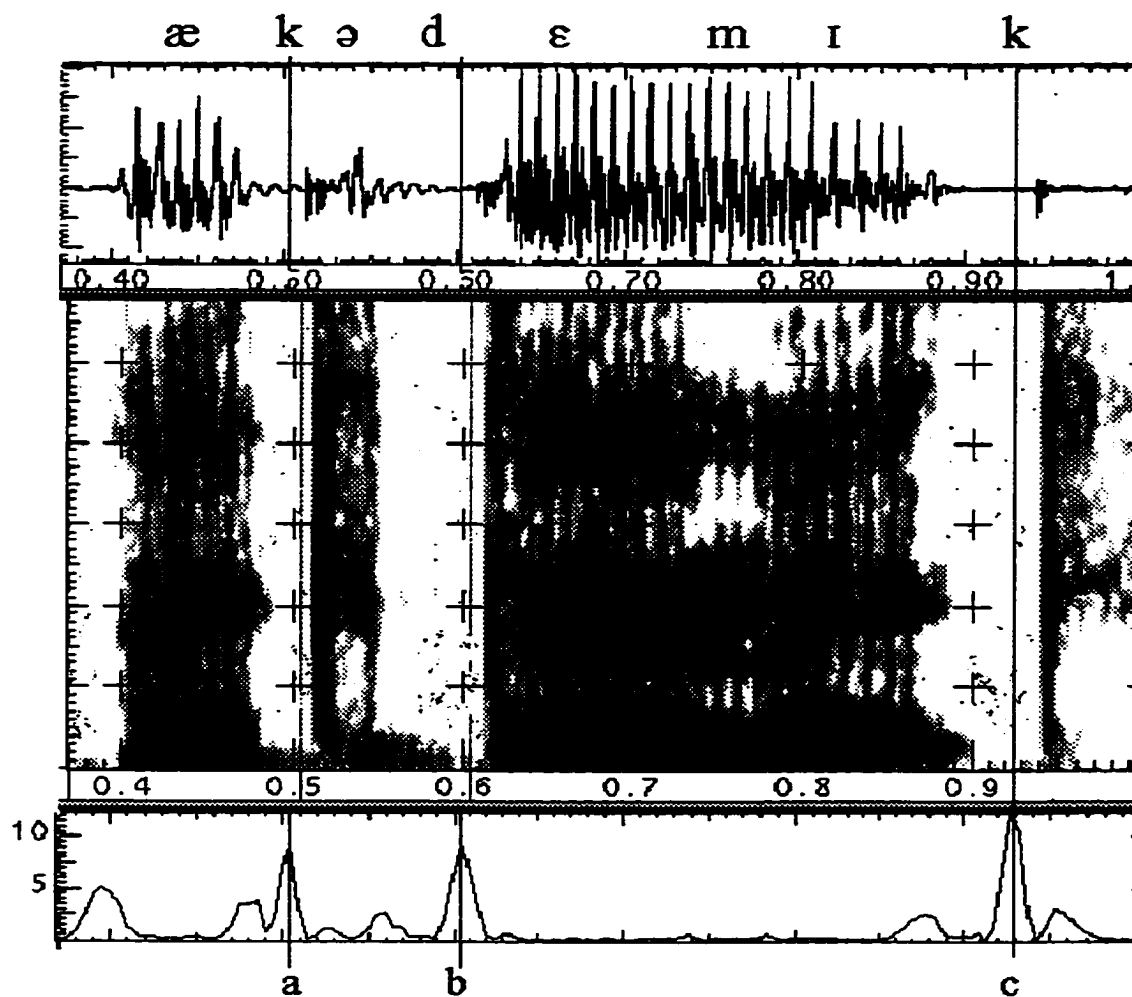


Figure 3.24. Waveform, spectrogram, and D for the word "academic" /kə/. Peaks in D for the burst of both /k/ ("a") and /d/ ("b") occur 8 ms before the burst. The peak for the burst of the final /k/ ("c") is also slightly early. (Boundaries of gated area not shown.)

D for each period of creaky voicing, shorter window lengths might produce separate peaks for each period of regular voicing, for example.

A second simple approach would involve changing slightly the formula for calculating D (shown in (2) at the beginning of the chapter, repeated as (3) here).

$$(3) D(t) = \frac{\sum_{i=1}^p a_i^2}{p}$$

As mentioned in Section 3.1.1, the probable purpose of squaring each  $a_i$  before summing them is to make all of the values of  $a_i$ <sup>11</sup> positive. This is because, in averaging the degree of change over time for the cepstral coefficients, it does not matter whether the change in the cepstral coefficients is in a positive or negative direction. However, one could also make all of the values of  $a_i$  positive by taking their absolute values, instead of by squaring them. The disadvantage of squaring is that it makes large numbers much larger, but makes small numbers greater than one only a little bit larger, and makes numbers less than one smaller. Therefore, if a particular window has relatively small changes in each cepstral coefficient over time, each order cepstral coefficient will have a small number for the slope of its regression line ( $a_i$ ). Squaring and summing these small numbers will result in a relatively small value for D over that time window. However, if the cepstral coefficients have a large change over a time window, they will have large numbers for the slopes of at least some of their regression lines, which when squared and summed will create a very large number.

That is, squaring the  $a_i$  values, while it was probably only intended to make negative slopes positive, also has the effect of emphasizing large changes and de-emphasizing small changes. Using the absolute value of  $a_i$  instead might help to make the values of D for smaller changes, such as the ones between a vowel and a nasal or a sonorant, more

---

<sup>11</sup> Recall that there is one  $a_i$  for each order ( $i$ ) cepstral coefficient.



prominent relative to large changes such as the ones at bursts or closures. The suggested revised equation for D is shown in (4). This approach has not been tested, however, and it probably would not help with changes which are too slow to be reflected by D.

$$(4) D(t) = \frac{\sum_{i=1}^p |a_i|}{p}$$

One promising approach may be to weight particular frequency bands or particular features of the signal more strongly than others in some way. One of the problems with the measure D is that it averages the degree of change over the entire spectrum, but most of the linguistically important changes occur in the lower frequencies (approximately 100-4000 Hz, or even less). Thus, changes in very high frequency noise may have undue influence on the values of D. One might be able to emphasize the more linguistically significant changes in the spectrum by weighting some order cepstral coefficients more strongly than others instead of simply averaging over all the orders.

Another possibility is to apply Furui's formulas for the measure D directly to the results of the FFT (the spectrogram), rather than to the cepstral coefficients, and to weight certain frequency ranges of the spectrum more strongly than others. To do this, one would substitute the value for the amount of energy<sup>12</sup> at a certain time within a certain frequency band in place of the value of a particular order cepstral coefficient in the formula for  $a_i$  (in (1) above). One could choose to weight most strongly the frequency bands which would be likely to contain the second formant, for example. This should reduce the problem of the measure not reacting to changes in formant transitions, since F2 shows many such changes. The concept of "critical bands" (Plomp 1964) might be helpful in determining what bandwidths to use for this weighting.

---

<sup>12</sup> The amount of energy at a certain time and a certain frequency is what is represented as darkness on a spectrogram.

What all of these weighting approaches share is the attempt to locate frequency bands of the signal which are the most important for the linguistic distinctions to which the existing measure D does not react well, and emphasize changes in these bands more strongly than changes in other bands. If most of the change for bursts and fricatives is at high frequencies, this would not be ignored by these weighting methods. However, since the measure D already reacts quite strongly to such changes, the band containing F2, which provides dynamic cues for many distinctions which D does not reflect well, would be emphasized. Instead of deciding to emphasize the band containing F2 because linguists believe it contains many important cues, one could also attempt to determine the relative weighting of bands through reference to the response of the auditory system to various frequencies. One further approach would be to use differential instead of non-differential cepstral coefficients.

#### 3.2.3.2. The issue of sampling rate

Furui (1986), in digitizing the recordings of the word for his experiment, bandlimited the signal at 4000 Hz and used a sampling rate of 8000 Hz. This means that the measure D in his study reflects only changes in the speech signal at frequencies between 0 Hz and 4000 Hz. Since most of the information used for speech perception is at frequencies less than 4000 Hz, this prevents any linguistically irrelevant changes in the very high frequencies (above 4000 Hz) from influencing the results for the measure D. There are some linguistically relevant changes above 4000 Hz: much of the noise in the fricative [s] and some stop bursts is above that frequency. If the noise of a production of [s] were to be entirely above 4000 Hz, a sampling rate of 8000 Hz would prevent the measure D from reflecting the onset and end of that noise well. However, since most high frequency sounds have at least some noise below 4000 Hz, this may only rarely pose a problem.

The data for the experiment reported in this dissertation, however, was sampled at 16,000 Hz rather than 8000 Hz<sup>13</sup>. This means that any changes in the signal at frequencies less than 8000 Hz had an influence in the results for D, since the cepstral values used to calculate D reflect the distribution of energy in the entire spectrum. If there were linguistically irrelevant changes at frequencies greater than 4000 Hz, these would have caused undesired increases in the value of D, and might have led to peaks at points in the signal with no linguistically relevant change. I therefore chose several words which were representative of the types of problems I had found with the measure D, such as its failure to react to changes in formant frequency for transitions into consonants or vowel-vowel transitions, and its failure to react to the change between vowels and sonorants. I also chose a few words which, in the original results, have large peaks of the measure D (at closures and releases of stops and onset and end of frication, for example). I downsampled these words to 8000 Hz (making them equivalent to Furui's data), and recalculated the measure D from these downsampled files, in order to determine whether the use of a higher sampling rate than Furui used caused the problems I found with the measure D.

The words used for this test were: "elevator" /ɛl/, "fair" /eɪr/, "biopsy" /aɪə/, /hatake/ /ak/ 'field,' and "Zachary" /zæ/. "Elevator, fair" represent the problem of D not reacting to changes into sonorants, "biopsy" represents the vowel-vowel problem, and /hatake/ exemplifies the failure of D to react to formant transitions into a consonant. "Zachary" has several large peaks of D in the original results (at the end of frication and the closure and release of /k/), and several segments in /hatake/ and "biopsy" also provide large peaks of D at closures and releases for comparison.

---

<sup>13</sup> This departure from Furui's (1986) methods was unintentional. Very few phonetic measures are influenced by sampling rate in this way.

There was surprisingly little difference between the original D results (for data sampled at 16,000 Hz) and the results from the downsampled data. The downsampled data had somewhat higher values of D overall, probably because the region between 4000 Hz and 8000 Hz has relatively little energy in it during most of the signal, so including it in the calculations lowers the average degree of spectral change results for the whole spectrum. However, the locations of peaks of the measure D were almost identical for the original and downsampled data. Some peak locations differed by 5 ms (one frame of the measure D, the smallest amount of time they could differ by). The original (16,000 Hz) data appears to give slightly more accurate results for the location of  $D_{\max}$  points when predominantly high frequency sounds such as [s] are involved, as one might expect. However, downsampling caused no improvement in the failure of the measure D to reflect changes in formant frequencies. Thus, this failure of D is not due to the inclusion of irrelevant high frequency data. In areas where the 16,000 Hz data has only very small peaks of D, such as the transition to /r/ in "fair," the location of peaks of D for the downsampled data differs in some cases by 5 ms, but it is no more clear in the downsampled data what change in the signal these very small peaks might reflect.

In sum, the use of a higher sampling rate lowers the overall values of D, but makes very little difference in location of peaks. The only case for which a principled difference in location of peaks for the 8000 Hz data and the 16,000 Hz data can be found is peaks near sounds with a strong high frequency component, such as [s], and in these cases the results for the higher sampling rate are the more accurate. Therefore, I did not extend the downsampling approach to the remainder of the words in the experiment.

### 3.2.3.3. Results of one alternative measure of degree of spectral change

I did test one alternative to D for a measure of degree of spectral change, applying it to the same words for which downsampling was tested above. I will refer to this alternative measure of degree of spectral change as D'. It involves using Furui's formula

for  $a_i$  (in 1 above) for each order cepstral coefficient, but rather than averaging the values of  $a_i$  for all order cepstral coefficients, which is approximately what the original measure D does, using instead the greatest  $a_i$  value directly as the measure of degree of spectral change<sup>14</sup>. That is, instead of calculating the average degree of change over the entire spectrum, one counts only the change of that aspect of the spectrum which shows the most change during that time window. In principle, if there is a large change in only one of pitch, distribution of formant frequencies, or amplitude, this measure should reflect that change, without reducing it by averaging with the lack of change in the other features of the spectrum.

However, when this measure D' was tested on the words "fair, elevator, biopsy, Zachary," and /hatake/, there was very little difference in location of peaks of the two measures. Similar to the result of the downsampling, the overall values for D' were higher than for D, because it includes only the greatest change in the spectrum and does not average it with cepstral values for aspects of the spectrum which are not changing. The shape of the responses of D and D', and the locations of their peaks, were very similar, though. D' did, at least in some cases, have a larger difference between peaks for relatively small but identifiable acoustic changes and the background level of D'. This might help an experimenter avoid locating spurious tiny peaks in transitions which neither D nor D' reflects well: it might be possible to set an absolute value of D' which must be exceeded in order for any point to be counted as a maximum of D'. D' does reflect some relatively small but sudden changes better than the original D does, such as the change from burst noise to aspiration noise in the release of a stop. However, D' does not show any identifiable peaks during the formant transitions into a stop or during transitions into sonorants or between two vowels.

---

<sup>14</sup> I thank Herb Clark for suggesting this approach.

The lack of substantial differences in the responses of  $D$  and  $D'$  may indicate that it is rare for two or more different orders of cepstral coefficients to have substantial change at the same time. If this is the case, the average change in all cepstral coefficients would mirror the maximum cepstral change, with smaller overall values, since the maximum would be reduced by averaging it with the very small values of other order cepstral coefficients. The approach to an alternative measure of degree of spectral change described here as  $D'$  seemed, a priori, to be one of the most promising alternative measures. However, it fails to react to exactly the same linguistically relevant changes as the original measure  $D$  does. Since  $D'$  emphasizes whatever aspect of the spectrum is changing the most, this indicates that the problem is probably not in failure to weight the linguistically most relevant band of the signal most strongly, but in the insensitivity of these measures to gradual changes. The measure of degree of spectral change cannot be made more sensitive to gradual changes by lengthening the window over which it is calculated, because the window is already too long for rapid changes such as voiceless stop bursts (as discussed in Section 3.2.2 above). Therefore, to find an alternative measure of degree of spectral change which does reflect both gradual and sudden linguistically relevant changes well, a very different approach will be necessary. Perhaps one could devise two separate measures, one sensitive only to gradual changes and the other only to rapid changes, and count as a point of maximal change a peak in either one. I leave this topic for future research.

### 3.3. Spectral change measurements for English and Japanese acoustic data

Using the criteria described above, I located the  $D_{\max}$  point or points within the gated area for each word. I did not record the raw values of  $D$  at the  $D_{\max}$  points, since the purpose of using  $D$  is only to locate the time points of greatest spectral change. I did, however, record which of the  $D_{\max}$  points (if there was more than one) was the largest. The results appear in Table 3.2.

Table 3.2. Locations of  $D_{\max}$  points for all words in the experiment.  
 $D_{\max}$  points are in milliseconds from the beginning of the word.

No.	Word	Trans.	$D_{\max}$ (largest)	$D_{\max}$ (others)	No.	Word	Trans.	$D_{\max}$ (largest)	$D_{\max}$ (others)
English					Japanese				
1	tip	tɪ	75		1	todana	[to]	25	
2	stiff	tɪ	150	180	2	tatoe'ru	[to]	165	190
3	Tibet	tɪ	15		3	ka'to	[to]	190	
4	petition	tɪ	115	160	4	kakari'iN	[ka]	60	
5	attic	tɪ	125		5	hakama'	[ka]	150	190
6	custom	kʌ	50		6	sya'kai	[ka]	255	285
7	skull	kʌ	205	240	7	dama'ru	[da]	15	
8	accompany	kʌ	80	100	8	midare'ru	[da]	110	
9	caboose	kə	15		9	ku'da	[da]	125	
10	academic	kə	105	125	10	hotoke'	[ot]	100	55
11	duck	dʌ	80		11	himoto'	[ot]	375	300
12	citizen	ɪt	185	165	12	hakobu	[ak]	150	105
13	fitness	ɪt	80	120	13	hatake	[ak]	330	275
14	Italian	ɪt	70	50	14	ha'yaku	[ak]	340	245
15	committee	ɪt	225	245	15	kadai	[ad]	140	120
16	unity	ɪt	255	235	16	hanada'yori	[ad]	260	
17	bucket	ʌk	125	85	17	ka'nada	[ad]	305	270
18	mechanical	ək	115	85	18	megumi	[me]	40	
19	indicate	ək	195	165	19	tomeru	[me]	145	
20	induction	ʌk	245		20	nemui	[ne]	30	
21	muddy	ʌd	150	170	21	kemuri	[em]	120	
22	cadenza	əd	120	80	22	tabemo'no	[em]	215	
23	medicine	mɛ	40	55	23	teni'motu	[en]	115	
24	immense	mɛ	90		24	soda'tu	[so]	35	80
25	remedy	ɛm	120		25	zabu'toN	[za]	40	
26	attempt	ɛm	275		26	syabe'ru	[ʃa]	95	115
27	negative	nɛ	65		27	hokeN	[ho]	45	
28	tenants	ɛn	135		28	zyosei	[os]	120	
29	saddle	sæ	100	115	29	kazari	[az]	125	
30	master	æs	180		30	basyo	[aʃ]	120	
31	Zachary	zæ	55		31	gohoo	[oh]	115	
32	asthma	æz	105		32	wahuku	[aɸ]	150	125
33	shell	ʃɛ	140		33	dohyoo	[oç]	130	
34	session	ɛʃ	220		34	harada'tu	[ra]	100	

No.	Word	Trans.	D <sub>max</sub> (largest)	D <sub>max</sub> (others)	No.	Word	Trans.	D <sub>max</sub> (largest)	D <sub>max</sub> (others)
35	fees	fi	60	15	35	yubi'	[ju]	55	
36	unfeeling	fi	190		36	kara'i	[ar]	135	155
37	leaf	if	250	230	37	huyoo	[uj]	100	
38	relief	if	360		38	mawari	[aw]	110	
39	vacuum	væ	50		39	tyazuke	[tʃa]	50	
40	ravish	æv	135		40	zyokyo'ozyu	[dʒo]	45	
41	trail	reɪ	125	160	41	mati'	[atʃ]	210	150
42	fair	eɪr	195	155 165	42	tozi'ru	[odʒ]	135	150
43	lever	le	75		43	haNtai	[nt]	300	220
44	elevator	el	60		44	kaNdoo	[nd]	305	285
45	yellow	je	60		45	teNkiN	[ŋk]	295	225
46	watch	wa	75		46	kaNzeN	[nz]	280	
47	chapel	tʃæ	85		47	seNsoo	[ns]	305	
48	latches	ætʃ	175	145	48	keNritu	[nr]	265	
49	jump	dʒʌ	35		49	koNyaku	[ŋj]	235	
50	judge	ʌdʒ	210	235 265	50	kiNtyoo	[ntʃ]	290	255
51	bent	nt	255	215	51	sukunai	[sk]	105	150
52	sentiment	nt	210		52	sikaku	[s'k]	125	150
53	reinterpret	nt	210	190	53	kitamuki	[k't]	105	80
54	band	nd	295	275	54	kokutetu	[kt]	210	250
55	wander	nd	280		55	kyaku	[kj]	60	
56	recondition- ed	nd	255		56	dakyoo	[kj]	190	240
57	axe	ks	245		57	hyoo	[çj]	105	
58	hacksaw	ks	165	190	58	ryokaN	[rj]	15	
59	unaccep- table	ks	165		59	mottaina'i	[tt]	265	
60	cats	ts	295		60	sakka	[kk]	380	
61	Betsy	ts	135		61	sassoku	[ss]	265	
62	stop	st	130	110	62	hassya	[ʃʃ]	270	230
63	based	st	400	375	63	teNmetu	[mm]	190	
64	pastime	st	290	260	64	aNnaizyo	[nn]	195	145 175
65	skate	sk	160	100	65	tootyaku	[oo]	125	105
66	mask	sk	450	405	66	keigo	[ee]	120	175
67	discount	sk	180	150	67	syuukaN	[uu]	220	245
68	train	tr	75		68	haori	[ao]	160	200
69	string	tr	160	195	69	siatu	[ia]	250	
70	Detroit	tr	105	210	70	kaeri'miti	[ae]	115	
71	crops	kr	80		71	taiko	[ai]	115	



No.	Word	Trans.	$D_{\max}$ (largest)	$D_{\max}$ (others)	No.	Word	Trans.	$D_{\max}$ (largest)	$D_{\max}$ (others)
72	scrap	kr	180	215	72	koibito	[oi]	105	
73	acrobat	kr	120	165	73	teNiN	[ēi]	195	295 320
74	drop	dr	30		74	hiN	[iŋ]	225	255
75	groan	gr	70	50 85	75	maNne'N- hitu	[an] <sup>15</sup>	140	195
76	plain	pl	95	40	76	seNmoN	[em]	205	
77	split	pl	230	260					
78	twelve	tw	50						
79	court	rt	295	245					
80	cork	rk	270	205					
81	help	lp	310	235					
82	fans	nz	415						
83	dance	ns	250						
84	fancy	ns	220						
85	uncon- cealed	ns	240						
86	snow	sn	170	130					
87	Disney	zn	185						
88	farm	rm	355	260					
89	corn	m	285						
90	film	lm	155						
91	ranch	nŋ	285	210					
92	flash	fl	45						
93	fragile	fr	45						
94	sleep	sl	135						
95	Iceland	sl	255						
96	swan	sw	160	125					
97	golf	lf	230						
98	wharf	rf	195						
99	false	ls	235						
100	calcium	ls	175						
101	cultural	ŋ	160						
102	marginal	rdʒ	185	160					
103	optical	pt	145						
104	pact	kt	330	265 295					
105	coughs	fs	300						
106	nerves	vz	400						

<sup>15</sup> For words 75 and 76 of the Japanese data, the nasal in this transition is the mora nasal.

No.	Word	Trans.	$D_{\max}$ (largest)	$D_{\max}$ (others)
107	amnesty	mn	175	
108	garlic	rl	130	
109	biopsy	a <sup>j</sup> a	190	
110	biography	a <sup>j</sup> a	100	
111	biotech	a <sup>j</sup> o <sup>w</sup>	165	
112	eon	ia	35	130
113	diagonal	a <sup>j</sup> æ	85	100
114	react	izæ	200	
115	tiger	ta <sup>j</sup>	120	85
116	bite	a <sup>j</sup> t	180	
117	data	de <sup>j</sup>	65	
118	fade	e <sup>j</sup> d	320	300
119	doubt	a <sup>w</sup> t	285	260
120	soybean	o <sup>b</sup>	300	260
121	toad	to <sup>w</sup>	65	
122	oats	o <sup>w</sup> t	255	180
123	courage	kə	80	
124	circle	ɔk	240	220
125	button	tɪ	120	
126	beetle	tɪ	140	
127	apple	pl	195	220

These results will be used for comparison with the perceptual results in Chapter 4.

#### 4. Results

##### 4.1. Method of analyzing the results

##### 4.1.1. Transcribing the responses

The handwritten responses from the 154 English speaking subjects and 120 Japanese subjects were transferred to a computer for analysis. Misspellings on the part of the English subjects were corrected, but heterographic homophones (responses from different subjects to the same stimulus which have different spellings but the same pronunciation, such as "baste" and "based") were not standardized in the original inputting of the data, so that the meaning of the response intended by the subject is recoverable. The experimenter asked subjects about any homographic heterophones (such as "read," /rid/ vs. /rɛd/) at the time of the experiment, and the pronunciation the subject gave was recorded along with the orthographic response. In a few cases, the experimenter did not notice such responses at the time of the experiment, and the subject's intended response is therefore not known. In these cases, it was assumed that the subject had intended the same pronunciation as was given by other subjects to the same stimulus.

Japanese subjects' responses were transcribed phonemically for purposes of data analysis. In a few cases, Japanese subjects used incorrect Chinese characters to write a word (the equivalent of a misspelling). Since subjects were asked to include the intended pronunciations using the syllabary for all words they wrote in Chinese characters, the intended pronunciation was usually clear, and this was used rather than the pronunciation of the characters the subject actually wrote. Several subjects did not follow the instructions to add syllabary notation to their Chinese characters, and Japanese has a large number of homographic heterophones when words are written in Chinese characters, particularly when the word consists of only one character. In most cases, another subject gave the same response (using both Chinese characters and syllabary) to the same stimulus, so the ambiguous response was assumed to match those of other subjects. A native speaker of Japanese also evaluated ambiguous cases, and was usually able to state that only one

pronunciation was possible in the context of the response. This left only two ambiguous responses. In both cases, the two possible pronunciations of the characters had very different segmental content, and the stimuli allowed subjects to hear two to three segments of the word, so the pronunciation with the same initial phonemes as the stimulus was chosen.

No attempt was made to record the location of the pitch accent for the Japanese responses. Orthography does not mark accent. In theory, accent information for subjects' responses could be recovered with a pitch accent dictionary, but there is considerable variation in lexical placement of accents in Japanese, certainly between dialects, and also between generations. Many of the subjects had lived for several years in more than one dialect area of Japan, and it is impossible to know where they place accents. Lexical placement of accent is also different in Nagoya and in Tokyo for some words. Furthermore, many subjects gave responses such as proper names and slang terms which accent dictionaries do not list. Naive speakers are not aware of where they place accents, so the only way to be sure of accent locations would be to record subjects' productions of each of their responses. Because reliable accent information was not available, any responses with identical phonemic content which differed in location of accent (and in orthography) were treated as the same response. However, except for dialect variation, cases in which a difference in intended accent placement was lost were probably rare, as subjects who gave responses with the same phonemic content usually all used the same Chinese characters for the word.

For the English data, homophonic heterographic responses were then standardized, so that "based" and "baste" were counted as the same response, for example. Because this experiment is about the timing of perception of segments and spoken word recognition relative to the timing of acoustic events, and no amount of phonetic information will help a listener decide whether a word in isolation is "baste" or "based," such words are the same for purposes of the experiment, even though listeners giving these responses have accessed

different lexical items<sup>1</sup>. As the Japanese data had already been transcribed phonemically, no further conversion of this sort was necessary. For both languages, however, different forms of the same word, such as "base" and "based," were counted as different responses.

#### 4.1.2. Conversion to numerical data

After the conversions discussed above, the data was evaluated in two ways: the number of different responses measure (#Resp) and the percent with second segment correct measure (%Corr)<sup>2</sup>.

##### 4.1.2.1. The number of responses (#Resp) measure

For the number of responses (#Resp) measure, I counted the number of different responses given to each stimulus. For the English data, this yields a result on a scale of one to eleven, as the pool of eleven subjects must give at least one unique response to the stimulus, and they can at most each give a unique response. These results were converted to a zero to ten scale by subtracting one from the raw data. For the Japanese data, the raw data scale was one to twelve, because there were twelve subjects in each condition instead of eleven. This was converted to a zero to ten scale by subtracting one and then multiplying by 10/11.

In the few remaining cases of subjects failing to give any response to a stimulus, or responding with non-words such as "sss" instead of real words (discussed in Sections 2.1.3 and 2.1.4 above), "no response" was counted as a different response from other, real words. That is, if only one subject (of the eleven or twelve to hear the stimulus) failed to give a response to a particular stimulus, this increased the total number of responses by one, just as if that subject had given a real word response which no other subject gave.

---

<sup>1</sup> Which word of the pair a particular listener accesses probably depends on the relative familiarity of the homophones to the particular listener, but it has no import for this experiment.

<sup>2</sup> To make it easier to keep track of the two types of data, I will also abbreviate "number of different responses" or "number of responses" as "#Resp" and "percent second segment of interest correct" or "percent correct" as "%Corr."

However, if more than one subject failed to give a response to the same stimulus (a situation which occurred only for the Japanese data and rarely there), all of the non-responses were counted as the same, so the total number of responses was still only increased by one.

The number of responses measure (#Resp) is intended to be a measure of progress toward spoken word recognition. In this experiment, even the longest gate often does not allow listeners to hear enough of the word to be sure what it is. Therefore, for the majority of the words in the experiment, the group of listeners hearing the stimulus do not all come to agree on the one correct response even by the final gate. In an experiment which involves gating through entire words, such as that by Grosjean (1980), one can determine the point at which all listeners have recognized the correct lexical item. What is of interest in the current experiment is not where in the signal listeners recognize the single correct word, but rather where in the signal they make the most progress toward recognizing the word. This is usually reflected by a relatively sudden reduction in the number of different responses, although exceptions will be discussed in Section 4.4.5 below. Thus, although I will refer to the concept of spoken word recognition throughout the subsequent discussion, what I am investigating is progress toward recognition of a spoken word, not necessarily successful recognition of a unique lexical item.

#### 4.1.2.2. The percent correct (%Corr) measure

For the percent correct (%Corr) measure, I recorded whether each response had the second phoneme of the two phoneme transition of interest correct. Percent correct was represented on a zero to one scale. Responses were evaluated relative to the second segment of the transition, regardless of which gate the subject heard, in order to determine at which gate listeners perceived the second segment. In order to be counted as correct, the response had to match the stimulus' segment at a subphonemic level. That is, "correct" responses to the stimulus "citizen" /sɪtəzən/ with the target transition [ɪr] included both

"citizen" and "city," but not "sit." For the Japanese stimulus /kitamuki/ [**k**<sup>h</sup>itamuki]<sup>3</sup> 'North face,' the response /kiti/ [kiti̯] 'military base' was not counted as correct, even though it contains the same phoneme as the stimulus, because /t/ is affricated before /i/, so this response shows that the listener had not yet perceived that the relevant segment was a stop. In the Japanese data, for VC transitions, responses were counted correct if they had the correct place, manner, and voicing of the consonant, but used a geminate instead of a singleton or vice versa, because the issue in perception of a VC transition is the perception of the quality of the consonant, not its duration. However, in CC geminate sequences, only geminates were counted as correct, since the issue is the timing of perception of distinctive length.

In some cases, a response contained the correct phoneme, but with a different number of phonemes before it than the stimulus word had. If the difference between response and stimulus involved inserting an unstressed vowel between the two segments of the gated transition, the response was counted as correct, as for example "coroner" and "quarantine" as responses for "corn" /kɔrn/, or "isolate" for the stimulus "Iceland" /a<sup>h</sup>islænd/. (A fast production of "coroner" could be homophonous with "corner," so this need not even be viewed as insertion of a vowel.) A similar case which was also counted as correct was "fantasy" for the stimulus "fancy" /fænsi/, since the /t/ in "fantasy" can be deleted in fast speech. If additional segments were inserted somewhere before the gated transition, but the phonemes in or around the transition of interest were the same, the response was also counted as correct, as in "affair" for the stimulus "fair" /fe<sup>h</sup>r/, "afraid" for "fragile" /frædʒl/, or "max" for "axe" /æks/. (One should note that these responses

---

<sup>3</sup> Recall that the transition of interest is printed in bold, as mentioned in Chapter 2. If the word is written only orthographically because it is not necessary to transcribe the entire word, then the transition of interest appears after the word.

would not even be present in the cohort of the stimulus word in a simple cohort model (Marslen-Wilson and Welsh 1978, Taft and Hambly 1986), and that they point to the importance of the stressed syllable, or perhaps its rhyme, for English listener's spoken word recognition. This will be discussed in Section 5.3.1 below.)

Such cases of misalignment were more numerous for the English data than for Japanese, presumably because of the possibilities for vowel reduction and consonant cluster simplification in English. Misalignments of correct phonemes in Japanese usually involved the substitution of a geminate for a singleton or vice versa at a point before the second segment of the transition, as in the response /sakkaa/ [sakkaa] 'soccer' for the stimulus /syakai/ [sakai] 'society.' Such cases were counted as correct.

In order for the second segment of the transition to be counted as correct, it was not necessary that the preceding segments all be correct also, or even that the first segment of the transition be correct. This is a logically separate situation from the misalignments discussed above: even if a response has the same number of segments before the target segment as the stimulus does, it is not necessary that all of the segments before the target segment be correct for the target segment itself to be counted as correct. If the response had the second segment in the correct position in the word and that segment was correct, it was counted as correct regardless of what preceded it. For example, the responses "morph" to the stimulus "wharf" /wɔːrf/ and "reduction" to "induction" /ɪndʌkʃən/ were counted as correct. Such cases were very rare in the Japanese data.

In a few cases, a response had both a problem in alignment of the correct segment (segments inserted or deleted before it) and in identity of preceding segments. If the response had a segment matching the second segment of the target transition in the same position of the same syllable as the stimulus word, this was still counted as correct. For example, the response "hungry" was counted as correct for the stimulus "unfeeling" /ʌnfɪlm/ because "hungry" has the target /i/ as the nucleus of the second syllable, as the



stimulus word does. A Japanese example is the response /ryohi/ [rjoçi] 'travel expenses' for /dohyoo/ [doçjoo] 'wrestling ring.' There were very few such cases, however, even in the English data. In this particular case, counting "hungry" as in any way correct for "unfeeling" may seem undesirable, but this allows for consistent treatment of the data in determining what is correct and what is not<sup>4</sup>.

For a few transitions in both languages, subjects never perceived the second segment of the transition well, even by the last gate. If correct perception of the second segment never exceeded 30%, the usual percent correct measure was replaced by a "partially correct" or "two features correct" measure. In these cases, the percent of responses in which the segment corresponding to the second segment of the transition had any two of place, manner, and voicing correct was calculated. If the second segment of the transition was a vowel, the percent of responses for which the corresponding vowel was within the same tense/lax pair was calculated. (This is of course only possible for English, as Japanese does not have its vowels in pairs, but none of the Japanese vocalic targets failed to reach 30% correct.)

The following eight English stimulus words and five Japanese words never reached 30% correct for the second segment: "ravish" /ræviʃ/, "asthma" /æzmə/, "muddy" /mʌdi/, "immense" /ɪmens/, "shell" /ʃel/, "chapel" /tʃæpəl/, "yellow" /jeləʊ/, "Disney" /dɪzni/, and "Tibet" /tɪbet/; /aŋnaizyo/, /sassoku/, /keŋritu/, /tenimotu/, and /hanadayori/. For the word "chapel" /tʃæpəl/, the target segment /æ/ is not part of a tense/lax pair, so the usual method for a vowel which fails to reach 30% correct (percent to give a member of the correct tense/lax pair) cannot be applied. However, the regular percent

---

<sup>4</sup> Furthermore, "hungry" and "unfeeling" do have the same vowel (except for stress) in the first syllable, as well as a nasal in the coda of the first syllable. However, only the identity of the second segment of interest enters into the decision to count this response as correct.

correct data for this word (percent with /æ/) has a negative overall slope, so this word is excluded from the percent correct calculation (as discussed in Section 4.1.3 below). Otherwise, the partially correct measure described above was applied to these words. In the tables reporting the percent correct (%Corr ) results below, Tables 4.1, 4.4, and 4.13, these words are marked with an asterisk, so that the reader can tell which results are evaluated on different criteria from the others.

The reason for substituting this partially correct measure for the usual percent completely correct measure in these cases is that when the listeners never did recognize a segment by the final gate, the area during which the usual percent correct measure shows the most improvement may not be very meaningful. If the improvement is from 9% to 18% correct, this represents an increase of only one listener perceiving the segment correctly, with 9 of 11 listeners still perceiving it incorrectly, for example. The partially correct measure, however, usually reflects a real change in how a large proportion of the listeners perceive the segment. For "ravish," for example, at early gates "rat" and "rap" are common responses, with a variety of others represented. By the final gate, 100% of subjects give responses with at least two features (considering voicing, place, and manner as the "features") correct, with most subjects responding "rabbit." Still, only 18% give responses with a /v/. The shift in responses from a wider variety of possibilities to a concentration on "rabbit" shows that subjects have perceived the voicing and place of the sound. When listeners do not perceive a segment correctly even by the last gate, the area during which the partially correct measure makes the most improvement reflects the point of the signal at which listeners gain enough perceptual cues to perceive distinctive features they did not perceive before then.

For /sassoku/, a measure of three features out of four correct was used, where the features are place, manner, voicing, and length. Since the target transition of /aNnaizyo/ [annaid̚<sub>30</sub>] consists of a mora nasal followed by a non-moraic nasal, at least a mora nasal

was necessary for a response to be counted as partially correct. This is because the issue in perception of the /Nn/ [nn] sequence is when the duration of the nasal becomes long enough for listeners to realize that it is long. In order for it to be long, it must contain a mora nasal. The requirement of a mora nasal in order for a response to be counted as partially correct is not inconsistent with the previously described criterion of evaluating correctness of responses at the sub-phonemic level, because length is the important sub-phonemic characteristic of the /Nn/ [nn] sequence.

#### 4.1.3. Data excluded for anomalous slopes

The procedures described in the previous section result in data which shows the number of responses and the percent correct as a function of gate number for each word. The slope of the linear regression line was calculated for each word for each of the two types of data (#Resp and %Corr). An example of the two types of data for the word "circle" /sɔ:kəl/, with linear regression lines, is shown in Figure 4.1. One would normally expect that as listeners hear more of a word and have more phonetic information available to them, their perception of the segment being gated through would improve. Thus, the percent correct measure (%Corr) should have a positive slope. For the number of responses data (#Resp), a negative slope is expected, because the cohort theory and other models of spoken word recognition (Marslen-Wilson and Welsh 1978, Taft and Hambly 1986, and many others) assume that as more phonetic information becomes available, listeners narrow down the group of words (the cohort) from which they access a lexical item.

For most of the words used for this experiment, the slope of the percent correct data (%Corr) was positive and the slope of the number of responses data (#Resp) negative, as predicted and as shown in Figure 4.1. However, for the percent correct data five English words and one Japanese word had zero or negative slopes; for the number of responses data, 13 English words and 24 Japanese words had zero or positive slopes. (Reasons for

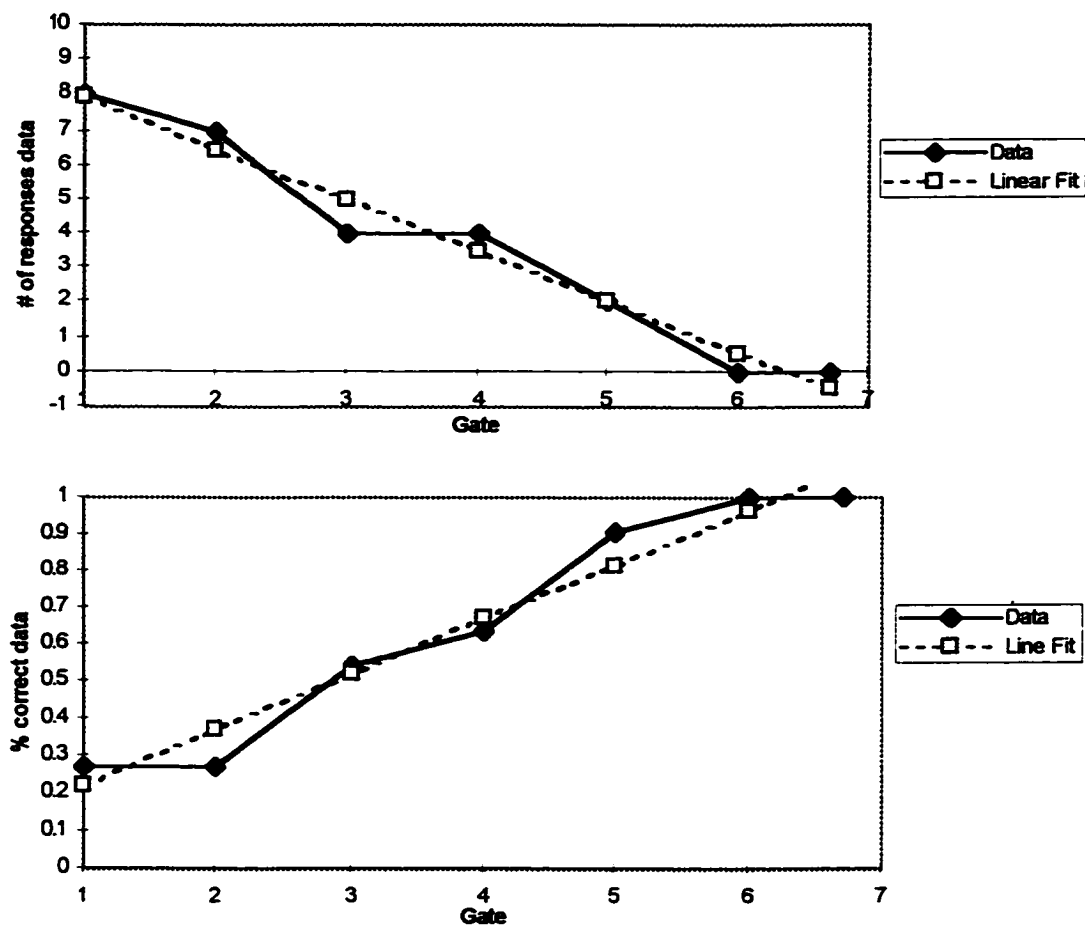


Figure 4.1. Number of responses and percent correct data for the word "circle" /ə:k/, with linear regression lines. The x-axis is gate number, which corresponds to time in 20 ms increments. The y-axis for the number of responses (#Resp) measure is the number of responses given by the group of subjects to hear the stimulus, converted to a 0 to 10 scale. The y-axis for the percent correct (%Corr) measure is the percent of responses to the stimulus which had the second segment of the transition of interest correct, calculated as explained in the text. Note that although the data itself cannot exceed the 0-10 or 0-1 scales, the linear regression lines can.

these discrepancies will be discussed in Section 5.3.2 below.) These words were excluded from further calculations for the type of analysis (%Corr or #Resp) for which they had an anomalous slope.

#### 4.1.4. Fitting of ogival curves

For the number of responses (#Resp) and the percent correct (%Corr) data for each remaining word, an ogival curve (the shape of curve used for categorical perception) was fit to the data. Open response data is usually rather noisy, and this is certainly the case for these results, especially for the number of responses (#Resp) data. Therefore, rather than attempt to define a recognition point based on the data itself, I believe it is better to fit a curve to the data and locate a point or an area of recognition based on the curve. An ogival shape was chosen as the type of curve to fit to the data because ogival curves are flat on both ends with a more rapidly rising or falling area in the middle. The prediction regarding use of dynamic cues is that if listeners are using primarily dynamic cues to perceive speech, their recognition of a segment or a word should improve rapidly near areas of great spectral change, but improve less quickly in steady state areas. This pattern would resemble an ogival curve if the beginning and end of the curve are placed at steady state areas, with the middle being the area of great spectral change.

The equation for the ogival curve which was fit to the data is shown in (1),

$$(1) y = \alpha + (\beta - \alpha) \cdot \text{NormSDist}(\gamma(x - \delta))$$

where the function NormSDist returns the standard normal cumulative distribution (the integral of the standard normal distribution), and is defined by the equation in (2).

$$(2) \text{NormSDist}(T) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^T e^{-\frac{t^2}{2}} \cdot dt$$

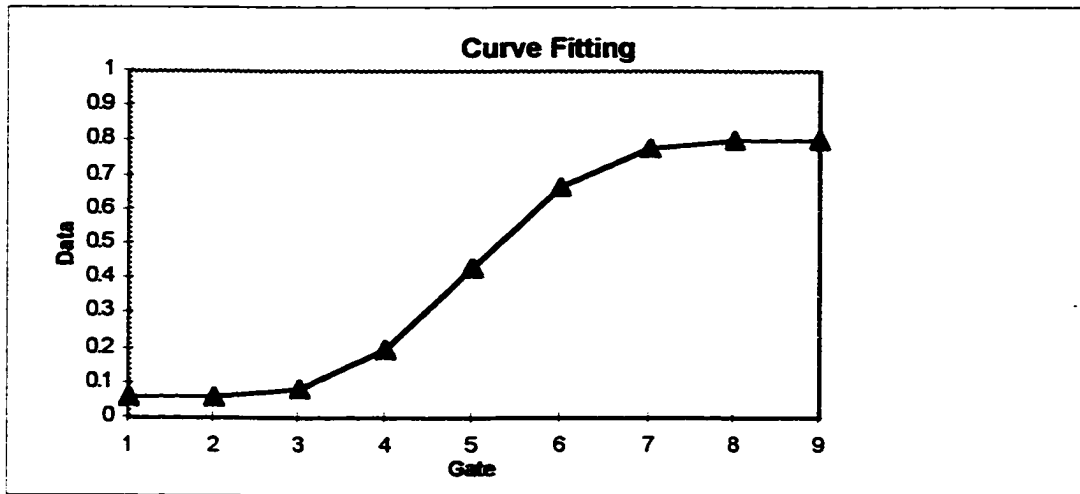
where t = time.

Equation (2) accounts for the ogival shape of the curve, and the parameters  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\delta$  in equation (1) control the exact position of the curve relative to the data points. Five examples of possible curves are shown in Figure 4.2 in order to demonstrate the manipulation of each parameter. The parameter  $\alpha$  controls the height (the value on the y axis) of the asymptote of the left side of the curve. The parameter  $\beta$  controls the height of the asymptote of the right side of the curve.  $\gamma$  controls the steepness of the rise or fall of the curve, and  $\delta$  controls where along the x axis (time, or gate) the rise or fall of the curve occurs.

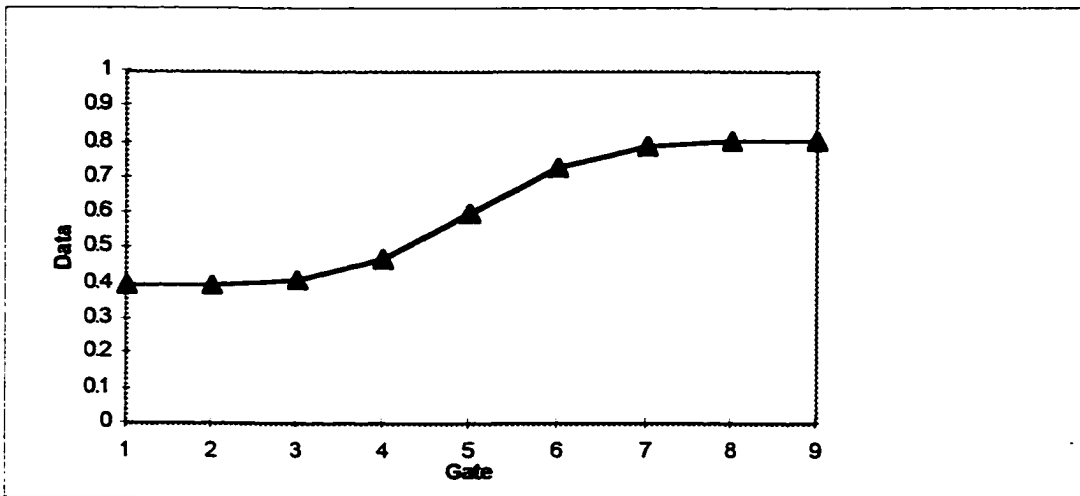
The linear regression line for each set of data, along with its least squares error, was also calculated. The ogival curves were fit to the data using an interactive method which allows the user to manipulate each of the four parameters while watching the resulting curve and observing the change in the least squares error for the curve. Using this method, I fit an approximate curve to the data by eye, and then manipulated the four parameters to minimize the error of the curve fit. This method proved to be more reliable in finding a curve with a low error than several automated methods which use the curve fitting functions of commercially available statistics packages. When comparing curves fit to the same data by the automated methods and by the interactive method, the curve produced by the interactive method frequently had a lower error than the curves produced by the automated methods. Figure 4.3 shows the screen used for fitting the curves by the interactive method.

#### 4.1.5. Exclusion of linear data

The least squares errors for the linear fit and the ogival fit were compared for each set of data. If the linear fit had a lower error than the ogival fit, the data for that word was excluded from further calculations (for either the number of responses (#Resp) or percent

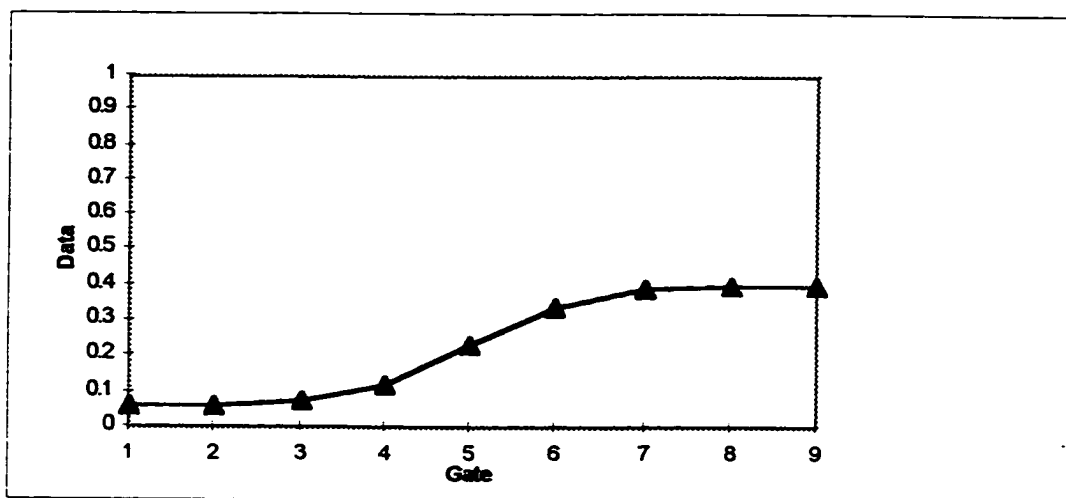


$$\alpha = 0.06 \quad \beta = 0.805 \quad \gamma = 0.9 \quad \delta = 5$$

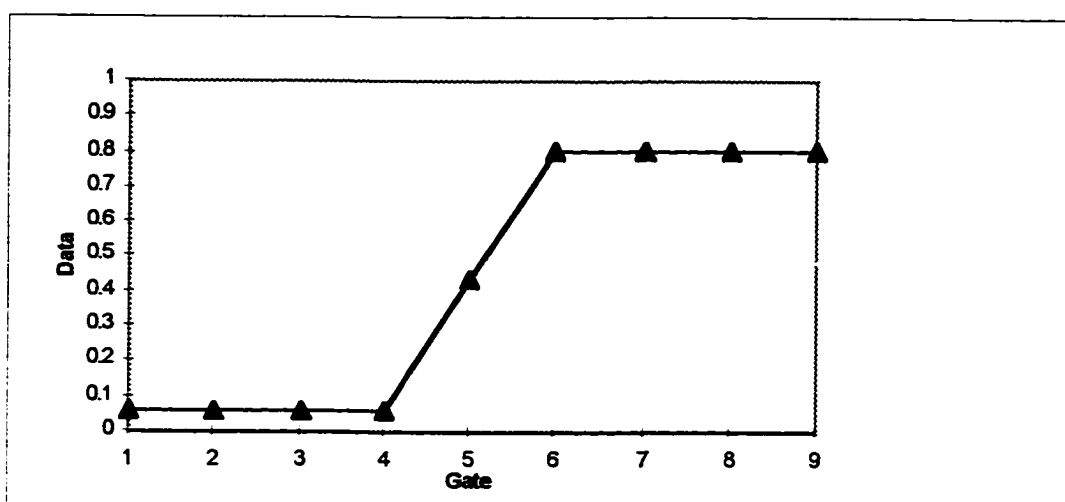


$$\alpha = 0.40 \quad \beta = 0.805 \quad \gamma = 0.9 \quad \delta = 5$$

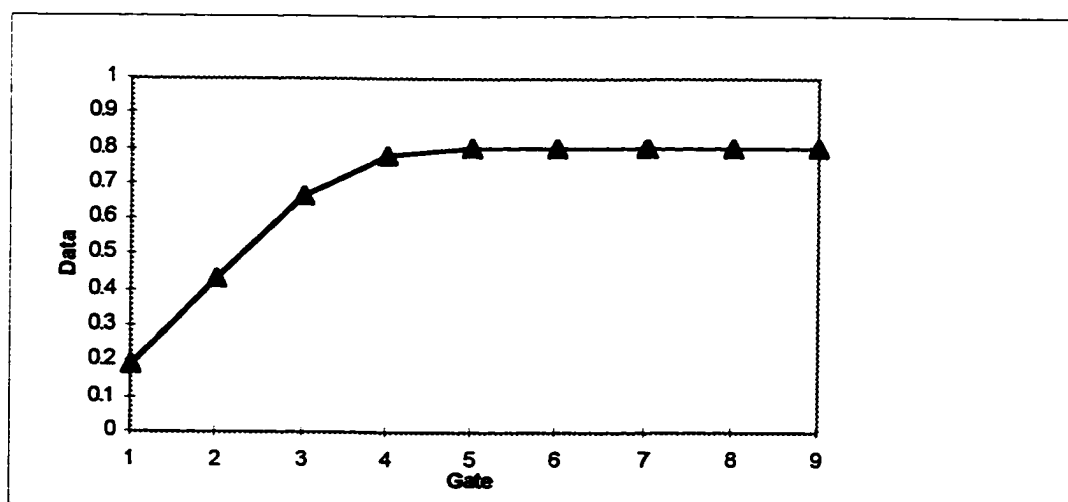
Figure 4.2. The first panel shows an example of an ogival curve fitting the equation given in the text. Each subsequent panel shows the effect of changing one parameter of the equation: the second panel shows the same curve with  $\alpha$  (left asymptote) raised. The third panel (top of next page) shows the original curve with  $\beta$  (right asymptote) lowered. The fourth panel shows  $\gamma$  (steepness) increased. The fifth panel shows the location of the rise in the curve moved to the early gates ( $\delta$  lowered), so that only part of the curve appears. One should note that in the fourth panel, where the slope is steep, the fitted curve is plotted only at the points on the x-axis for which perceptual data exists. That is, the equation for this curve has the same ogival shape as the other sample curves do, but in this graph it is only plotted at nine points, and the entire rise takes place between three of these points, so the curve appears to have sharp corners.



$\alpha = 0.06$      $\beta = 0.40$      $\gamma = 0.9$      $\delta = 5$







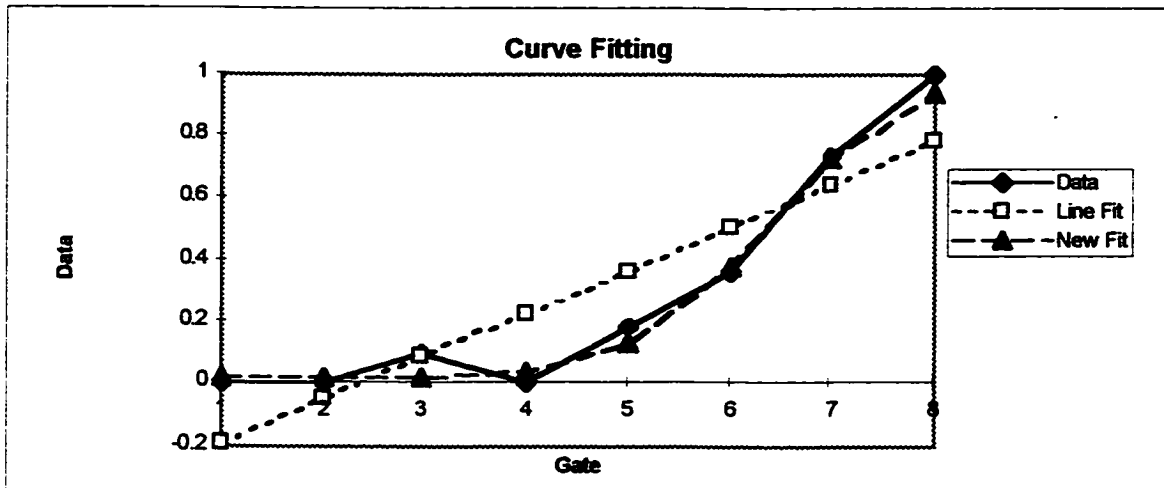
$\alpha = 0.06$      $\beta = 0.805$      $\gamma = 6.40$      $\delta = 5$



$\alpha = 0.06$      $\beta = 0.805$      $\gamma = 0.9$      $\delta = 2$



Param	Est	Min	Max	Scroll bars	Scroll	Current
Alpha	0	0	1		2	0.02
Beta	1	0.2	1.2		82	1.02
Gamma		0	10		9	0.9
Delta	4.5	0	10		64	6.4



Line: 0.445  
 Ogive: 0.116  
 Winner: Ogive  
 Beg. max. 6

Figure 4.3. The screen used for fitting curves to data. The scroll bars allow the user to manipulate each parameter while watching the effect on the ogival error, below the graph, and comparing this to the linear error. The beginning of the area of maximal slope also appears. Parts of the spreadsheet which are not displayed here calculate the best linear fit and the linear and ogival errors.

correct (%Corr), whichever was linear). This is because, if improvement in the perceptual measure is linear, there is no area during which perception of the segment (for %Corr) or recognition of the word (for #Resp) becomes accurate more quickly than at other times. For the percent correct data (%Corr), five English words and two Japanese words were better fit by a line than by an ogival curve. For the number of responses (#Resp) data, this was the case for 12 English words and three Japanese words. This means that, after exclusion of words with anomalous slopes described above, 95.9% (117/122) of the English words and 97.3% (73/75) of the Japanese words were better fit by an ogival curve than by a line for their %Corr data. For the #Resp data, 89.5% (102/114) of English words and 94.5% (52/55) of Japanese words were more ogival than linear. Thus, a very high percentage of words which have an overall decrease in the number of responses (#Resp) or an increase in percent correct (%Corr) have this decrease or increase concentrated in some part of the signal: they rise or fall more quickly at some parts of the signal than at others.

The word "diploma" /dɪpləˈmɑ/ was also excluded from all calculations because the initial syllable was so short that it had only two gates, and therefore its data must always be linear. Exclusion of this word, the data with anomalous slopes, and the data best fit by a line left, for the percent correct (%Corr) data, 117 English words and 73 Japanese words, and for the number of responses (#Resp) data, 102 English words and 52 Japanese words. (The entire experiment included 128 English words and 76 Japanese words.) Appendix B shows the graphs of each set of data, with their regression lines. Except for the words with anomalous slopes, the fitted ogival curve is also shown, and the least squares errors for both the linear and the ogival fits are included. The area of maximal slope of the fitted curve (which will be discussed below) is also indicated in the graphs.

#### 4.2. Comparison of location of maximal change in the perceptual measures to location of point of maximal spectral change

For each of the remaining sets of data, the two contiguous points of the fitted curve between which the slope of the curve was the greatest were located. (The maximal negative slope was found for number of responses (#Resp) data and the maximal positive slope for percent correct (%Corr) data.) If the curve fits the data points closely, this is the area during which the number of different responses given decreases the most, or during which the percent of listeners giving responses with the second segment of the transition of interest correct increases the most. This area was assumed to be the 20 ms window during which listeners' perception of the segment or recognition of the word improved the most. This method diverges from the usual approach to evaluating such data in two ways, first in the use of the fitted curve instead of the data itself for determining an area of recognition, and second in the use of the area of greatest slope.

Figure 4.4 demonstrates the importance of evaluating the perceptual data by means of fitted curves, rather than directly from the data. Figure 4.4 shows the #Resp data for the Japanese word /kiNtyoo/ [kintʃoo] 'nervousness.' As is usually the case with open response data, this data is somewhat noisy. If one were to look for the area of most change in the number of responses (#Resp) using the data itself, one would find that the area between the second and third gates and the area between the seventh and eighth gates have an equally great negative slope. One would have to conclude either that it is impossible to tell whether the area during which listeners make the most progress toward recognizing the word is between the second and third or between the seventh and eighth gates, or that there are two areas during which listeners make an equal amount of progress toward word recognition.

However, visual examination shows that the data, while noisy, is relatively flat through the sixth gate, and then falls from the sixth through ninth gates. Using only the data, there is no principled way to exclude the area of the second to third gate as the fall.

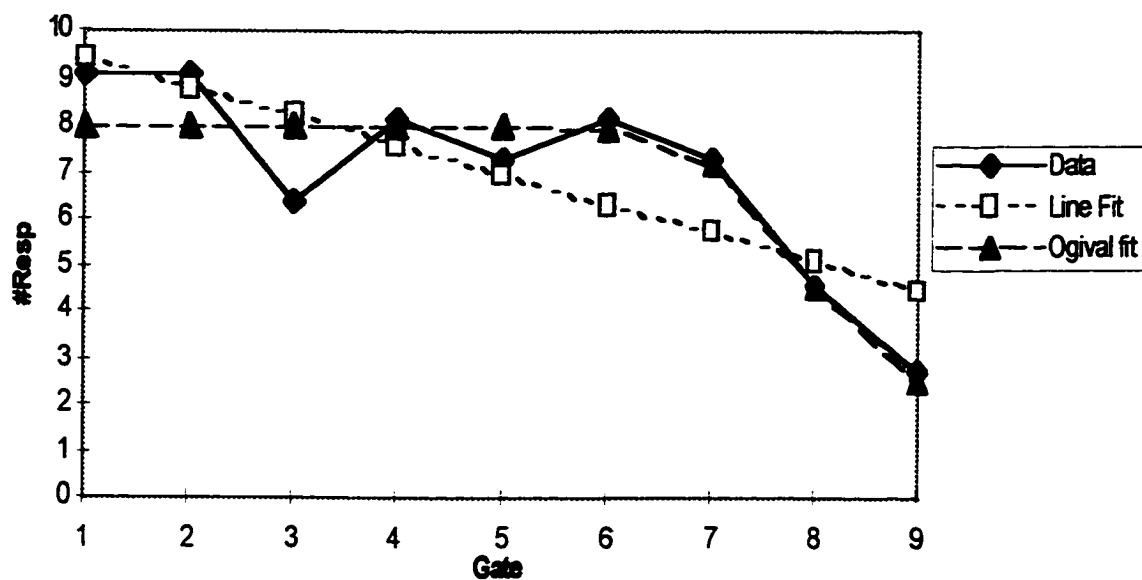


Figure 4.4. Data, best linear fit, and fitted ogival curve for the number of responses data for /kiNtyoo/ [kintʃoo] 'nervousness.' The fitted ogival curve smoothes out the noisiness of the raw data and allows for location of the area of maximal change, which is between gates 7 and 8. The least squares error of the linear regression line is 3.634, while the least squares error of the fitted ogival curve is 2.400.

The fitted ogival curve, however, has only one area of maximal negative slope, which is between the seventh and eighth gates. The fitted curve treats the falls and rises of the early gates as noise, and therefore has the flat left tail of the curve at a level intermediate between the early data points. It fits the later data points quite closely. The fitted ogival curve has a lower error than the linear fit, showing that this data can be more accurately represented as falling quickly in some areas and slowly in others than as a line, which would have the fall evenly distributed. Furthermore, the fitted curve makes it possible to locate a unique area of maximal slope, which should correspond to the time period during which listeners' recognition of the word became accurate most quickly. Such cases, in which fitted curves are essential for interpreting noisy data, are quite common in the results of this experiment.

One should also note in this figure that it is possible for the fitted curve to have only one flat end of the curve present in the data. The fall of the curve is at the later gates, and the flat right tail of the curve is beyond the end of the data. This type of curve is very common in this data: perceptual cues for segments are located at various time points in the signal relative to the segment itself, and recognition of a segment may begin before the preceding segment or not be successful until after a segment ends. Furthermore, in a transition from an inherently short segment to an inherently long one, such as from a voiceless unaspirated stop to a vowel, only one or two gates are likely to fall within the first segment. This accounts for the fact that the steepest fall of the fitted curve is often not in the middle of the data points.

For data which is represented in terms of percent of answers correct, it is common in the literature to set an arbitrary cut-off point, such as 80% correct or 50% correct, as the recognition point. This is the usual method for categorical perception experiments, for example, particularly if they use a two way forced choice response. I chose to use the area of maximal slope of the fitted curve for the percent correct (%Corr) measure as well as for the number of responses (#Resp) measure, rather than setting such an arbitrary recognition point, for several reasons. First, it is desirable to have the area chosen as the recognition

area or point be readily comparable between the two perceptual measures (#Resp and %Corr), so the same method should be used for locating this recognition area for both. Second, because of the wide variety of transitions and environments used, the %Corr data often does not cover the entire 0% to 100% range. Some segments which are perceived quite early relative to their acoustic onsets (nasals, for example), may already be at 80% correct at the first gate, and increase to 100% correct. Some other segments (e.g. /v/, some vowels) still are not recognized accurately at the final gate, so that the %Corr measure changes from 0% to 30% correct. With this degree of variation, the point at which some arbitrary cut-off is crossed would be meaningless for many words. Such a method is more useful for categorical perception data because categorical perception usually involves stimuli on a continuum from one phoneme to another, so that responses must range between 0% and 100%.

For these reasons, locating the area of maximal slope was considered to be the best choice for the data in this experiment. This area of maximal slope for each curve was compared to the location of the point(s) of maximal spectral change ( $D_{\max}$  points) obtained from the acoustic data, which were described in Chapter 3.

#### 4.2.1. Observed results

For each word, the location of the 20 ms<sup>5</sup> area of maximal slope for both the number of responses measure (#Resp) and the percent correct measure (%Corr) was compared to the location of the  $D_{\max}$  point or points. If a  $D_{\max}$  point falls within the 20 ms area of maximal slope of one of the perceptual measures, this shows that the most improvement in perception is made at the same time as the greatest spectral change takes place, since the smallest time unit that can be examined in the perceptual measure is 20 ms, the gating interval. If the beginning of the 20 ms area of maximal slope fell within 5 ms

---

<sup>5</sup> This area can also be less than 20 ms, if it is between the last two gates of the word and the endpoints of the last two gates are separated by less than 20 ms. However, this area can never be more than 20 ms long.

after the  $D_{\max}$  point, this was also considered as being close enough to show the perceptual change happening at the same time as the spectral change. As discussed in Section 3.2.2, the peak of the measure  $D$  frequently occurs 5-10 ms earlier than the acoustic event it reflects, especially in the case of voiceless stop bursts. Thus, if the area of maximal slope of the perceptual measure begins very shortly after the measured  $D_{\max}$  point, the acoustic event with which the  $D_{\max}$  point is associated may actually fall within the area of maximal slope. Even when the  $D_{\max}$  point is not early relative to the acoustic change, listeners might need to hear a few milliseconds after the point of maximal change to use the perceptual cues available there. The results of the comparison between perceptual and acoustic data are shown in Table 4.1.

Table 4.1. Comparison of location of  $D_{\max}$  point(s) and location of area of maximal slope of the two perceptual measures. All measurements are in milliseconds from the beginning of the word. Where there is more than one local maximum of  $D$ , the time of the largest appears first. Maximal slopes marked with an asterisk are those for which %Corr never exceeded 30%, so that the percent partially correct measure was used.

No.	Word	Trans.	$D_{\max}$	#Resp. max. slope	Does #Resp. max slope overlap or begin within 5ms after $D_{\max}$ ?	%Corr. max. slope	Does %Corr. max slope overlap or begin within 5ms after $D_{\max}$ ?
English							
1	tip	tɪ	75	42-62	No	62-82	Yes
2	stiff	tɪ	150, 180	188-208	No	168-88	Yes
3	Tibet	tɪ	15	0 slope	N/A	37-57*	No
4	petition	tɪ	115, 160	139-59	No	199-212	No
5	attic	tɪ	125	145-65	No	125-45	Yes
6	custom	kʌ	50	17-37	No	57-77	No
7	skull	kʌ	205, 240	242-262	Yes	262-82	No
8	accompany	kʌ	80, 100	129-47	No	129-47	No
9	caboose	kə	15	14-34	Yes	34-53	No
10	academic	kə	105, 125	128-46	Yes	128-46	Yes
11	duck	dʌ	80	positive slope	N/A	40-60	No
12	citizen	ɪt	185, 165	189-95	Yes	189-95	Yes
13	fitness	ɪt	80, 120	86-106	No	106-26	Yes-smaller $D_{\max}$
14	Italian	ɪt	70, 50	114-38	No	34-54	Yes-smaller $D_{\max}$

No.	Word	Trans.	D <sub>max</sub>	#Resp. max. slope	Does #Resp. max slope overlap or begin within 5ms after D <sub>max</sub> ?	%Corr. max. slope	Does %Corr. max slope overlap or begin within 5ms after D <sub>max</sub> ?
15	committee	ɪt	225, 245	242-62	Yes-smaller D <sub>max</sub>	242-62	Yes-smaller D <sub>max</sub>
16	unity	ɪt	255, 235	251-269	Yes	251-69	Yes
17	bucket	ʌk	125, 85	121-41	Yes	121-41	Yes
18	mechanical	ək	115, 85	128-48	No	68-88	Yes
19	indicate	ək	195, 165	196-216	Yes	linear	N/A
20	induction	ʌk	245	positive slope	N/A	negative slope	N/A
21	muddy	ʌd	150, 170	linear	N/A	170-83*	Yes
22	cadenza	əd	120, 80	109-29	Yes	89-109	No
23	medicine	mɛ	40, 55	25-45	Yes	45-65	Yes
24	immense	mɛ	90	97-117	No	117-37*	No
25	remedy	ɛm	120	0 slope	N/A	101-21	Yes
26	attempt	ɛm	275	314-23	No	274-94	Yes
27	negative	nɛ	65	70-90	Yes	70-90	Yes
28	tenants	ɛn	135	positive slope	N/A	129-149	Yes
29	saddle	sæ	100, 115	linear	N/A	124-44	No
30	master	æs	180	154-74	No	174-94	Yes
31	Zachary	zæ	55	81-101	No	61-81	No
32	asthma	æz	105	30-50	No	110-130	Yes
33	shell	ʃɛ	140	182-92	No	negative slope*	N/A
34	session	ɛʃ	220	253-73	No	233-53	No
35	fees	fi	60, 15	36-56	No	36-56	No
36	unfeeling	fi	190	181-201	Yes	181-201	Yes
37	leaf	if	250, 230	295-315	No	295-315	No
38	relief	if	360	346-66	Yes	366-86	No
39	vacuum	væ	50	75-95	No	35-55	Yes
40	ravish	æv	135	149-69	No	149-69*	No
41	trail	reɪ	125, 160	161-81	Yes	161-81	Yes
42	fair	eɪr	195, 155, 165	112-32	No	132-52	No
43	lever	le	75	positive slope	N/A	111-31	No
44	elevator	ɛl	60	19-39	No	19-39	No
45	yellow	je	60	85-105	No	65-85*	Yes
46	watch	wa	75	50-70	No	70-90	Yes



No.	Word	Trans.	D <sub>max</sub>	#Resp. max. slope	Does #Resp. max slope overlap or begin within 5ms after D <sub>max</sub> ?	%Corr. max. slope	Does %Corr. max slope overlap or begin within 5ms after D <sub>max</sub> ?
47	chapel	tʃæ	85	85-105	Yes	negative slope	N/A
48	latches	ætʃ	175, 145	103-23	No	183-203	No
49	jump	dʒʌ	35	18-38	Yes	18-38	Yes
50	judge	ʌdʒ	210, 235, 265	167-87	No	167-87	No
51	bent	nt	255, 215	252-72	Yes	252-72	Yes
52	sentiment	nt	210	210-30	Yes	210-30	Yes
53	reinterpret	nt	210, 190	216-36	No	196-216	Yes
54	band	nd	295, 275	292-312	Yes	292-312	Yes
55	wander	nd	280	0 slope	N/A	linear	N/A
56	recondition ed	nd	255	206-26	No	246-63	Yes
57	axe	ks	245	285-305	No	265-85	No
58	hacksaw	ks	165, 190	216-36	No	196-216	No
59	unacceptable	ks	165	200-20	No	180-200	No
60	cats	ts	295	320-40	No	320-40	No
61	Betsy	ts	135	positive slope	N/A	153-73	No
62	stop	st	130, 110	132-52	Yes	132-52	Yes
63	based	st	400, 375	linear	N/A	397-417	Yes
64	pastime	st	290, 260	278-98	Yes	278-98	Yes
65	skate	sk	160, 100	147-67	Yes	167-87	No
66	mask	sk	450, 405	linear	N/A	450-70	Yes
67	discount	sk	180, 150	173-93	Yes	133-53	Yes
68	train	tr	75	38-58	No	linear	N/A
69	string	tr	160, 195	linear	N/A	160-80	Yes
70	Detroit	tr	105, 210	linear	N/A	111-31	No
71	crops	kr	80	73-93	Yes	13-33	No
72	scrap	kr	180, 215	linear	N/A	200-20	Yes
73	acrobat	kr	120, 165	172-86	No	112-32	Yes
74	drop	dr	30	62-74	No	62-74	No
75	groan	gr	70, 50, 85	59-79	Yes	59-79	Yes
76	plain	pl	95, 40	39-59	Yes	39-59	Yes
77	split	pl	230, 260	236-56	No	236-56	No
78	twelve	tw	50	80-90	No	40-60	Yes

No.	Word	Trans.	$D_{max}$	#Resp. max. slope	Does #Resp. max slope overlap or begin within 5ms after $D_{max}$ ?	%Corr. max. slope	Does %Corr. max slope overlap or begin within 5ms after $D_{max}$ ?
79	court	rt	295, 245	207-27	No	227-47	Yes
80	cork	rk	270, 205	294-314	No	294-314	No
81	help	lp	310, 235	linear	N/A	289-309	No
82	fans	nz	415	positive slope	N/A	421-41	No
83	dance	ns	250	239-59	Yes	239-59	Yes
84	fancy	ns	220	238-58	No	218-38	Yes
85	uncon- cealed	ns	240	positive slope	N/A	241-61	Yes
86	snow	sn	170, 130	166-86	Yes	166-86	Yes
87	Disney	zn	185	155-75	No	175-95*	Yes
88	farm	rm	355, 260	252-72	Yes	252-72	Yes
89	corn	m	285	260-80	No	0 slope	N/A
90	film	lm	155	positive slope	N/A	0 slope	N/A
91	ranch	ntʃ	285, 210	269-89	Yes	309-29	No
92	flash	fl	45	56-76	No	56-76	No
93	fragile	fr	45	25-45	Yes	45-65	Yes
94	sleep	sl	135	139-59	Yes	139-59	Yes
95	Iceland	sl	255	205-25	No	linear	N/A
96	swan	sw	160, 125	100-20	No	100-120	No
97	golf	lf	230	positive slope	N/A	301-21	No
98	wharf	rf	195	222-42	No	262-82	No
99	false	ls	235	284-304	No	264-84	No
100	calcium	ls	175	linear	N/A	185-205	No
101	cultural	ʃtʃ	160	187-207	No	147-67	Yes
102	marginal	rdʒ	185, 160	193-213	No	173-93	Yes
103	optical	pt	145	138-58	Yes	138-58	Yes
104	pact	kt	330, 265, 295	337-57	No	317-37	Yes
105	coughs	fs	300	347-67	No	327-47	No
106	nerves	vz	400	459-79	No	399-419	Yes
107	amnesty	mn	175	183-203	No	183-203	No
108	garlic	rl	130	138-58	No	118-38	Yes
109	biopsy	a'a	190	161-81	No	161-81	No
110	biography	a'a	100	75-95	No	135-55	No

No.	Word	Trans.	D <sub>max</sub>	#Resp. max. slope	Does #Resp. max slope overlap or begin within 5ms after D <sub>max</sub> ?	%Corr. max. slope	Does %Corr. max slope overlap or begin within 5ms after D <sub>max</sub> ?
111	biotech	a <sup>o</sup> w	165	positive slope	N/A	156-76	Yes
112	eon	ia	35, 130	156-76	No	156-76	No
113	diagonal	a <sup>i</sup> æ	85, 100	182-98	No	122-42	No
114	react	iæ	200	183-203	Yes	123-43	No
115	tiger	ta <sup>i</sup>	120, 85	linear	N/A	linear	N/A
116	bite	a <sup>i</sup> t	180	129-49	No	169-89	Yes
117	data	de <sup>i</sup>	65	41-61	No	41-61	No
118	fade	e <sup>i</sup> d	320, 300	231-51	No	231-51	No
119	doubt	a <sup>w</sup> t	285, 260	linear	N/A	193-213	No
120	soybean	o <sup>b</sup>	300, 260	185-205	No	245-65	Yes
121	toad	to <sup>w</sup>	65	120-140	No	120-140	No
122	oats	o <sup>w</sup> t	255, 180	95-115	No	75-95	No
123	courage	kə	80	140-59	No	100-20	No
124	circle	ək	240, 220	linear	N/A	208-28	Yes-smaller D <sub>max</sub>
125	button	t <sup>n</sup>	120	119-39	Yes	119-39	Yes
126	beetle	t <sup>l</sup>	140	147-67	No	127-47	Yes
127	apple	pl	195, 220	204-24	Yes	204-24	Yes

## Japanese

1	todana	[to]	25	0 slope	N/A	20-40	Yes
2	tatoe'ru	[to]	165, 190	185-205	Yes-smaller D <sub>max</sub>	185-205	Yes-smaller D <sub>max</sub>
3	ka'to	[to]	190	227-47	No	207-27	No
4	kakari'iN	[ka]	60	positive slope	N/A	21-41	No
5	hakama'	[ka]	150, 190	187-207	Yes-smaller D <sub>max</sub>	187-207	Yes-smaller D <sub>max</sub>
6	sya'kai	[ka]	255, 285	268-88	Yes-smaller D <sub>max</sub>	268-88	Yes-smaller D <sub>max</sub>
7	dama'ru	[da]	15	positive slope	N/A	15-35	Yes
8	midare'ru	[da]	110	positive slope	N/A	122-42	No
9	ku'da	[da]	125	131-51	No	131-51	No
10	hotoke'	[ot]	100, 55	53-73	Yes-smaller D <sub>max</sub>	53-73	Yes-smaller D <sub>max</sub>
11	himoto'	[ot]	375, 300	positive slope	N/A	373-93	Yes

No.	Word	Trans.	D <sub>max</sub>	#Resp. max. slope	Does #Resp. max slope overlap or begin within 5ms after D <sub>max</sub> ?	%Corr. max. slope	Does %Corr. max slope overlap or begin within 5ms after D <sub>max</sub> ?
12	hakobu	[ak]	150, 105	153-73	Yes	73-93	No
13	hatake	[ak]	330, 275	233-53	No	233-53	No
14	ha'yaku	[ak]	340, 245	198-218	No	218-38	No
15	kadai	[ad]	140, 120	131-51	Yes	131-51	Yes
16	hanada'yo ri	[ad]	260	positive slope	N/A	232- 52*	No
17	ka'nada	[ad]	305, 270	241-61	No	221-41	No
18	megumi	[me]	40	77-90	No	37-57	Yes
19	tomeru	[me]	145	143-63	Yes	123-43	No
20	nemui	[ne]	30	positive slope	N/A	linear	N/A
21	kemuri	[em]	120	104-124	Yes	104-24	Yes
22	tabemo'no	[em]	215	188-208	No	168-88	No
23	teni'motu	[en]	115	positive slope	N/A	120- 37*	Yes
24	soda'tu	[so]	35, 80	113-24	No	73-93	Yes-smaller D <sub>max</sub>
25	zabu'toN	[za]	40	38-58	Yes	38-58	Yes
26	syabe'ru	[ja]	95, 115	103-23	Yes	103-23	Yes
27	hokeN	[ho]	45	0 slope	N/A	42-62	No
28	zyosei	[os]	120	95-115	No	75-95	No
29	kazari	[az]	125	136-54	No	136-54	No
30	basyo	[aʃ]	120	112-32	Yes	112-32	Yes
31	gohoo	[oh]	115	137-57	No	157-77	No
32	wahuku	[aɸ]	150, 125	positive slope	N/A	114-34	Yes-smaller D <sub>max</sub>
33	dohyoo	[oɕ]	130	142-62	No	162-81	No
34	harada'tu	[ra]	100	87-107	Yes	107-27	No
35	yubi'	[ju]	55	150-70	No	150-70	No
36	kara'i	[ar]	135, 155	144-57	Yes-smaller D <sub>max</sub>	144-57	Yes-smaller D <sub>max</sub>
37	huyoo	[uj]	100	117-29	No	117-29	No
38	mawari	[aw]	110	119-31	No	119-31	No
39	tyazuke	[tʃa]	50	linear	N/A	55-75	Yes
40	zyokyo'oz yu	[dʒo]	45	positive slope	N/A	17-37	No
41	mati'	[atʃ]	210, 150	116-36	No	116-36	No
42	tozi'ru	[odʒ]	135, 150	132-52	Yes	132-52	Yes
43	haNtai	[nt]	300, 220	284-304	Yes	244-64	No

No.	Word	Trans.	D <sub>max</sub>	#Resp. max. slope	Does #Resp. max slope overlap or begin within 5ms after D <sub>max</sub> ?	%Corr. max. slope	Does %Corr. max slope overlap or begin within 5ms after D <sub>max</sub> ?
44	kaNdoo	[nd]	305, 285	positive slope	N/A	negative slope	N/A
45	teNkiN	[ŋk]	295, 225	304-24	No	264-84	No
46	kaNzeN	[nz]	280	302-22	No	282- 302	Yes
47	seNsoo	[ns]	305	positive slope	N/A	303-23	Yes
48	keNritu	[nr]	265	positive slope	N/A	229- 49*	No
49	koNyaku	[ŋj]	235	linear	N/A	236-56	Yes
50	kiNtyoo	[ntʃ]	290, 255	290-310	Yes	310-30	No
51	sukunai	[sk]	105, 150	positive slope	N/A	131-51	Yes-smaller D <sub>max</sub>
52	sikaku	[sʰk]	125, 150	152-72	Yes-smaller D <sub>max</sub>	152-72	Yes-smaller D <sub>max</sub>
53	kitamuki	[kʰt]	105, 80	119-28	No	119-28	No
54	kokutetu	[kt]	210, 250	213-33	Yes	linear	N/A
55	kyaku	[kj]	60	37-57	No	17-37	No
56	dakyoo	[kj]	190, 240	179-99	Yes	179-99	Yes
57	hyoo	[çj]	105	133-52	No	93-113	Yes
58	ryokaN	[rj]	15	positive slope	N/A	14-34	Yes
59	mottaina'i	[tt]	265	255-75	Yes	195-215	No
60	sakka	[kk]	380	330-50	No	310-30	No
61	sassoku	[ss]	265	232-52	No	232-52*	No
62	hassya	[ʃʃ]	270, 230	positive slope	N/A	270-89	Yes
63	teNmetu	[mm]	190	positive slope	N/A	171-91	Yes
64	aNnaizyo	[mn]	195, 145, 175	positive slope	N/A	178-97*	Yes
65	tootyaku	[oo]	125, 105	105-25	Yes	85-105	Yes-smaller D <sub>max</sub>
66	keigo	[ee]	120, 175	152-72	No	132-52	No
67	syuukaN	[uu]	220, 245	255-73	No	215-35	Yes
68	haori	[ao]	160, 200	110-30	No	150-70	Yes
69	siatu	[ia]	250	207-27	No	207-27	No
70	kaeri'miti	[ae]	115	positive slope	N/A	151-71	No
71	taiko	[ai]	115	127-47	No	87-107	No
72	koibito	[oi]	105	151-70	No	131-51	No

No.	Word	Trans.	D <sub>max</sub>	#Resp. max. slope	Does #Resp. max slope overlap or begin within 5ms after D <sub>max</sub> ?	%Corr. max. slope	Does %Corr. max slope overlap or begin within 5ms after D <sub>max</sub> ?
73	teNiN	[ēi]	195, 295, 320	210-30	No	210-30	No
74	hiN	[ij]	225, 255	305-25	No	185-205	No
75	maNne'N-hitu	[an]	140, 195	positive slope	N/A	168-88	No
76	seNmoN	[em]	205	linear	N/A	194-214	Yes

Excluding the data with anomalous slopes or a linear rather than ogival fit, I calculated the number and percentage of the remaining words for which the area of maximal change in the perceptual measure surrounds a D<sub>max</sub> point. These results appear in Table 4.2.

Table 4.2. Number and percentage of words for which the area of maximal change in the fitted curve surrounds a D<sub>max</sub> point, for each perceptual measure and each language. Number of words is shown relative to the number of words for that category which are more ogival than linear and do not have anomalous slopes.

Type of data	English		Japanese	
	Proportion	Percentage	Proportion	Percentage
#Resp	36/102	35.3%	22/52	42.3%
%Corr	63/117	53.8%	34/73	46.6%

The percent correct (%Corr) results for both languages have the area of maximal slope of the perceptual results surrounding a D<sub>max</sub> point (or beginning within 5 ms after it) for approximately half of the data. Figures 4.5 and 4.6 show examples in which the area of maximal perceptual change (for both perceptual measures) surrounds a D<sub>max</sub> point for the English word "band" /bænd/ and the Japanese word /basyo/ [baʃo] 'place.' The number of responses (#Resp) results for both languages, however, have the area of maximal perceptual change surrounding or beginning immediately after a D<sub>max</sub> point in only thirty-five to forty-two percent of the cases.

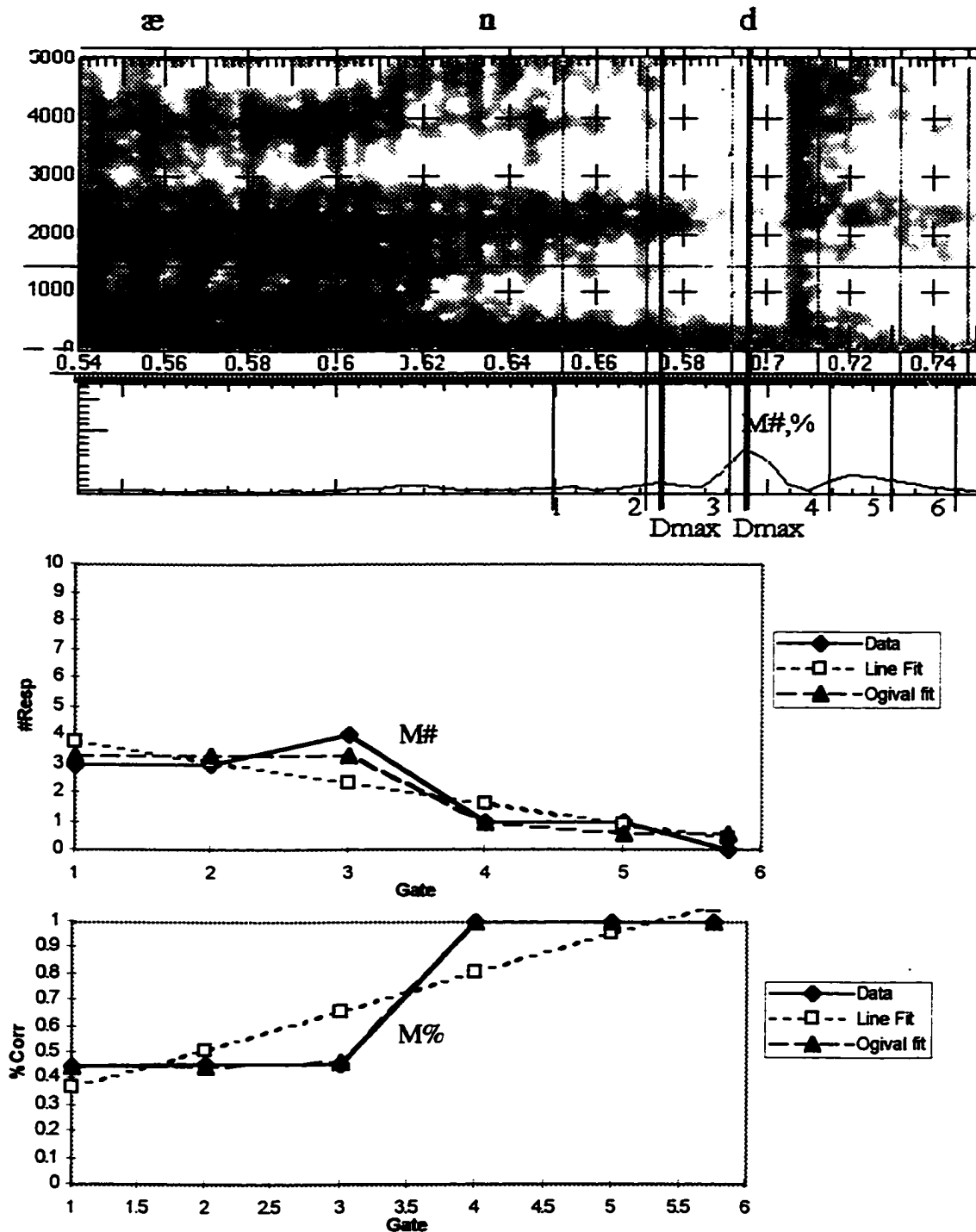


Figure 4.5. Spectrogram, measure D, and the two perceptual measures for the word "band" /bænd/. For this and subsequent such figures, the numbered vertical lines in the spectrogram and measure D field are the endpoints of each gate. Darker vertical lines are  $D_{max}$  points. Area of maximal change of the #Resp measure is indicated in the graph of the perceptual data and in the measure D field with an area marked M#, area of maximal change of %Corr with an M%. In this word, the area of maximal change is in the same location for both perceptual measures, and surrounds the  $D_{max}$  point.

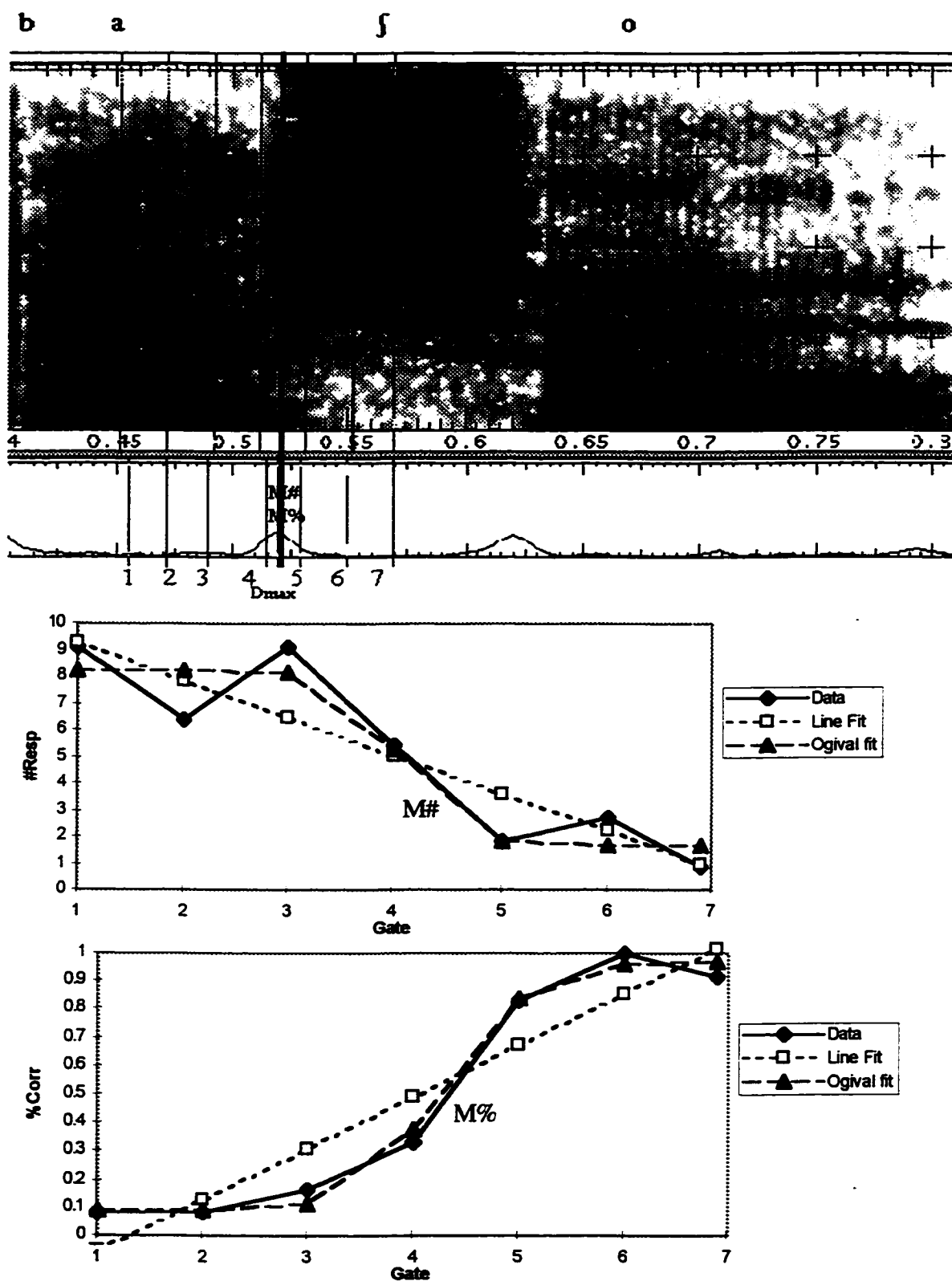


Figure 4.6. Spectrogram, measure D, and the two perceptual measures for /basyo/ [baʃo] 'place.' Area of maximal change for both perceptual measures (M#, M%) surrounds the  $D_{max}$  point.



I report no statistical tests of the difference between the two languages because the words used for each language were chosen to represent the possible transitions of that language, and the word lists are therefore strongly influenced by the phonology of the language. As will be discussed in section 4.3 below, the type of transition (e.g. vowel to fricative, consonant to vowel, etc.) has a strong effect on the likelihood of the area of maximal slope surrounding the  $D_{\max}$  point. Thus, the differences between the two languages for the overall results presented here are more likely to reflect language specific biases of the word lists than actual differences in subjects' behavior for equivalent stimuli. In particular, the lower result for the percent correct data for Japanese than for English may reflect the inclusion of geminate segments in the Japanese word list. Geminate segments are not expected to be perceived based on spectral change. Analysis of language specific effects is possible, however, when comparing the data for specific types of transitions. Such comparisons will be discussed in section 4.3.

#### 4.2.2. Probability of overlap by chance

In order to evaluate the results presented above, it is necessary to determine how often the area of maximal perceptual change would surround a  $D_{\max}$  point if the area of maximal perceptual change were distributed randomly. I calculated the chance probability of the area of maximal perceptual change being classified as surrounding a  $D_{\max}$  point separately for each word, taking into account for each word the duration of the area which was gated through and the number and location of  $D_{\max}$  points relative to the beginning and end of the gated area.

Beginning with a simple case, in which the gated area<sup>6</sup> contains only one  $D_{\max}$  point, and that  $D_{\max}$  point does not fall near the beginning or end of the gated area, the probability of the beginning of the 20 ms area of maximal spectral change falling within 20 ms before or 5 ms after the  $D_{\max}$  point for that word is equal to 25 divided by the total

---

<sup>6</sup> The area between the endpoint of the first gate and the endpoint of the last gate

duration of the gated area in milliseconds. (Since the area of maximal perceptual change is counted as surrounding the  $D_{\max}$  point even if it begins within 5 ms after the  $D_{\max}$  point, this is equivalent to the beginning of the area of maximal perceptual change falling within 20 ms before or 5 ms after the  $D_{\max}$  point.) A hypothetical example of such a case is illustrated in Figure 4.7. Because the difference between the penultimate and ultimate gates is frequently less than 20 ms (as discussed in Section 2.1.2.2), so that not all gates are of equal duration, I chose to calculate the probabilities based on durations in milliseconds rather than on numbers of gates.

When a word has more than one  $D_{\max}$  point within the gated area, or when a  $D_{\max}$  point falls near the beginning or end of the gated area, the calculation of the probability becomes more complicated. Figure 4.8 shows a hypothetical example in which there is exactly one  $D_{\max}$  point, but it falls less than 20 ms after the beginning of the gated area (between the endpoints of the first and second gates). In such a case, the probability that the beginning of the area of maximal perceptual change will fall 20 ms before the  $D_{\max}$  point is zero, since that lies outside the area which was gated. However, the beginning of the area of maximal change could fall anywhere from the endpoint of the first gate to the  $D_{\max}$  point, or within 5 ms after the  $D_{\max}$  point, and be counted as surrounding the  $D_{\max}$  point. Therefore, the chance probability for that word of the area of maximal change being classified as surrounding the  $D_{\max}$  point is the number of milliseconds from the endpoint of the first gate to 5 ms after the  $D_{\max}$  point divided by the total duration of the gated area, as shown in the figure.

Similarly, the location of a  $D_{\max}$  point relative to the end of the gated area must be considered. If a word has exactly one  $D_{\max}$  point, which falls between the penultimate and ultimate gates of the word, the beginning of the area of maximal perceptual change cannot possibly fall within the normally allowed 5 ms window after the  $D_{\max}$  point, since the beginning of the area of maximal change must be followed by at least one more gate endpoint. This is shown in Figure 4.9. In this case, since the beginning of the area of

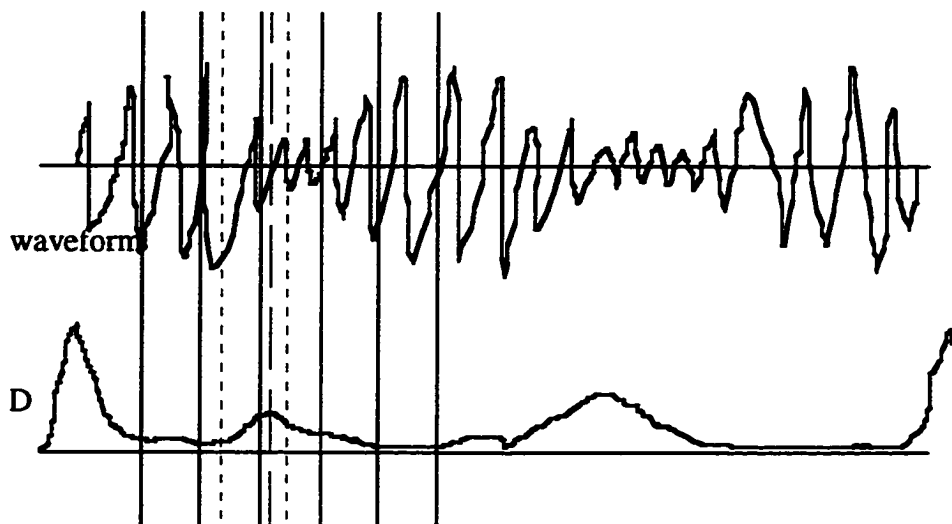


Figure 4.7. Drawing of a hypothetical case in which there is exactly one  $D_{\max}$  point, and it is not located within 20 ms of the endpoint of the first gate or 25 ms of the endpoint of the final gate. Solid lines represent endpoints of gates. The dashed line is the  $D_{\max}$  point within the gated area, and the dotted lines appear 20 ms before and 5 ms after the  $D_{\max}$  point. If the beginning of the area of maximal perceptual change falls between the dotted lines, the area of maximal perceptual change will be counted as surrounding  $D_{\max}$ . Since the gated area here is 100 ms long (assuming the last two gates are separated by 20 ms), the probability of that happening by chance is 25/100.

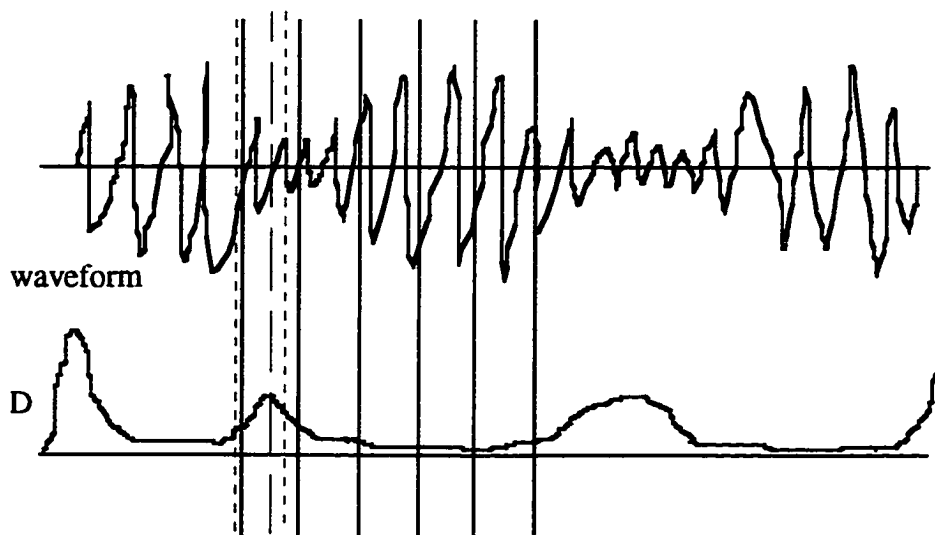


Figure 4.8. Drawing of a hypothetical case in which there is exactly one  $D_{\max}$  point, and it is located 10 ms after the endpoint of the first gate. Solid lines represent endpoints of gates. The dashed line is the  $D_{\max}$  point within the gated area, and the dotted lines appear at the end of the first gate and 5 ms after the  $D_{\max}$  point. If the beginning of the area of maximal perceptual change falls between the dotted lines, the area of maximal perceptual change will be counted as surrounding  $D_{\max}$ . Since the gated area here is 100 ms, as in Figure 4.7, the probability of that happening by chance is 15/100.

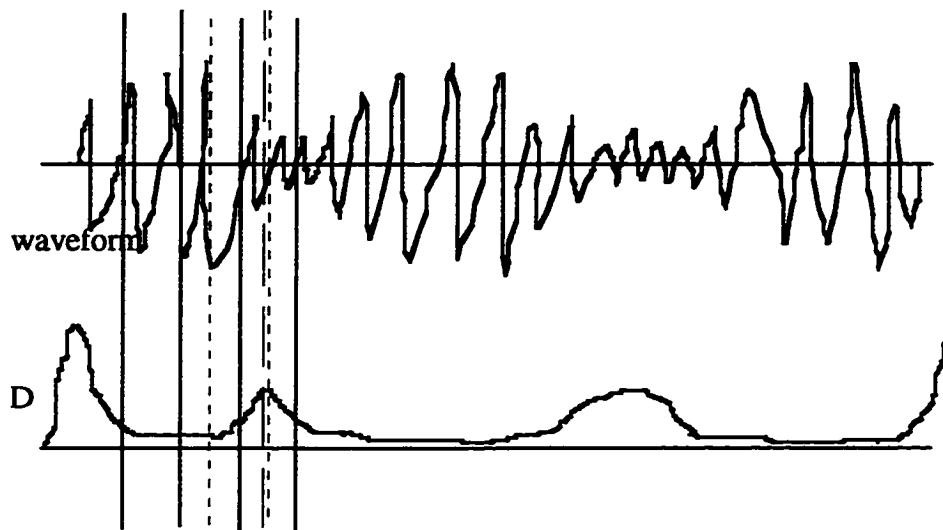


Figure 4.9. Drawing of a hypothetical case in which there is exactly one  $D_{\max}$  point, and it is located between the penultimate and ultimate gate endpoints, 10 ms from the endpoint of the final gate. Dotted lines appear 20 ms before the  $D_{\max}$  point and at the  $D_{\max}$  point. If the beginning of the area of maximal perceptual change falls between the dotted lines, the area of maximal perceptual change will be counted as surrounding  $D_{\max}$ . Since the gated area here is 60 ms, the probability of that happening by chance is 20/60.

maximal change may fall anywhere in the 20 ms before the  $D_{\max}$  point, but not after it, the probability of the area of maximal spectral change being classified as surrounding the  $D_{\max}$  point is 20 ms divided by the total duration of the gated area. Furthermore, if the only  $D_{\max}$  point falls less than 5 ms before the endpoint of the penultimate gate, the normally 5 ms window after the  $D_{\max}$  point is shortened to end at the penultimate gate's endpoint (Figure 4.10). Therefore, the probability is 20 ms plus the number of milliseconds between the  $D_{\max}$  point and the end of the penultimate gate, divided by the total duration of the gated area.

If a word has more than one  $D_{\max}$  point, and the  $D_{\max}$  points are separated from each other by at least 25 ms, the same procedures are followed as for a single  $D_{\max}$  point, and the probability of the area of maximal change surrounding any  $D_{\max}$  point is the sum of the time windows for each  $D_{\max}$  point divided by the total duration of the gated area. This is exemplified in Figure 4.11. If, however, the  $D_{\max}$  points are separated by less than 25 ms, their windows within which the beginning of the area of maximal change could fall overlap. Since the area of maximal perceptual change will be counted as surrounding a  $D_{\max}$  point regardless of which point it surrounds, a given area is only counted once in determining the probability of this happening, even if that area falls within the windows which surround two different points. This is shown in Figure 4.12. Thus, the probability of the area of maximal change being counted as surrounding a  $D_{\max}$  point is the number of milliseconds which fall within the relevant window for at least one  $D_{\max}$  point, divided by the total duration of the gated area.

Using these methods, I calculated, for each word separately, the probability of the area of maximal change for the perceptual measures being counted as surrounding a  $D_{\max}$  point by chance. This probability is the same for a given word for both the number of responses measure (#Resp) and the percent correct measure (%Corr), since both use the same gate endpoints and the same  $D_{\max}$  points. I then calculated the average probability across all words of the area of maximal change being counted as surrounding a  $D_{\max}$  point.

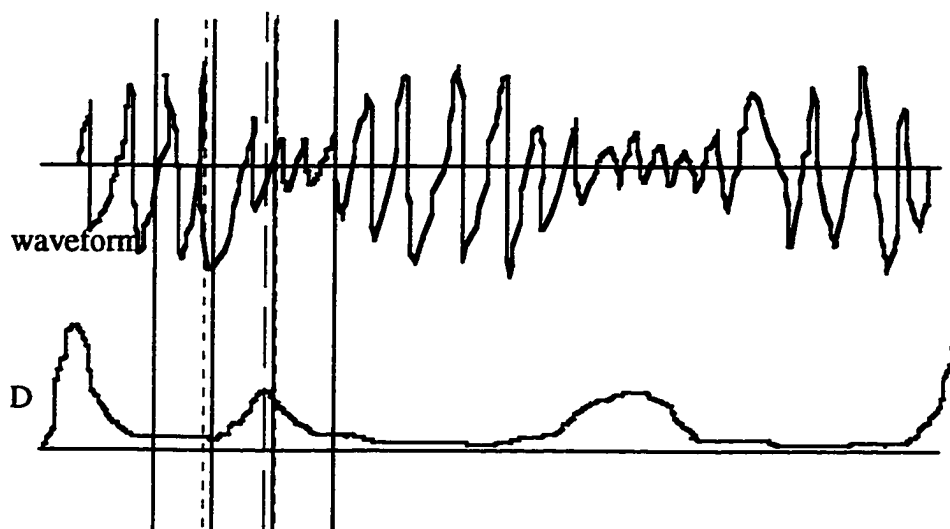


Figure 4.10. Drawing of a hypothetical case in which there is exactly one  $D_{\max}$  point, and it is located 2 ms before the endpoint of the penultimate gate. Dotted lines appear 20 ms before the  $D_{\max}$  point and at the endpoint of the penultimate gate. If the beginning of the area of maximal perceptual change falls between the dotted lines, the area of maximal perceptual change will be counted as surrounding  $D_{\max}$ . Since the gated area here is 60 ms, the probability of that happening by chance is  $22/60$ .

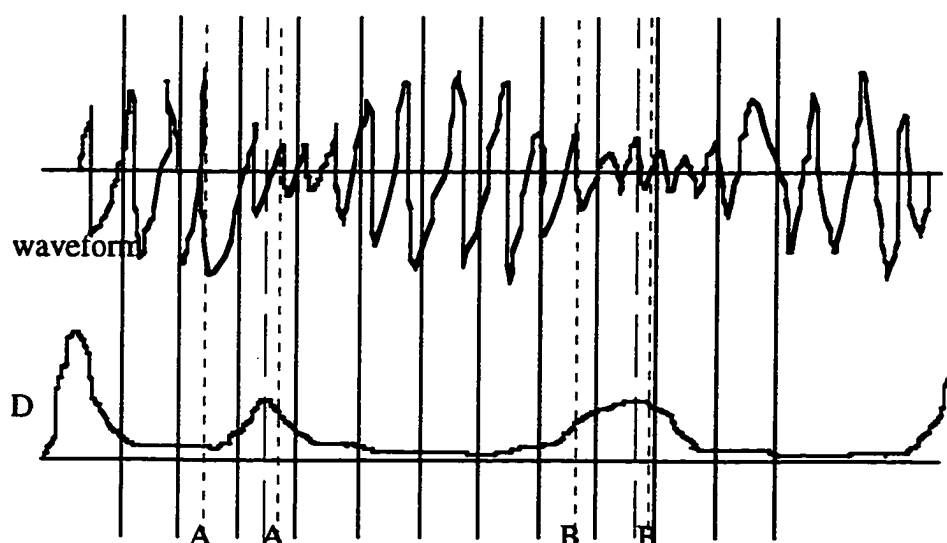


Figure 4.11. Drawing of a hypothetical case in which there are two  $D_{\max}$  points which are separated from each other by more than 25 ms and not near the initial or final gates. Dotted lines appear 20 ms before and 5 ms each  $D_{\max}$  point. If the beginning of the area of maximal perceptual change falls between either the two dotted lines labeled A or the two labeled B, the area of maximal perceptual change will be counted as surrounding  $D_{\max}$ . Since the gated area here is 220 ms, the probability of that happening by chance is  $50/220$ .

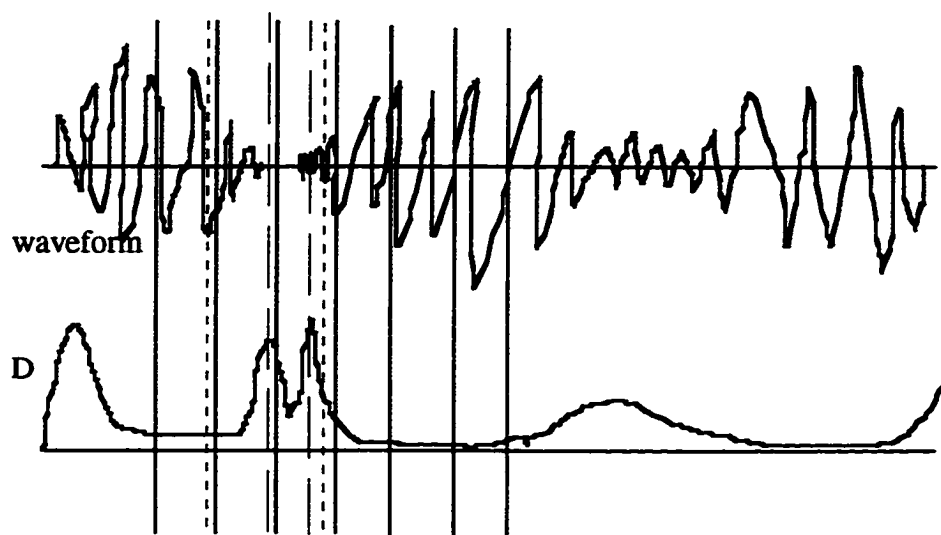


Figure 4.12. Drawing of a hypothetical case in which there are two  $D_{\max}$  points which are separated from each other by 10 ms and not near the initial or final gates. Dotted lines appear 20 ms before the earlier  $D_{\max}$  and 5 ms after the later one. The area between the two  $D_{\max}$  points falls within the 25 ms window of one or both  $D_{\max}$  points. If the beginning of the area of maximal perceptual change falls between the dotted lines, it will be counted as surrounding  $D_{\max}$ . Since the gated area here is 120 ms, the probability of that happening by chance is  $35/120$

I calculated this average probability separately for the %Corr measure and the #Resp measure, because in the calculation of the actual number of words for which the area of maximal change did surround a  $D_{\max}$  point, words with anomalous slopes were excluded only for the perceptual measure for which their data was anomalous. For example, the word "duck" /dʌk/ (word number 11 in Table 4.1) has an anomalous slope for the #Resp measure. The number of different responses given to it tends to increase as listeners hear more of the word instead of decreasing. This word was therefore excluded from the comparison of maximal perceptual change and  $D_{\max}$  for the #Resp measure. However, it had the expected slope (positive) for the %Corr measure, so it was not excluded from calculations involving that measure. One should remember that the number of responses (#Resp) and percent correct (%Corr) measures, while they have a theoretical relationship in the cohort model, are separate measures which are often not closely dependent on each other.

Thus, in order to calculate the average probability across all words of the area of maximal change surrounding a  $D_{\max}$  point for the number of responses (#Resp) data, I excluded the individual word probabilities for words which had non-negative slopes or a better linear than ogival fit for the #Resp data from the calculation. For the average probability for the percent correct (%Corr) data, I excluded the individual word probabilities for words with anomalous slopes or better linear fits in the %Corr data. That is, in order to calculate the average probabilities for each type of perceptual measure, only those words which contributed to the observed results for that perceptual measure were used. This average probability represents the percentage of words which would be counted as having the area of maximal change surrounding  $D_{\max}$  by chance, if there were no relationship between  $D_{\max}$  and when perception takes place. By multiplying by the total number of words under consideration, one can obtain the number of words which would be predicted to have this situation by chance. The resulting average probabilities and



predicted number of words appear in Table 4.3. The actual results, from Table 4.2, are repeated for clarity.

Table 4.3. Actual and predicted (chance) numbers and percentages of words with area of maximal perceptual change surrounding  $D_{max}$  for each perceptual measure and each language. Number of words is shown relative to the number of words for that category which are more ogival than linear and do not have anomalous slopes.

Type of data		English		Japanese	
		Proportion	Percentage	Proportion	Percentage
#Resp	Actual	36/102	35.3%	22/52	42.3%
	Predicted by chance	34.1/102	33.4%	18.7/52	36.0%
%Corr	Actual	63/117	53.8%	34/73	46.6%
	Predicted by chance	39.0/117	33.3%	25.6/73	35.1%

For example, if dynamic vs. static cues had no influence on where a segment or a word is recognized, and thus perceptual improvement had no relationship to the location of the point(s) of maximal spectral change, the area of maximal change in the percent correct (%Corr) measure would still, by chance, surround a  $D_{max}$  point in 33.3% of the English words.

I used the chi-squared test for goodness of fit to determine whether the actual number of words in each category which had the area of maximal perceptual change surrounding  $D_{max}$  differed significantly from the number which would be predicted to do so by chance. The difference was significant for the English %Corr measure ( $\chi^2$  (1, N=117)=22.15,  $p<.001$ ) and the Japanese %Corr measure ( $\chi^2$  (1, N=73)=4.24,  $p<.04$ ). The difference was not significant for the English #Resp measure ( $\chi^2$  (1, N=102)=0.16,  $p>.05$ ) or the Japanese #Resp measure ( $\chi^2$  (1, N=52)=0.91,  $p>.05$ ). Thus, for the percent correct measure (%Corr) in both languages, the area of maximal change in the perceptual data surrounds a  $D_{max}$  point significantly more often than it would by chance. For the number of responses measure (#Resp) in both languages, the area of maximal perceptual change surrounds  $D_{max}$  slightly more often than chance, but not significantly so.

#### 4.3. Analysis of cases in which perceptual measures are not related to the $D_{\max}$ point

Although the results for the %Corr measure in both languages are significantly greater than chance, the fact that the area in which listeners' perception becomes accurate most quickly only surrounds a  $D_{\max}$  point approximately half of the time seems surprising under the prediction that dynamic cues are the most important. However, examination of the cases in which the area of maximal change of the perceptual measures fails to surround  $D_{\max}$  reveals several categories of transitions for which there are logical reasons for their failure to follow the hypothesis.

##### 4.3.1. Categories of exceptions

###### 4.3.1.1. Postvocalic stops

For stops after stressed vowels in English (or after any vowel in Japanese), listeners frequently make the most progress toward recognition of the stop at a point in the preceding vowel, well before the  $D_{\max}$ . Figure 4.13 shows an example of this using the word "oats" /oʷts/, and Figure 4.14 shows the Japanese example /hatake/ 'field.' The  $D_{\max}$  point in such cases falls at the beginning of the closure for the stop, but it is well known that formant transitions within the preceding vowel provide at least some cues for place of articulation of a postvocalic stop, and that listeners can often recognize the place of a following stop from cues in the vowel (Repp 1980, some conditions of the experiment reported by Takehi et al. 1996). Although formant transitions are a dynamic cue, since they are inherently changing, they do not produce a high value of the measure  $D$ , because  $D$  is far more sensitive to changes in amplitude at any part of the spectrum than to slower changes in frequency of a part of the signal, as discussed in section 3.2.1.1 above. The beginning of a stop closure, especially for voiceless stops, produces a very large value of  $D$ , though.

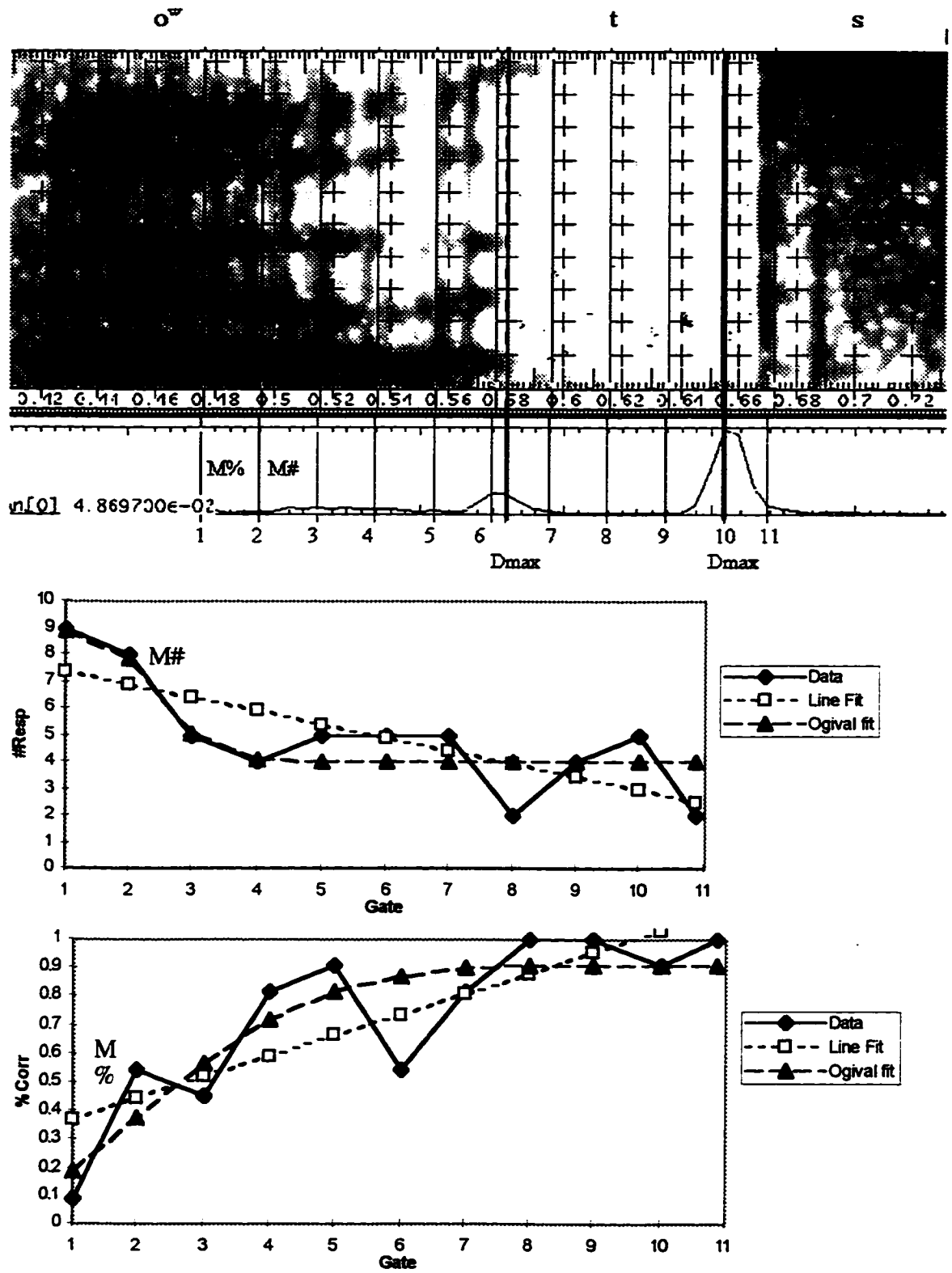


Figure 4.13. Spectrogram, measure D, and the two perceptual measures for "oats" /oʊts/. Area of maximal change for each perceptual measure is several gates before the  $D_{max}$  point.

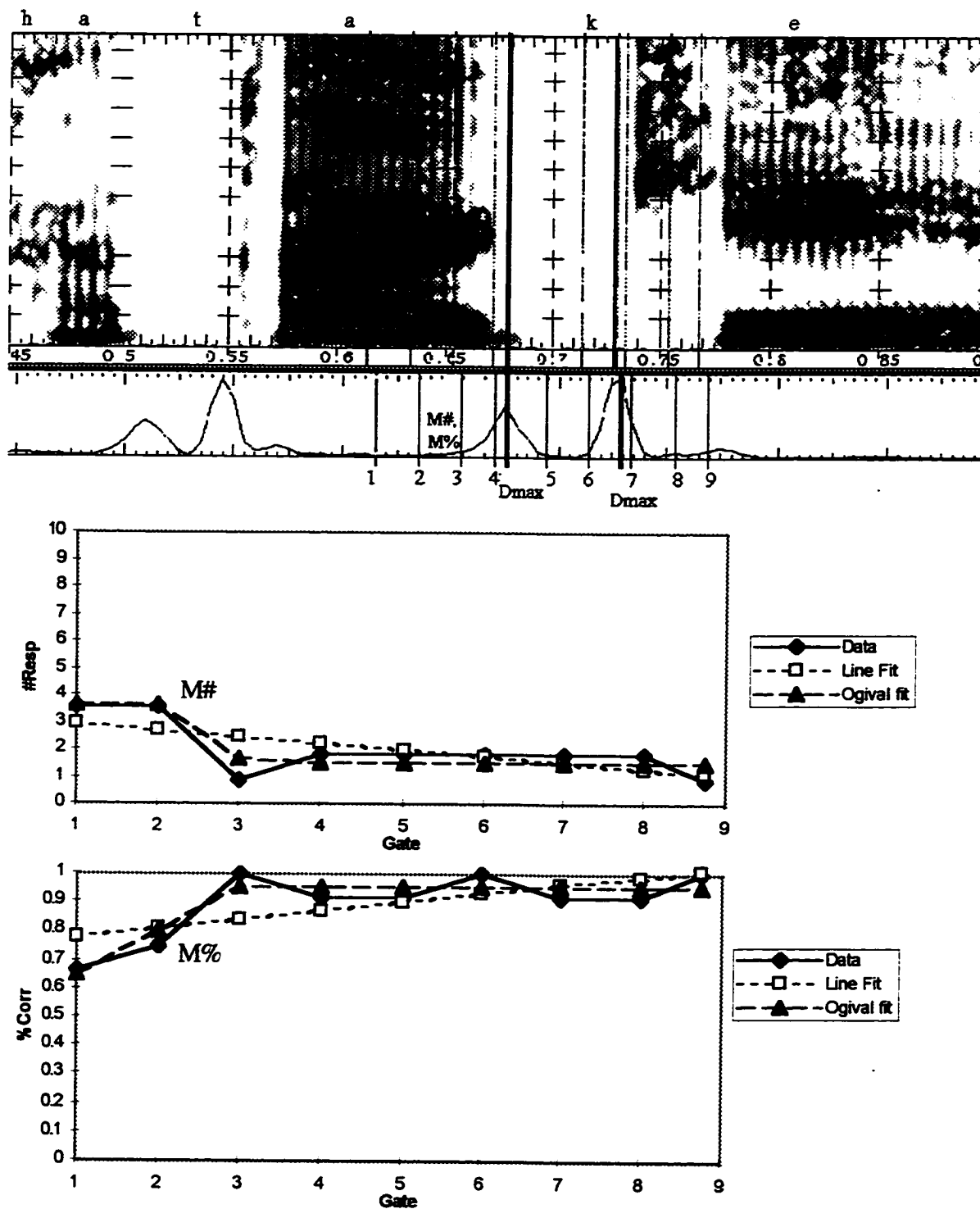


Figure 4.14. Spectrogram, measure D, and the two perceptual measures for /hatake/ [hatake] 'field.' Area of maximal change for the perceptual measures occurs before the  $D_{max}$  point.

Listeners recognize the place of articulation of the postvocalic stop from the low-D formant transitions, and the duration of the vowel up to the gate provides sufficient cues for listeners to decide whether the postvocalic stop is voiced or voiceless. This is clearly exemplified by the word "fade" /f<sup>e</sup>ɪd/ in the English data: at early gates, 90% or more of listeners gave responses with a voiceless postvocalic obstruent. Common responses were "fate, face, fake, faith." At gates shortly before the /d/ closure, however, 90% or more of listeners gave responses with a voiced postvocalic obstruent, either "fade" or "phase." Perception of voicing based on preceding vowel duration will be discussed in detail in section 5.7.1 below, but this brief example demonstrates how listeners can recognize both the place and voicing of a postvocalic stop before reaching the beginning of the stop or the  $D_{\max}$  point. This leaves only the manner of the postvocalic consonant to identify, and the cohort of the word may leave only the stop as a possibility in many cases. Thus, listeners have a good chance of identifying a postvocalic stop correctly based only on cues within the preceding vowel, so it is logical that recognition of postvocalic stops would often take place before reaching the  $D_{\max}$  point at the end of the vowel.

#### 4.3.1.2. Transitions into vowels

A second common category in which the area of maximal change in the perceptual measures often fails to surround  $D_{\max}$  is transitions from any consonant into a vowel. Although Furui (1986) found that the Japanese listeners in his experiment became able to identify vowels immediately after the  $D_{\max}$  point, listeners for both languages in my experiment often fail to identify the vowel until after the vowel has begun. Figures 4.15 and 4.16 demonstrate this with the words "toad" /toʊd/ and /kato/ 'crossing.' The  $D_{\max}$  point in a consonant vowel transition usually occurs at the point one would define as the onset of the vowel, that is, at onset of voicing after a voiceless consonant or onset of higher formants after most voiced consonants. In my experiment, the English listeners often do

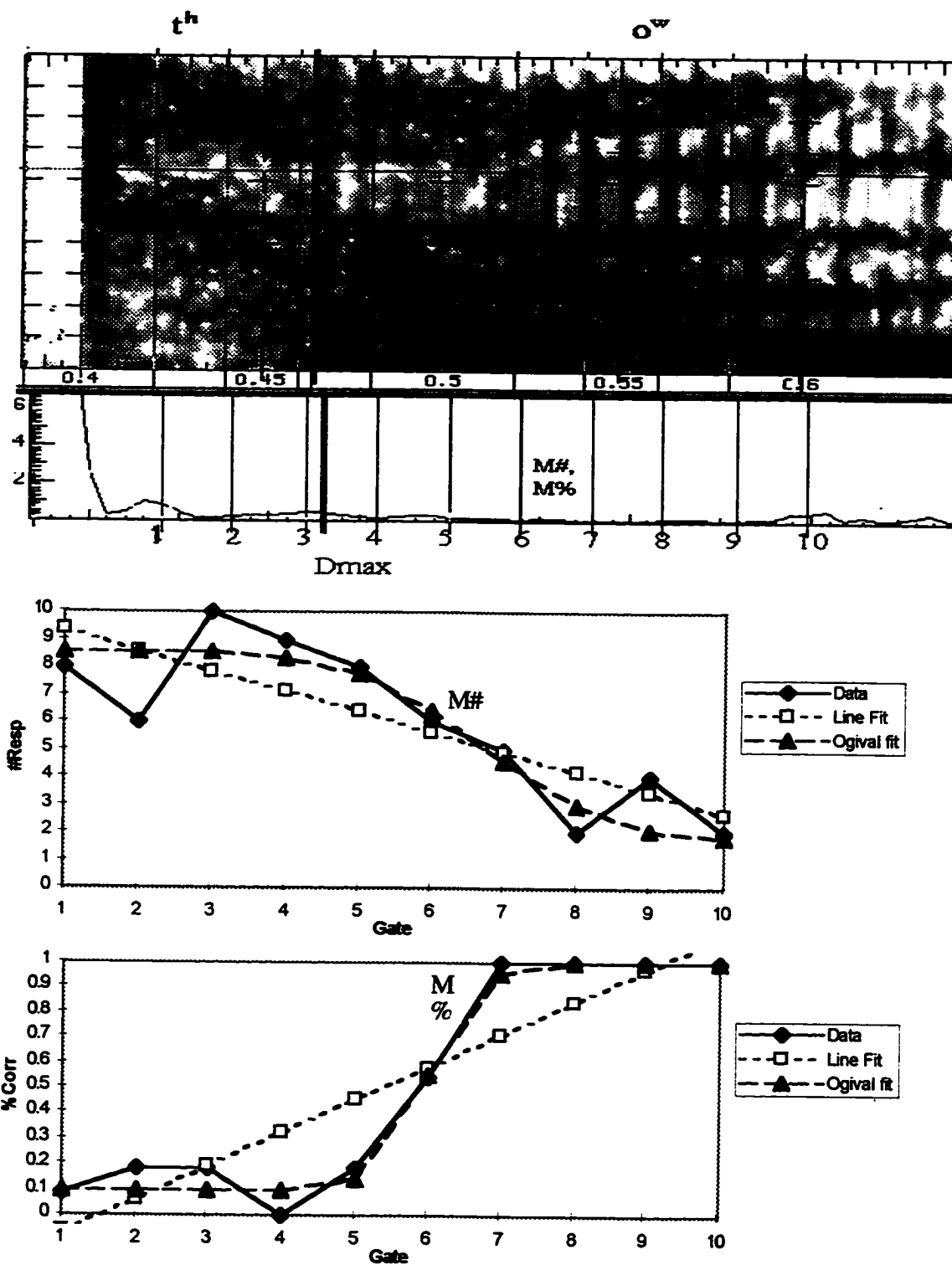


Figure 4.15. Spectrogram, measure D, and the two perceptual measures for "toad" /to<sup>ʷ</sup>d/. Area of maximal change for the perceptual measures occurs after the  $D_{max}$  point.

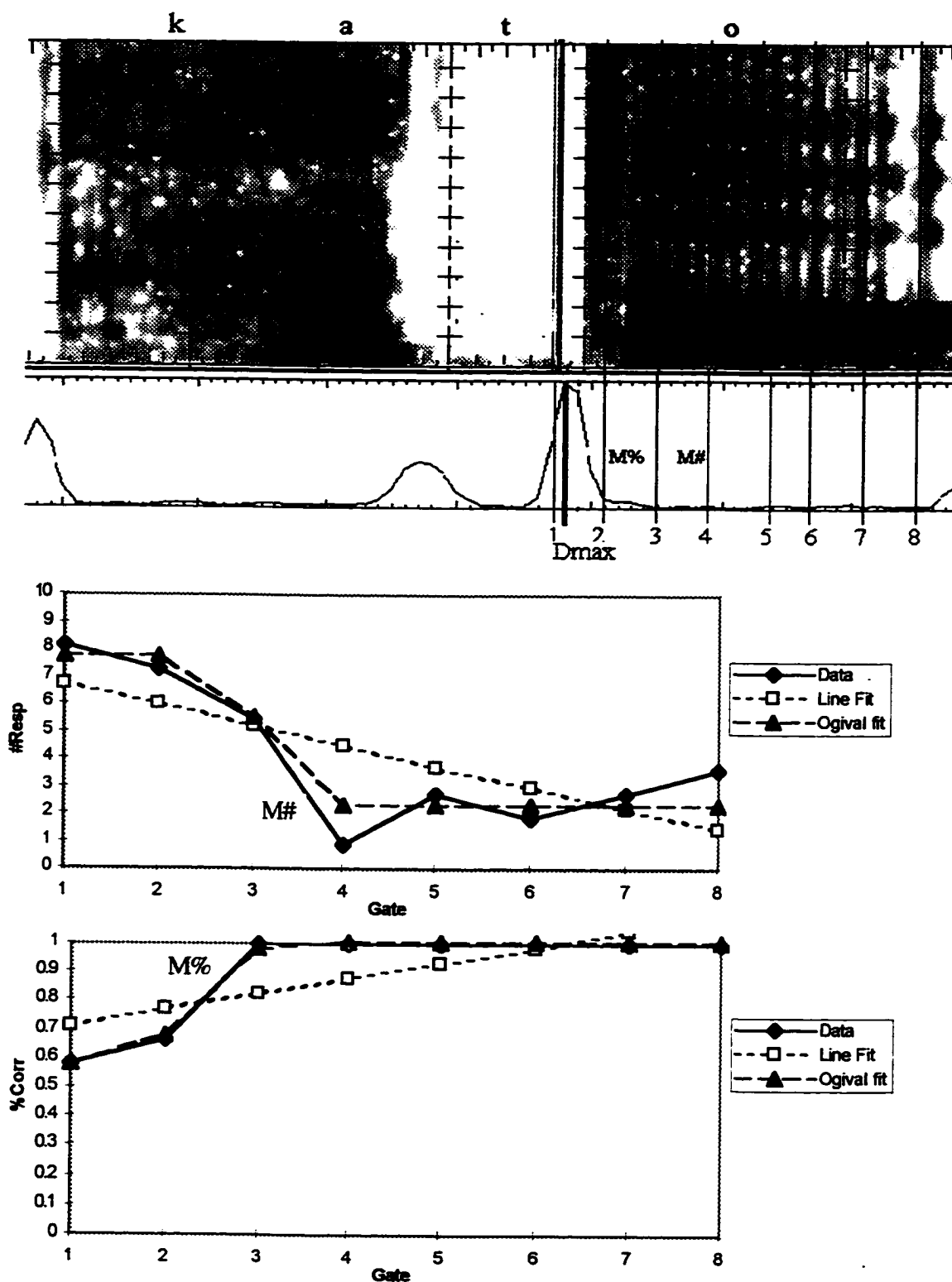


Figure 4.16. Spectrogram, measure D, and the two perceptual measures for /kato/ [kato] 'crossing.' Area of maximal change for the perceptual measures occurs after the  $D_{max}$  point.

not identify the vowel correctly until several gates into the vowel, and to some extent, the Japanese listeners show the same pattern. For the English results, this probably reflects the relatively large vowel inventory of English. The gate surrounding the  $D_{\max}$  point allows the listener to hear at most approximately 20 ms of the vowel, and this is not enough for English listeners to distinguish /e/ from /æ/ or /a/ from /ʌ/, for example. Confusions between these pairs are common in the data, and reflect the same pattern found by Lang and Ohala (1996). Reasons for the discrepancy between my Japanese results and Furui's results on this point will be discussed below.

#### 4.3.1.3. Vowel-Vowel transitions

A third category of transitions for which the area of maximal perceptual change often does not surround  $D_{\max}$  is vowel-vowel transitions. In Figure 4.17, the word "diagonal" /dɪˈæɡənəl/ demonstrates this. As discussed with regard to VC formant transitions above, the measure  $D$  is not very sensitive to changes in formant frequencies, and of course, in a vowel-vowel transition, formant frequency changes are the relevant feature of the signal. Thus, a VV transition tends to have extremely low values of the measure  $D$  throughout the two vowels, as was shown in Figure 3.13 in Chapter 3. Whatever point within the gated area has the largest value of  $D$  is established as the  $D_{\max}$  point, but this point often has an extremely low  $D$  value, despite being the maximum, and is not much greater than the low  $D$  values of several other points. Furthermore,  $D_{\max}$  points in such cases are often not associated with any clear change in the acoustic signal, as was discussed in Chapter 3. Therefore, in VV transitions, the location of the  $D_{\max}$  point may be somewhat arbitrary, so it is not surprising that the area of maximal change in the perceptual measures is often not near the  $D_{\max}$  point.



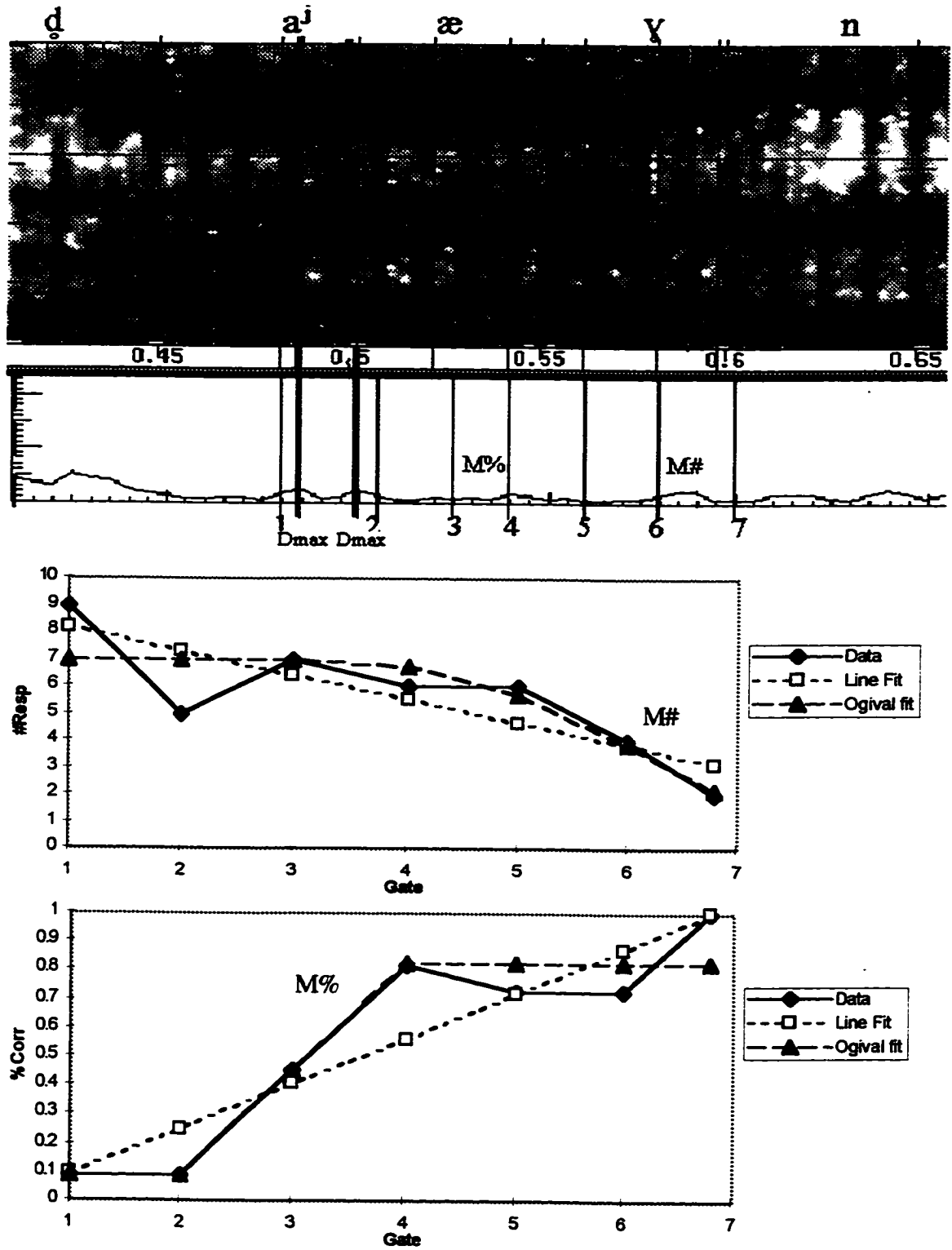


Figure 4.17. Spectrogram, measure D, and the two perceptual measures for "diagonal" /daːgənəl/. Areas of maximal change for the perceptual measures occur after the D<sub>max</sub> points.

#### 4.3.1.4. Transitions into sonorants

Transitions into sonorants, especially after segments which allow cues for the sonorant to spread into the preceding segment (vowels, aspirated stops, some fricatives), have a similar problem. Nasalization, of course, spreads far into a preceding vowel, and English /r/ has a strong effect on many preceding segments (lowering of the third formant in vowels and the frequency of aspiration or frication noise). Because the spread of nasalization or a lowered third formant, however, creates slow changes in the acoustic signal and is largely confined to changes in formant frequencies, not amplitude, the measure  $D$  is not likely to show any elevated values during such parts of the signal. The  $D_{\max}$  will fall at the onset of the sonorant, if there is a sudden change in the signal there, well after cues to the sonorant become available. Listeners for both languages often identified sonorants before reaching the  $D_{\max}$  point if the environment allowed cues for the sonorant to spread into the preceding segment. Figures 4.18 and 4.19 show the examples "fair" /fe<sup>i</sup>r/ and /kyaku/ 'guest.'

#### 4.3.1.5. Fricatives

Fricatives present a different problem for the hypothesis. Furui (1986) found that when he gated fricative-vowel transitions from the beginning of the signal (initial gating) instead of the end, listeners did not identify the fricative correctly if they heard only from shortly before the  $D_{\max}$  point to the end of the syllable. (For other types of consonants, listeners were able to identify the consonant at comparable points in the initial gating condition.)  $D_{\max}$ , in a fricative-vowel transition, usually falls at the end of frication and onset of voicing<sup>7</sup>. Listeners could only identify initial-gated fricatives correctly if they heard somewhat more of the fricative than the part within a short window of the  $D_{\max}$  point. Furui concluded that perception of fricatives depends not on dynamic cues but on having at

---

<sup>7</sup> Although there may be two separate  $D_{\max}$  points because of changes in the quality of frication during the fricative, one  $D_{\max}$  point will always be located at the onset of voicing for the vowel.

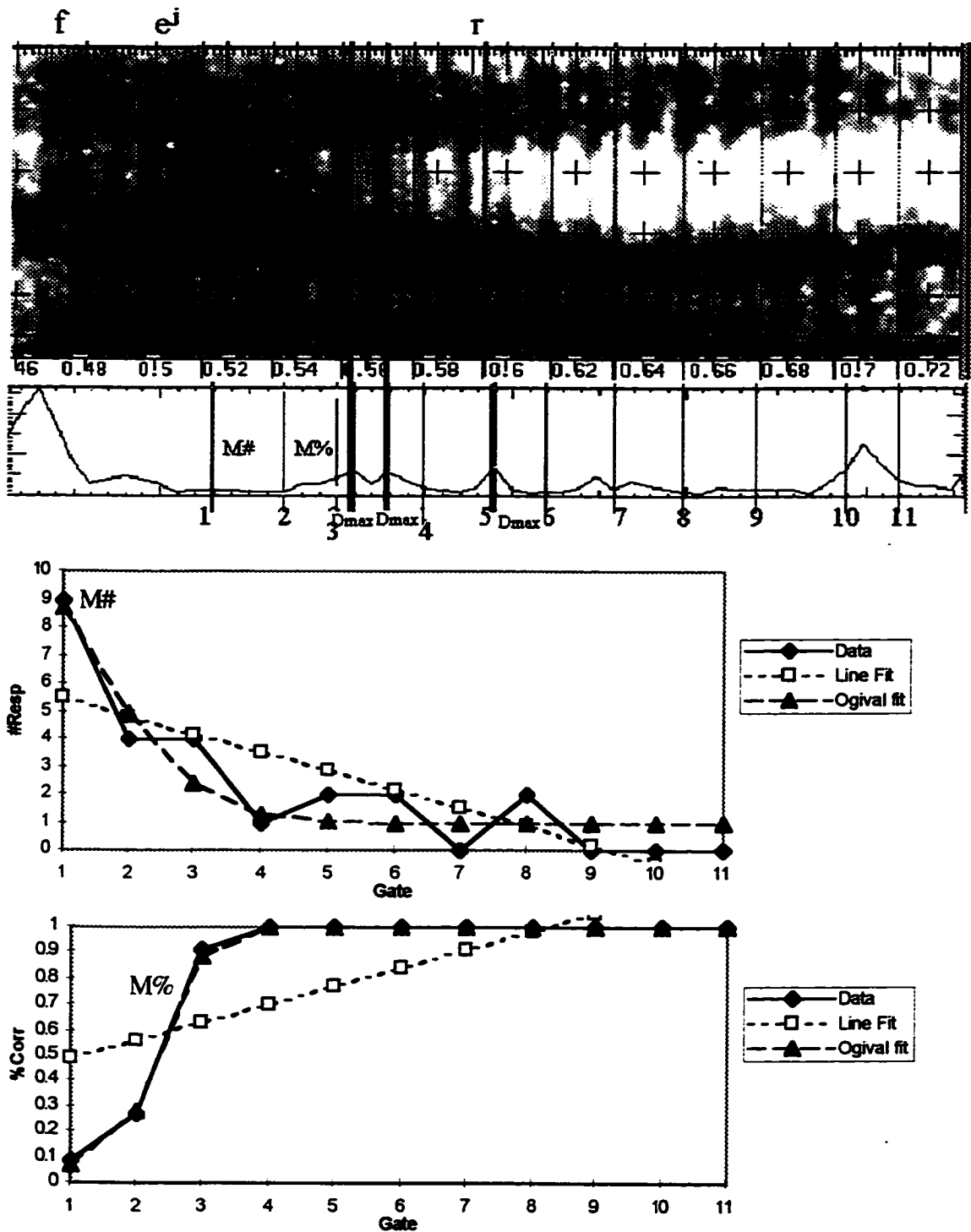


Figure 4.18. Spectrogram, measure D, and the two perceptual measures for "fair" /fer/. Areas of maximal change for the perceptual measures occur before any  $D_{max}$  point.

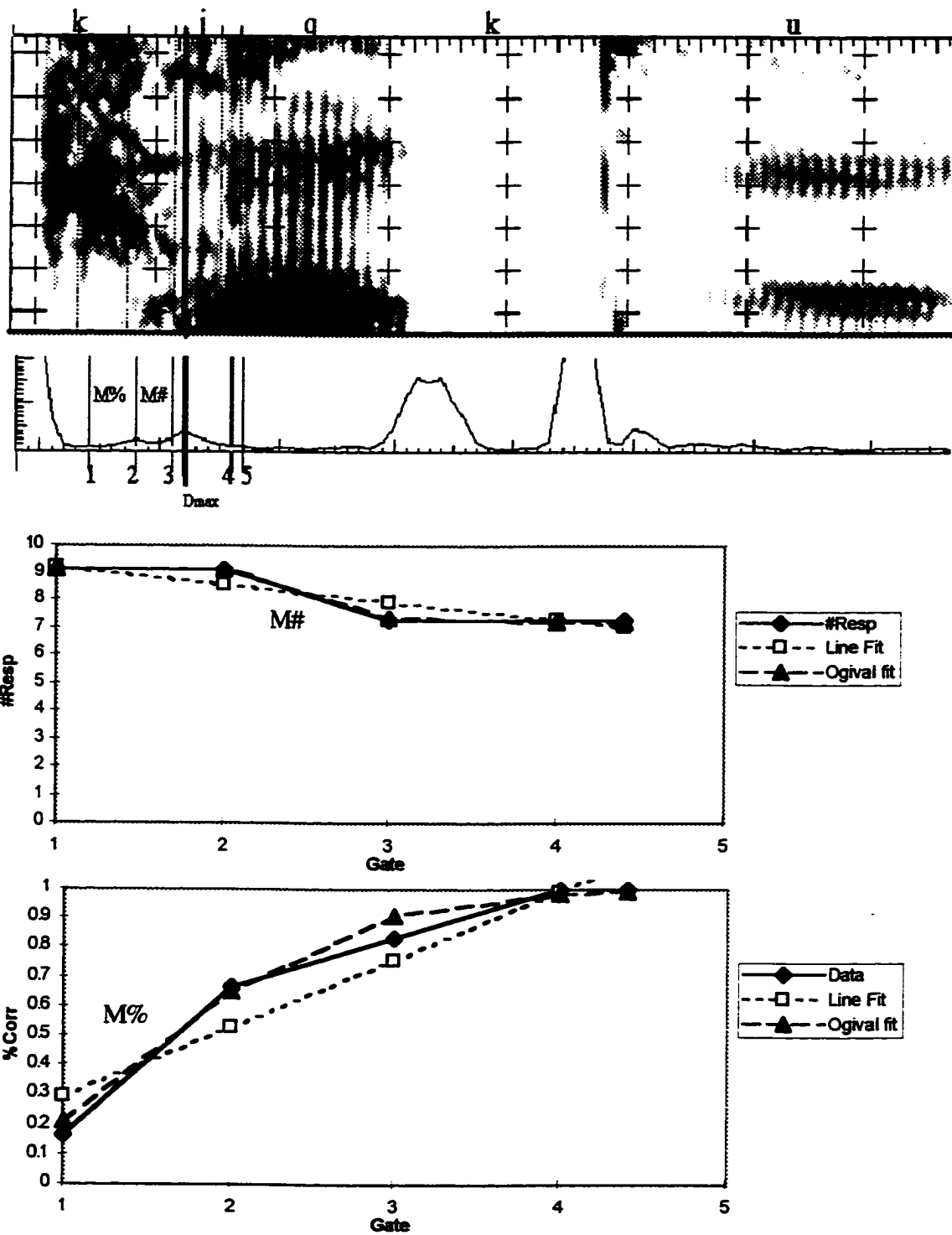


Figure 4.19. Spectrogram, measure D, and the two perceptual measures for /kyaku/ [kjaku] 'guest.' Area of maximal change for the perceptual measures occurs before the  $D_{max}$  point.

least a certain duration of noise and possibly on the rise time of the amplitude of that noise. Kluender and Walsh (1992) also find that duration of noise is a crucial perceptual cue to fricatives. Because Furui used only CV transitions, this point only applied to initial gated syllables in his experiment. My experiment involves only final gating, but since VC transitions were used, the same problem occurs: listeners for both languages tend to recognize fricatives much later than the  $D_{\max}$  point, at a point somewhere during the fricative where the acoustic signal does not show any noticeable changes, and the measure D is very low. This is shown in Figure 4.20 for the word leaf /lif/.

#### 4.3.1.6. Exceptions specific to Japanese

Two other effects occur only in the Japanese data. First, the geminates and long vowels of the Japanese data often do not have the area of maximal change of the perceptual measures surrounding the  $D_{\max}$ . This is not surprising, as one does not expect spectral change to be an important cue in the perception of distinctive length, for which duration is the most likely cue. For geminate voiceless stops, there is a large peak of the measure D at the burst of the long stop. For geminate fricatives and long vowels, as the sound is quite steady throughout, there are only very low values of D. In either case, however, the location of  $D_{\max}$  is not related to the time at which the gate includes enough of the long segment for listeners to judge that it is long. Figure 4.21 shows the example /keego/ 'honorific language.' This effect, of course, does not occur in English, since English has no geminates or monophthongal long vowels.

Secondly, the Japanese data contains two examples of transitions into medial glides, /mawari/ 'surroundings' and /huyoo/ 'unnecessary.' In both cases, the glide is recognized after the gate surrounding the  $D_{\max}$  point, which is associated with the dip in amplitude and weakening of formants for the glide. Listeners seem to require more of the signal than this (and may perhaps need to hear the next vowel) in order to recognize the glide and determine that it is not a syllabic nucleus itself. This will be discussed further in

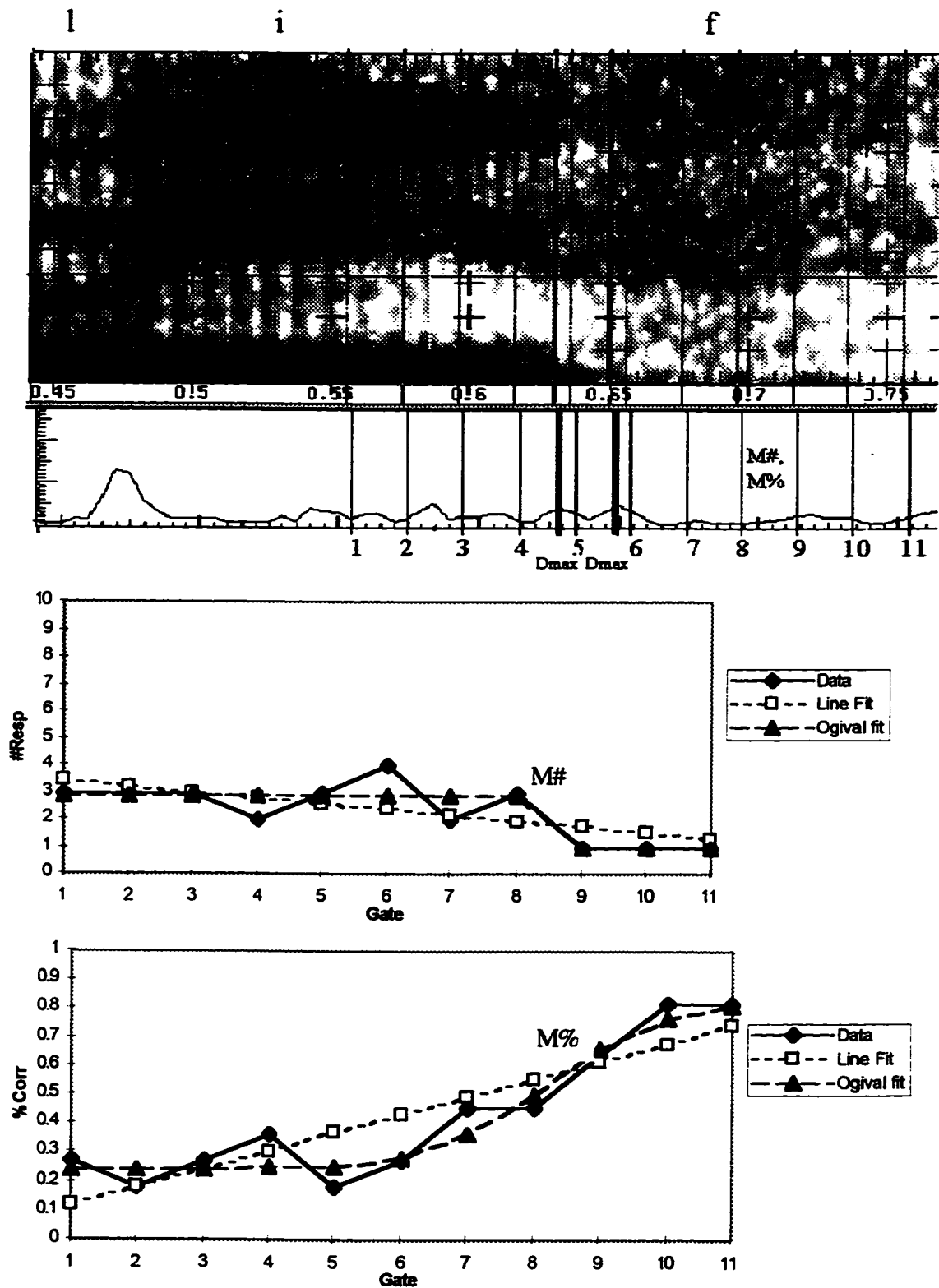


Figure 4.20. Spectrogram, measure D, and the two perceptual measures for "leaf" /lif/. Area of maximal change for the perceptual measures occurs after the  $D_{\max}$  points.

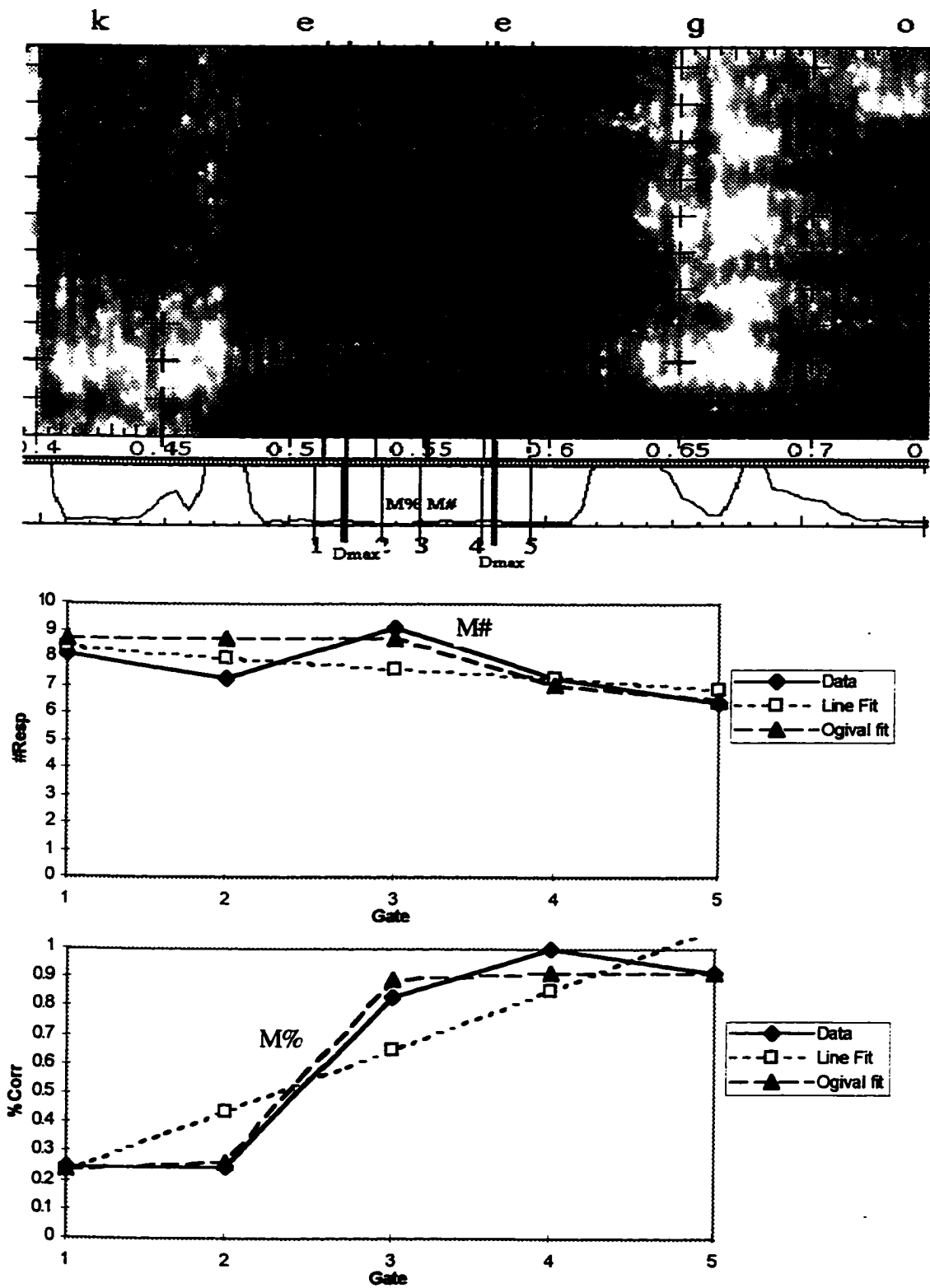


Figure 4.21. Spectrogram, measure D, and the two perceptual measures for /keego/ [keego] 'honorific language.' Areas of maximal change for the perceptual measures occur between the two D<sub>max</sub> points.

Section 5.6.2 below. Figure 4.22 shows the example /huyoo/ 'useless.' The English data contains no transitions into medial glides, as postvocalic glides are usually part of a diphthong.

#### 4.3.2. Classification of words by potential exception category

After these categories of transitions which are likely not to follow the hypothesis of recognition happening near  $D_{max}$  were identified, all of the words used in the experiment were classified as to which category, if any, their transition of interest would fall into. For this purpose, the availability of the cues which lead to the failure to follow the hypothesis was considered for the particular environment of each transition. Stops after a vowel and a sonorant, as in "court" /kɔrt/ or "help" /hɛlp/, were included in the category which was described above as postvocalic stops, even though they are not postvocalic, because they also have the potential for formant transitions for the stop to serve as cues. Stops after a nasal of the same place of articulation as the stop ("band" /bænd/, "sentiment" /sentəment/, /haŋtai/ 'opposite'), however, were not counted in this category, because the nasal could not have formant transitions for the place of articulation of the following stop, since it is already at the same place<sup>8</sup>. For the English data, stops after unstressed vowels were not included in the postvocalic stop category, because the vowels appear to be too short for listeners to identify the postvocalic stop from cues in the vowel. Decisions such as this were made on a post hoc basis, but this method of identifying categories which fail to follow a hypothesis is in itself post hoc, so this part of the analysis should be considered exploratory. As I will discuss below, several aspects of the results which I discovered through this post hoc analysis should be clarified with additional experiments in future work.

---

<sup>8</sup> To the extent that nasals must be homorganic with a following stop, the place of the nasal should serve as a cue to the place of the stop. The use of phonotactic constraints in perception is a different topic from the use of phonetic cues which spread in a signal, however.



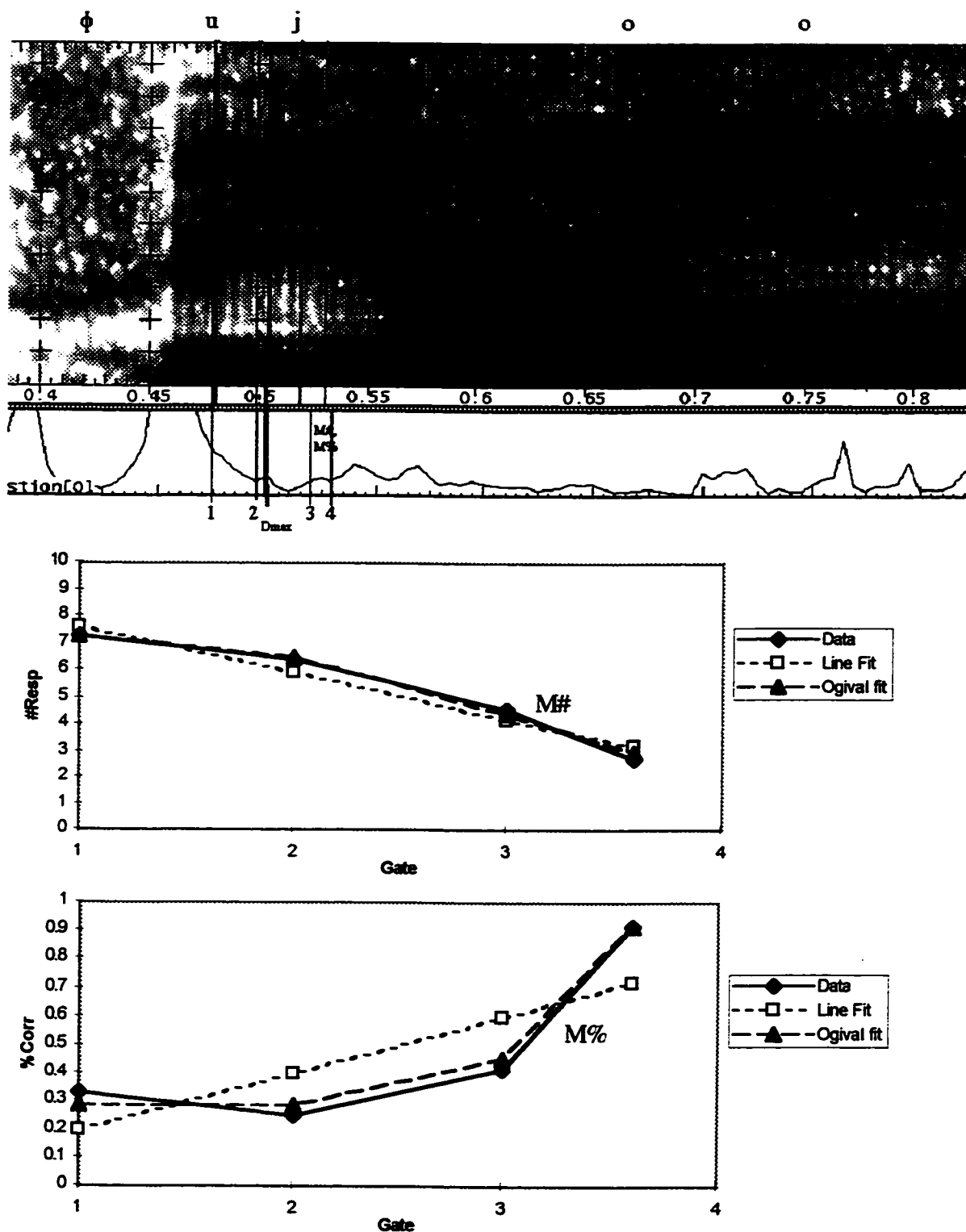


Figure 4.22. Spectrogram, measure D, and the two perceptual measures for /huyoo/ [ɸujoo] 'unnecessary.' Area of maximal change for the perceptual measures occurs after the  $D_{max}$  point.

The environment for the category of transitions into sonorants is somewhat complicated. Postvocalic sonorants, of course, have cues which spread into the preceding vowel. Sonorants after aspirated stops frequently also have a strong influence on the quality of the aspiration noise, which creates the potential for early cues to the sonorant in words like "plain" /pleɪn/, "crops" /kraps/, and /kyaku/ 'guest.' Similarly, sonorants can affect the quality of frication noise of some fricatives, as in "swan" /swan/, "fragile" /frædʒəbl/, "flash" /flæʃ/. Therefore, these transitions were included in the pre-sonorant category.

However, in the sequence /sl/, the /l/ cannot influence the quality of the /s<sup>h</sup>/, because these two sounds require the same articulators (the tip and sides of the tongue), and /s/ requires the tongue to be down in the middle of the tongue and up on the sides, while /l/ requires the reverse. Therefore, "sleep" /slip/ and "Iceland" /aɪslənd/ were not included in this category. Similarly, little or no nasalization can spread into a preceding fricative (Ohala 1975), so "Disney" /dɪzni/ was not included in this category. Nasalization or lateralization from a syllabic nasal or lateral cannot spread into a preceding voiceless stop, so "button" and "apple" were not included in this category either. Furthermore, sonorants after voiceless unaspirated or voiced stops may have some influence on the quality of the burst of the stop, but since the burst of such stops is likely to fall into the same gate as the beginning of the sonorant, and therefore the  $D_{max}$ , this will not lead to early recognition of the sonorant. The same is true of the sequence /ry/ [rj] in Japanese, because whatever the realization of the /r/<sup>10</sup>, it is extremely short. Therefore, "groan" /groʊn/, "acrobat" /ækroʊbæt/, "split" /splɪt/, /ryokan/ [rjokan] 'inn,' and other such words were not included in this category. The two Japanese words with transitions into

---

<sup>9</sup> At least not until very late in the /s/.

<sup>10</sup> Japanese /r/ word initially before /y/ sounds somewhat different from the usual flap, and may be affricated.

medial glides were not included in this category either, as they formed a separate category, discussed above.

The remaining categories are simpler. All transitions into vowels were classified as having the potential for late recognition of the vowel. All vowel-vowel transitions were classified as belonging to that category. All transitions into fricatives were classified as having the potential for late recognition of the fricative, and all geminate or long segments in Japanese were classified as belonging to that group. Only one type of transition has the potential to belong to more than one of the categories: transitions into affricates. Postvocalic affricates were sometimes recognized early, like postvocalic stops, and sometimes recognized late, like fricatives. Either of these effects is logical in consideration of the timing of perceptual cues. In a vowel-affricate transition, formant transitions and vowel duration should supply cues to the place and voicing of the postvocalic affricate, just as with a postvocalic stop. Only the manner is left to identify. Listeners' behavior in this situation is probably determined by the cohort for the particular word.

The only affricates which were identified early, like postvocalic stops, were those in "judge" /d͡ʒʌd͡ʒ/ and /mati/ [mat͡ʃi] 'city.' For the former, since there is no word \*/d͡ʒʌz/, the only real words with a voiced alveolar obstruent after the vowel are "judge," "Jud" (the proper name), and morphologically related words like "judgment." The target, "judge," is much more common than the stop-final possibility, although a few listeners did answer "Jud." Thus, the cohort in this case may lead listeners to respond with an affricate, once they can recognize that the next segment is voiced and alveolar. The explanation for /mati/ is less clear, since there are relatively common real words with a postvocalic voiceless alveolar stop (/mata/ [mata] 'again,' /matto/ [matto] 'mat') or a different postvocalic affricate (/matumoto/ [matsumoto] (common proper name), and these words do appear in the responses at the first gate, but are replaced by words with the affricate [t͡ʃ] well before the closure for the affricate, let alone its release. For both languages,

transitions into affricates were classified with the fricatives except for these two cases, which were classified with the postvocalic stops. This choice is obviously post hoc. A further study could be designed to test the timing of perception of postvocalic affricates, giving very careful attention to the cohorts of each word, or perhaps using non-words. Transitions which fell into none of the categories described here were classified as having no potential reason for the hypothesis not to apply.

#### 4.3.3. Classification of words' results into these categories

The actual percent correct (%Corr) results for each word which did not have the area of maximal change in the %Corr measure surrounding  $D_{\max}$  were then classified using the same categories. If a transition which had been classified as having the potential for early recognition because of a postvocalic stop was actually recognized early, it was classified as having failed to have the area of maximal change in the %Corr measure surrounding  $D_{\max}$  for that reason. If it was recognized late (the opposite of the proposed effect for its category), it was classified simply as "other," since this does not conform to the proposed explanation for deviations from the hypothesis. If it did have the area of maximal improvement in the %Corr measure surrounding  $D_{\max}$ , then it was classified as following the hypothesis. Similarly for the other categories, if a fricative was recognized (using the %Corr measure) late, it was classified as having the "fricative effect," but if it was recognized at  $D_{\max}$ , it was classified as meeting the hypothesis, and if it was unexpectedly recognized early, it was classified as "other." To be classified as showing the sonorant effect, a sonorant had to be recognized early relative to  $D_{\max}$ , and to be classified as showing the vowel effect, a vowel had to be recognized late relative to  $D_{\max}$ , not early. Transitions which were classified as having the vowel-vowel effect or the geminate (or long vowel) effect could be recognized either early or late relative to the  $D_{\max}$ , since the problem in these cases is that the location of  $D_{\max}$  is arbitrary<sup>11</sup>. Transitions which were not

---

<sup>11</sup> Except for the geminate voiceless stops, which were only recognized early, never late.

predicted to fall into any of the categories, if they did not have the area of maximal change in the %Corr measure surrounding  $D_{\max}$ , were simply classified as "other."

For all transitions which did not have the area of maximal change in the %Corr measure surrounding a  $D_{\max}$  point, the number of milliseconds from the  $D_{\max}$  point to the nearer edge of the area of maximal change was calculated. This is a measure of how much early or late the segment was recognized. For words with more than one  $D_{\max}$  point, the  $D_{\max}$  point which would be expected to be relevant for the perception of such segments was used. For example, in a transition into a postvocalic stop, the stop is usually recognized closer to the  $D_{\max}$  point for the closure of the stop than to that at the release of the stop, so distance from the  $D_{\max}$  point of the closure was calculated. In cases where it is not clear what acoustic change the  $D_{\max}$  points are associated with, as in "fair" /feɪr/, the distance to the nearest  $D_{\max}$  point was used. The resulting classifications (both potential and actual) and distances from  $D_{\max}$  are shown in Table 4.4.

Table 4.4. Categorization of words by effects which can lead to the area of most change in the %Corr measure not surrounding  $D_{\max}$ . Abbreviations for effects: VC=postvocalic stop effect, V=vowel effect, VV=vowel-vowel transition effect, R=sonorant effect, F=fricative effect, LL=long segment effect, G=medial glide effect, Yes=area of maximal perceptual change did surround  $D_{\max}$ . Words evaluated using the partially correct measure are marked with an asterisk in the actual effect category column.

No.	Word	Trans.	Potential effect category	Actual effect category for %Corr data	Time difference from $D_{\max}$ to area of max. change of %Corr
English					
1	tip	tɪ	V	Yes	
2	stiff	tɪ	V	Yes	
3	Tibet	tɪ	V	V*	22
4	petition	tɪ	V	V	39
5	attic	tɪ	V	Yes	
6	custom	kʌ	V	V	7
7	skull	kʌ	V	V	22
8	accompany	kʌ	V	V	49
9	caboose	kə	V	V	19
10	academic	kə	V	Yes	
11	duck	dʌ	V	V	20

No.	Word	Trans.	Potential effect category	Actual effect category for %Corr data	Time difference from $D_{max}$ to area of max. change of %Corr
12	citizen	ɪt	VC	Yes	
13	fitness	ɪt	VC	Yes	
14	Italian	ɪt		Yes	
15	committee	ɪt	VC	Yes	
16	unity	ɪt		Yes	
17	bucket	ʌk	VC	Yes	
18	mechanical	ək		Yes	
19	indicate	ək		linear fit	N/A
20	induction	ʌk	VC	wrong slope	N/A
21	muddy	ʌd	VC	Yes*	
22	cadenza	əd		Other	9
23	medicine	mɛ	V	Yes	
24	immense	mɛ	V	V*	27
25	remedy	ɛm	R	Yes	
26	attempt	ɛm	R	Yes	
27	negative	nɛ	V	Yes	
28	tenants	ɛn	R	Yes	
29	saddle	sæ	V	V	9
30	master	æs	F	Yes	
31	Zachary	zæ	V	V	6
32	asthma	æz	F	Yes	
33	shell	ʃɛ	V	wrong slope*	N/A
34	session	ɛʃ	F	F	13
35	fees	fi	V	Other	-4
36	unfeeling	fi	V	Yes	
37	leaf	if	F	F	45
38	relief	if	F	F	6
39	vacuum	væ	V	Yes	
40	ravish	æv	F	F*	14
41	trail	reɪ	V	Yes	
42	fair	eɪr	R	R	-3
43	lever	le	V	V	36
44	elevator	ɛl	R	R	-21
45	yellow	je	V	Yes*	
46	watch	wa	V	Yes	

No.	Word	Trans.	Potential effect category	Actual effect category for %Corr data	Time difference from $D_{max}$ to area of max. change of %Corr
47	chapel	tjæ	V	wrong slope	N/A
48	latches	ætʃ	F	F	8
49	jump	dʒʌ	V	Yes	
50	judge	ʌdʒ	VC	VC	-23
51	bent	nt		Yes	
52	sentiment	nt		Yes	
53	reinterpret	nt		Yes	
54	band	nd		Yes	
55	wander	nd		linear fit	N/A
56	reconditioned	nd		Yes	
57	axe	ks	F	F	20
58	hacksaw	ks	F	F	6
59	unacceptable	ks	F	F	15
60	cats	ts	F	F	25
61	Betsy	ts	F	F	18
62	stop	st		Yes	
63	based	st		Yes	
64	pastime	st		Yes	
65	skate	sk		Other	7
66	mask	sk		Yes	
67	discount	sk		Yes	
68	train	tr	R	linear fit	N/A
69	string	tr		Yes	
70	Detroit	tr	R	R	-79
71	crops	kr	R	R	-47
72	scrap	kr		Yes	
73	acrobat	kr		Yes	
74	drop	dr		Other	32
75	groan	gr		Yes	
76	plain	pl	R	Yes	
77	split	pl		Other	-4
78	twelve	tw	R	Yes	
79	court	rt	VC	Yes	
80	cork	rk	VC	Other	24
81	help	lp	VC	Other	54
82	fans	nz	F	F	6

No.	Word	Trans.	Potential effect category	Actual effect category for %Corr data	Time difference from $D_{max}$ to area of max. change of %Corr
83	dance	ns	F	Yes	
84	fancy	ns	F	Yes	
85	unconcealed	ns	F	Yes	1
86	snow	sn		Yes	
87	Disney	zn		Yes*	
88	farm	rm	R	Yes	
89	corn	m	R	wrong slope	N/A
90	film	lm	R	wrong slope	N/A
91	ranch	nɪʃ	F	F	24
92	flash	fl	R	Other	11
93	fragile	fr	R	Yes	
94	sleep	sl		Yes	
95	Iceland	sl		linear fit	N/A
96	swan	sw	R	R	-40
97	golf	lf	F	F	71
98	wharf	rf	F	F	67
99	false	ls	F	F	29
100	calcium	ls	F	F	10
101	cultural	lɪʃ	F	Yes	
102	marginal	rɪdʒ	F	Yes	
103	optical	pt		Yes	
104	pact	kt		Yes	
105	coughs	fs	F	F	27
106	nerves	vz	F	Yes	
107	amnesty	mn		Other	8
108	garlic	rl	R	Yes	
109	biopsy	a <sup>j</sup> a	VV	VV	-9
110	biography	a <sup>j</sup> a	VV	VV	35
111	biotech	a <sup>j</sup> o <sup>w</sup>	VV	Yes	
112	eon	ia	VV	VV	26
113	diagonal	a <sup>j</sup> æ	VV	VV	22
114	react	iæ	VV	VV	-57
115	tiger	tæ <sup>j</sup>	V	linear fit	N/A
116	bite	a <sup>t</sup>	VC	Yes	
117	data	de <sup>j</sup>	V	Other	-4
118	fade	e <sup>d</sup>	VC	VC	-49



No.	Word	Trans.	Potential effect category	Actual effect category for %Corr data	Time difference from $D_{max}$ to area of max. change of %Corr
119	doubt	a <sup>w</sup> t	VC	VC	-47
120	soybean	o <sup>b</sup>	VC	Yes	
121	toad	to <sup>w</sup>	V	V	55
122	oats	o <sup>w</sup> t	VC	VC	-85
123	courage	kə	V	V	20
124	circle	ək	VC	Yes	
125	button	tɪ		Yes	
126	beetle	tɪ		Yes	
127	apple	pl		Yes	
<b>Japanese</b>					
1	todana	[to]	V	Yes	
2	tatoe'ru	[to]	V	Yes	
3	ka'to	[to]	V	V	17
4	kakari'iN	[ka]	V	Other	-19
5	hakama'	[ka]	V	Yes	
6	sya'kai	[ka]	V	Yes	
7	dama'ru	[da]	V	Yes	
8	midare'ru	[da]	V	V	12
9	ku'da	[da]	V	V	6
10	hotoke'	[ot]	VC	Yes	
11	himoto'	[ot]	VC	Yes	
12	hakobu	[ak]	VC	VC	-12
13	hatake	[ak]	VC	VC	-22
14	ha'yaku	[ak]	VC	VC	-7
15	kadai	[ad]	VC	Yes	
16	hanada'yori	[ad]	VC	VC*	-8
17	ka'nada	[ad]	VC	VC	-29
18	megumi	[me]	V	Yes	
19	tomeru	[me]	V	Other	-2
20	nemui	[ne]	V	linear fit	N/A
21	kemuri	[em]	R	Yes	
22	tabemo'no	[em]	R	R	-27
23	teni'motu	[en]	R	Yes*	
24	soda'tu	[so]	V	Yes	
25	zabu'toN	[za]	V	Yes	
26	syabe'ru	[ja]	V	Yes	

No.	Word	Trans.	Potential effect category	Actual effect category for %Corr data	Time difference from $D_{max}$ to area of max. change of %Corr
27	hokeN	[ho]	V	Yes	
28	zyosei	[os]	F	Other	-25
29	kazari	[az]	F	F	11
30	basyo	[aʃ]	F	Yes	
31	gohoo	[oh]	F	F	42
32	wahuku	[aϕ]	F	Yes	
33	dohyoo	[oç]	F	F	32
34	harada'tu	[ra]	V	V	7
35	yubi'	[ju]	V	V	95
36	kara'i	[ar]	R	Yes	
37	huyoo	[uj]	G	G	17
38	mawari	[aw]	G	G	9
39	tyazuke	[tʃa]	V	Yes	
40	zyokyo'ozyu	[dʒo]	V	Other	-8
41	mati'	[atʃ]	VC	VC	-14
42	tozi'ru	[odʒ]	F	Yes	
43	haNtai	[nt]		Other	-36
44	kaNdo	[nd]		wrong slope	N/A
45	teNkiN	[ŋk]		Other	-11
46	kaNzeN	[nz]	F	Yes	
47	seNsoo	[ns]	F	Yes	
48	keNritu	[nr]		Other*	-16
49	koNyaku	[nj]	R	Yes	
50	kiNtyoo	[ntʃ]	F	F	20
51	sukunai	[sk]		Yes	
52	sikaku	[s'k]		Yes	
53	kitamuki	[k't]		Other	14
54	kokutetu	[kt]		linear fit	N/A
55	kyaku	[kj]	R	R	-23
56	dakyoo	[kj]	R	Yes	
57	hyoo	[çj]	R	Yes	
58	ryokaN	[çj]		Yes	
59	mottaina'i	[t]	LL	LL	-50
60	sakka	[kk]	LL	LL	-50
61	sassoku	[ss]	LL	LL*	-13
62	hassya	[ʃʃ]	LL	Yes	

No.	Word	Trans.	Potential effect category	Actual effect category for %Corr data	Time difference from $D_{max}$ to area of max. change of %Corr
63	teNmetu	[mm]	LL	Yes	
64	aNnaizyo	[mn]	LL	Yes*	
65	tootyaku	[oo]	LL	Yes	
66	keigo	[ee]	LL	LL	12
67	syuukaN	[uu]	LL	Yes	
68	haori	[ao]	VV	Yes	
69	siatu	[ia]	VV	VV	-23
70	kaeri'miti	[ae]	VV	VV	36
71	taiko	[ai]	VV	VV	-8
72	koibito	[oi]	VV	VV	26
73	teNiN	[ēi]	V	V	15
74	hiN	[ij]	R	R	-20
75	maNne'Nhitu	[an]	R	R	-7
76	seNmoN	[em]	R	Yes	

It is also possible to evaluate the #Resp measure for whether it shows these effects when the maximal change in number of responses fails to surround the  $D_{max}$  point. However, when the area of maximal change in the number of responses (#Resp) is not at the same time as the maximal change in the percent correct (%Corr), it is difficult to interpret these results. If the maximal change of the %Corr measure surrounds the  $D_{max}$ , but the maximal change in #Resp falls before the  $D_{max}$ , one cannot readily attribute this to early recognition of a sonorant, even if the transition is into a postvocalic sonorant, because the %Corr measure shows that the sonorant was not recognized early. When the two perceptual measures agree, categorizing the #Resp measure adds no new information. Therefore, rather than categorize the cases in which the #Resp measure failed to have its area of maximal change surrounding  $D_{max}$ , I will analyze the cases in which the two perceptual measures did not have their areas of maximal change at the same location, in section 4.4 below.

## 4.3.4. Statistical tests of results

Table 4.5 summarizes the proportion of words which actually showed a particular effect out of the number which were classified as having the potential for that effect. It also shows the average number of milliseconds by which the area of maximal perceptual change was separated from the  $D_{\max}$ . (As in Table 4.4, negative numbers mean that the word was recognized before reaching  $D_{\max}$ , while positive numbers mean it was recognized after  $D_{\max}$ .)

Table 4.5. Summary of results for potential and actual effects. Number and percentage of words observed to have a particular effect out of the total number of words which could potentially have that effect, and average time from  $D_{\max}$  to the area of maximal perceptual change for words showing that effect.

Effect	English		Japanese	
	Proportion (percentage)	Average time difference (ms)	Proportion (percentage)	Average time difference (ms)
postvocalic stop (VC)	4/16 (25%)	-51	6/9 (67%)	-15
transition into vowel (V)	12/31 (39%)	24	6/20 (30%)	16 <sup>12</sup>
transition into sonorant (R)	5/17 (29%)	-38	4/11 (36%)	-19
transition into fricative (F)	17/25 (68%)	24	4/10 (40%)	26
vowel-vowel (VV)	5/6 (83%)	3	4/5 (80%)	8
geminate segment (LL)	N/A		4/9 (44%)	-25
medial glide (G)	N/A		2/2 (100%)	13

## 4.3.4.1. Percent of words showing each effect

Using a chi-squared test of independence, I tested each effect (except the effects which only apply to Japanese) for whether it is statistically more likely to happen in one language than the other. Because the numbers of words which even have the potential to

<sup>12</sup> For this average and subsequent calculations, the word /yubi/ [jubi] 'finger' was excluded from the group of words showing the late vowel recognition effect. In this word, the area of maximal change in percent correct fell 95 ms after the  $D_{\max}$  point, much later than any other Japanese word with the vowel effect. This is because the  $D_{\max}$  point for this word is at the onset of the /y/, not the change from /y/ to the vowel.

show most of these effects is small, these statistical tests should be considered preliminary. The only effect for which the difference between the two languages in the proportions of words showing the effect was significant was the postvocalic stop effect: postvocalic stops are more likely to be recognized before  $D_{\max}$  in Japanese than in English ( $\chi^2$  (1,  $N=25$ )=4.17,  $p<.05$ ). I believe this reflects only the fact that more of the Japanese words which had postvocalic stops as the second segment of interest had these stops following the vowel /a/, which is rather long. All of the Japanese words for which postvocalic stops were recognized early had this stop after the vowel /a/. Most of the English words which demonstrated the postvocalic stop effect had the stop after diphthongs, as in "fade, doubt, oats." The postvocalic stop effect was never observed in words such as "citizen, fitness, committee," where the postvocalic stop follows a rather short vowel. As /a/ is a very common vowel in Japanese<sup>13</sup>, many of the transitions used were between /a/ and a consonant. Thus, the fact that postvocalic stops are recognized early more often in Japanese than in English is probably an artifact of the composition of the word lists.

The differences between the two languages in proportion of words to show a given effect was not significant for the remaining effects (for the sonorant effect:  $\chi^2$  (1,  $N=28$ )=0.148,  $p>.05$ ; for the vowel effect:  $\chi^2$  (1,  $N=51$ )=0.404,  $p>.05$ ; for the fricative effect:  $\chi^2$  (1,  $N=35$ )=2.33,  $p>.05$ ; for the vowel-vowel effect:  $\chi^2$  (1,  $N=11$ )=0.02,  $p>.05$ ). Thus, neither language is significantly more likely than the other to have vowels or fricatives recognized late, sonorants recognized early, or vowels after other vowels recognized somewhere other than at  $D_{\max}$ . In these cases, also, the phonological systems of the languages caused such differences in the types of transitions included in the word lists that the number of cases showing a particular effect is not entirely comparable.

---

<sup>13</sup> Although long /aa/ is quite uncommon, short /a/ is very common.

#### 4.3.4.2. Distance of area of maximal perceptual change from $D_{max}$

However, for words which do show these effects, one can safely compare the degree to which they are recognized early or late, with less interference from differences in the characteristics of the word lists. Recall from the average times in Table 4.5 that when these effects are present, the postvocalic stop effect, the vowel effect, the sonorant effect, and the fricative effect are all stronger in English than in Japanese. That is, when these effects take place, the area of maximal change in the %Corr measure is on the average farther away from the  $D_{max}$  point in English than in Japanese. I used analyses of variance, with the analysis of weighted means to correct for the unequal sample sizes for the two languages, to compare the amount of time by which the segments were recognized before or after the  $D_{max}$  in the two languages. I tested each effect separately. Thus, each test is a one factor ANOVA, in which the only factor is language. I used the between subjects ANOVA (not repeated measures), since the "subjects" are the individual words which showed these effects, and the words are different in the two languages.

The ANOVA with weighted means showed a significant difference between the two languages for the postvocalic stop effect ( $F(1,8)=10.49, p<.02$ ). That is, when postvocalic stops are recognized early, the tendency for them to be recognized earlier relative to  $D_{max}$  in English than in Japanese is statistically significant. The vowel effect was not significant ( $F(1,15)=1.63, p>.05$ ), but vowels which are recognized late do seem to be recognized later relative to the  $D_{max}$  point in English than in Japanese, with the exception of a very few words. This effect might be significant if tested with more words, and with the environment in each word more closely matched across the two languages. The fricative and sonorant effects were not significantly different in the two languages ( $F(1,19)<1$  for the fricative effect;  $F(1,7)=1.56, p>.1$  for the sonorant effect). Thus, there is only a non-significant tendency for fricatives to be recognized later and sonorants to be recognized earlier in English than in Japanese. I did not test the difference between the two languages for the vowel-vowel effect, since this effect can cause segments to be recognized either

early or late, and is the result of a problem with the measure  $D$ , not of facts about the languages.

#### 4.3.4.3. Results for words not in any category of exceptions

The words which are not predicted to fall into any of the categories of exceptions (those with no notation for the expected category in Table 4.4 above) can provide a further test of the overall hypothesis that listeners make disproportionate use of cues near changing parts of the signal. These are the transitions for which there is no reason to expect the second segment of interest to be recognized anywhere other than at a  $D_{\max}$  point, because cues for the second segment of interest cannot spread into preceding segments, and the second segment of interest is not a type that requires a long portion of the segment for listeners to perceive it (not a fricative or vowel). Consonant-consonant transitions in which cues for the second C cannot spread into the first are frequently of this type. Since these are exactly the transitions for which there is no reason to expect them to deviate from the hypothesis, a larger proportion of these words should follow the hypothesis (by having the area of maximal improvement in perception surrounding a  $D_{\max}$  point) than for the entire experiment.

The results for these words are shown in Table 4.6. The probability of the area of maximal change surrounding a  $D_{\max}$  point by chance was calculated in the same way described in section 4.2.2 above, using only the words which do not fall into any of the categories of exceptions and are more ogival than linear and do not have anomalous slopes. The chance probabilities are also shown in Table 4.6.

Table 4.6. Actual and predicted (chance) numbers and percentages of words which do not fall into any of the exception categories with area of maximal perceptual change surrounding  $D_{\max}$  for each language. Only percent correct (%Corr) measure is included. Number of words is shown relative to the number of words which do not fall into any of the exception categories discussed above and which are more ogival than linear and do not have anomalous slopes.

Type of data		English		Japanese	
		Proportion	Percentage	Proportion	Percentage
%Corr	Actual	25/30	83.3%	3/7	42.9%
	Predicted by chance	12.1/30	40.3%	2.1/7	30.4%

For the English data, the percentage of these words fulfilling the hypothesis is considerably higher than in the experiment overall (the results shown in Table 4.3 above). The fact that 83.3% of these words do have the area of maximal perceptual change surrounding a  $D_{\max}$  point is an encouraging result for the hypothesis of speech perception through dynamic cues. For the English data, the number of words fulfilling the hypothesis is significantly greater than would be expected by chance ( $\chi^2(1, N=30)=23.05, p<.001$ ). For the Japanese data, however, examining only transitions which are not predicted to fall into any of the exceptional categories still results in only 42.9% of such words following the hypothesis that the most improvement in perception will happen at a  $D_{\max}$  point. This is not significantly higher than chance ( $\chi^2(1, N=7)=0.55, p>.05$ ), and is a lower percentage of words than for the entire Japanese experiment.

However, there are very few Japanese words (seven) in which the transition of interest does not fall into any of the exceptional categories. This is because Japanese permits only a few consonant-consonant clusters, which comprise many of the transitions which are not subject to the explanations for the exceptions. Furthermore, in the English wordlist, postvocalic stops after unstressed vowels are not considered to fall into the postvocalic stop exception category, because the unstressed vowels are too short for listeners to recognize the following stop based on cues in the vowel. In Japanese, however, because unaccented moras do not show significant reduction of their vowels, all



postvocalic stops fall into the postvocalic stop exception category. Thus, there are very few words in the Japanese corpus which allow for this test of the hypothesis, and the total number of such words may be too small for a reasonable evaluation of the claim that the area of maximal perceptual change should surround  $D_{\max}$  more often in such words than in the experiment overall.

Furthermore, of the four Japanese words which are not predicted to fall into any of the exceptional categories, but still do not have the area of maximal perceptual change surrounding a  $D_{\max}$  point, three have the mora nasal followed by a consonant as the transition of interest, namely /haNtai/, /teNkiN/, and /keNritu/. These three constitute all of the mora nasal-consonant transition words which are not predicted to fall into any of the exceptional categories. All three are recognized slightly early relative to a  $D_{\max}$  point: /haNtai, teNkiN/ have the most progress in recognition of the stop during the closure for the stop, between the  $D_{\max}$  for the closure and  $D_{\max}$  for the release. /keNritu/ has the most progress toward recognizing the /r/ slightly before the  $D_{\max}$  for the flap (of which there is only one in this case). The /r/ in /keNritu/ was never recognized more than 20% correct, and the data for both /haNtai/ and /teNkiN/ is somewhat noisy. There may be a systematic reason for the early recognition of the post-nasal consonants in these three words, or this may simply be experimental error. However, if these three words are excluded from consideration, 75% of the remaining four Japanese words which do not belong to any of the classes of exceptions do have the area of maximal improvement in perception surrounding a  $D_{\max}$  point<sup>14</sup>. This is more similar to the English result of 83.3%. With such a small number of words involved and no clear explanation for the behavior of the /Nt, Nk, Nr/ transitions, any conclusions on this point are obviously speculative.

---

<sup>14</sup> Despite the extremely low number of tokens, this result is significantly greater than would be expected by chance for these words ( $\chi^2$  (1, N=4)=3.86,  $p<.05$ ). However, with such a small number of tokens, the chi-squared test is probably not reliable.

What is clear from the analysis of transitions which are not predicted to fall into any of the categories of exceptions is that for the English data, when there is no reason based in the availability of perceptual cues for listeners to behave otherwise, the area in which listeners quickly become able to perceive the segment is very often the area of maximal spectral change in the acoustic signal. This result from a subset of the data strengthens the results from the data for all words in the experiment, which showed that the area of maximal perceptual change is at the point of maximal spectral change more often than it would be by chance, even though there are reasons to expect many types of transitions to show different results.

#### 4.4. Analysis of cases in which the perceptual measures differ

##### 4.4.1. Rationale for comparing the timing of perception of segments and recognition of words

Models of spoken word recognition, while they differ on the question of what group of words listeners choose a lexical item from (cohorts versus lexical neighborhoods, for example), all assume that listeners gain phonetic information from a signal over time<sup>15</sup> and in some way use this information to narrow down the group of words from which they choose. Some propose that this involves (at least partly) perception at the level of the phoneme (McClelland and Elman 1986), others propose that perception takes place through distinctive features without an intermediate level for recognition of phonemes (Lahiri and Jongman 1990, Warren and Marslen-Wilson 1987, 1988). However, in any model of spoken word recognition, one would expect listeners' progress in processing of phonetic information to result in progress toward recognizing the word. That is, when listeners become able to perceive some phonological unit, this should, at approximately the same point in time, lead to a change in the words they think they might be hearing.

---

<sup>15</sup> Lexical neighborhoods do not provide a model of how spoken word recognition takes place over time, but they do involve listeners extracting phonetic information from the signal to recognize a word.

#### 4.4.2. Overall results of comparing the two perceptual measures

As discussed above, I analyzed my data in two ways, through the number of responses (#Resp) measure, which is intended to measure spoken word recognition, and through the percent correct (%Corr) measure, which reflects perception of individual segments. This allows for comparison of the timing of perception of the target segment (the second segment of the transition of interest) and recognition of the word in each case<sup>16</sup>. To do this, I compared the location of the area of maximal change in the %Corr measure to the area of maximal change in the #Resp measure. Table 4.7 shows the results of this comparison.

Table 4.7. Number of gates separating the area of maximal change in %Corr from the area of maximal change in #Resp. Results are in percent of words in the experiment to fall into each category. In parentheses: the percentage of words which have an area of maximal change for both perceptual measures (those not in the "not applicable" category) in each category.

Number of gates apart the areas of maximal change in the 2 perceptual measures are	English	Japanese
0 (areas of max. change are the same)	41 (42.7%)	24 (47.1%)
1 (areas of max. change are contiguous)	32 (33.3%)	16 (31.4%)
2 (20 ms intervene between the two areas of max. change)	11 (11.5%)	8 (15.7%)
3 (40 ms intervene)	10 (10.4%)	1 (2.0%)
4+ (60 ms or more intervene)	2 (2.1%)	2 (3.9%)
not applicable (one or both of the perceptual measures is linear or has the wrong slope)	31	25

Words which have an anomalous slope or which are better fit by a line than by an ogival curve for *either* perceptual measure are excluded from this calculation, because both perceptual measures must have an area of maximal improvement in order to compare the locations of these areas. Of the remaining words, 42.7% of the English words and 47.1% of the Japanese words have the areas of most improvement in perception of the target

<sup>16</sup> By "recognition of the word" here, I do not mean that listeners become sure of exactly which word they hear. I mean only that they narrow down the words which are possibilities, not that they all agree on a single correct response. This is a matter of progress toward recognizing the word, rather than of complete recognition.

segment (%Corr) and recognition of the word (#Resp) at the same place. Figures 4.5, 4.6, 4.14, and 4.15 above are examples in which the area of most change in the two perceptual measures is at the same time<sup>17</sup>. In these cases, one can assume that listeners' success in narrowing in on a smaller number of lexical items is a result of their progress toward correctly perceiving the target segment (the second segment of the transition of interest).

In a further 33.3% of the English words and 31.4% of the Japanese words, the areas of maximal change in the two perceptual measures are at neighboring gates, as in Figures 4.18, 4.19, and 4.21 above, for example. Even though the area in which listeners make the most progress toward perceiving the segment is not the same as the area in which they make the most progress toward recognizing the word in these cases, I do not believe that these show spoken word recognition happening over a different time course than perception of individual segments. The method of locating the area of maximal change of the fitted ogival curve is meant to provide a unique area which can be considered as the area during which listeners make the most progress in perception (or word recognition). The need to use the fitted curve in order to locate this unique area was discussed in Section 4.1.4 above. However, improvement in perception or word recognition is often relatively gradual. While in some words, nearly all the improvement in perception takes place between one gate and the next (a period of 20 ms), in many other words, this improvement takes place over a period of several gates. This is demonstrated in Figure 4.23. When the improvement in perception is gradual, locating a unique 20 ms area during which the fitted curve has maximal change may be misleading, as it eliminates the difference between words of the sort shown in Figure 4.23. This implies that the unique 20 ms area of maximal change in the perceptual measure is the only area important for perception, while it may be only the center of a larger area, all of which is important for perception. Furthermore, when the slope of the data is gradual, small adjustments in the fitting of the

---

<sup>17</sup> In investigating the agreement of the two perceptual measures, location of the  $D_{\max}$  point is not at issue, but the figures illustrating comparison with  $D_{\max}$  location can also be used for this purpose.

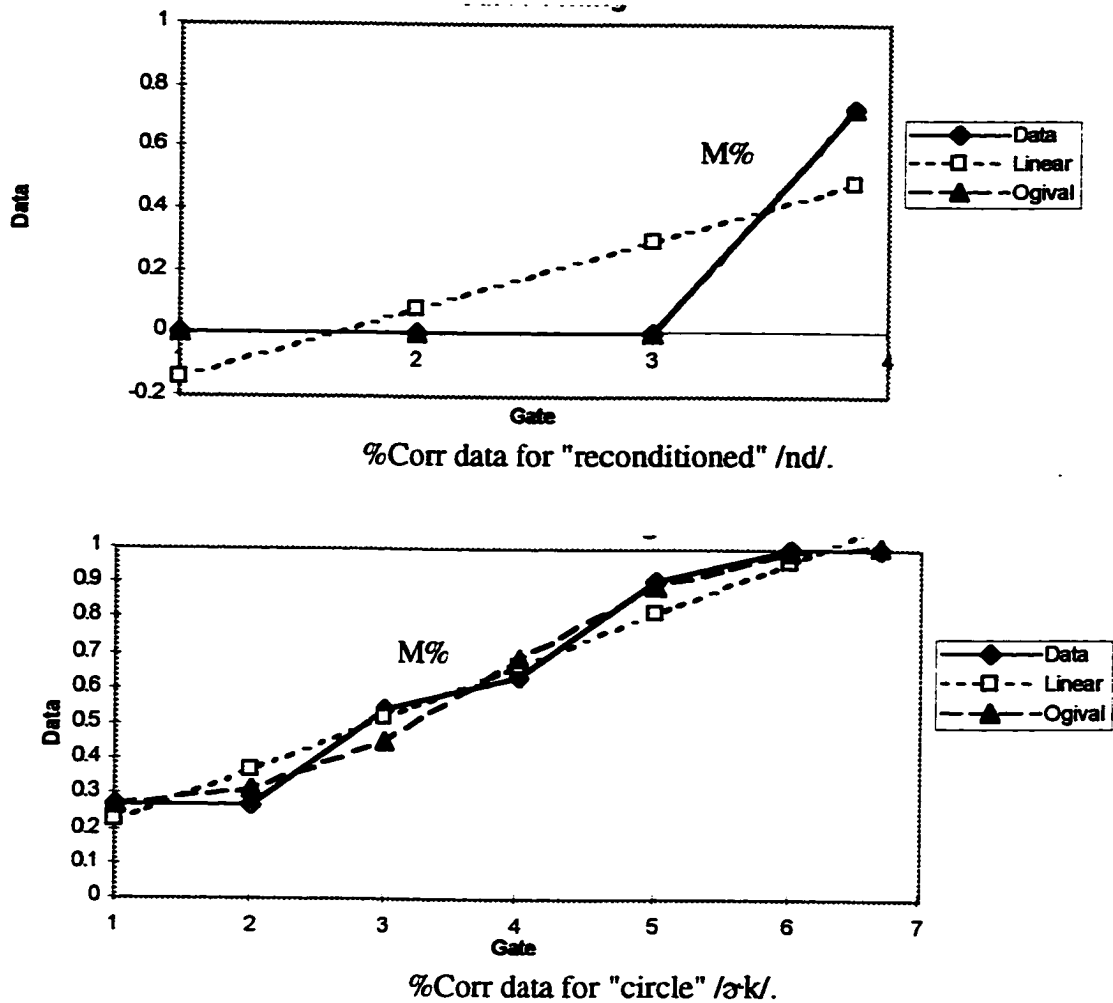


Figure 4.23. Percent correct (%Corr) data for two words. "Reconditioned" /nd/ has all of the improvement in perception between two contiguous gates, while "circle" /ɔ:k/ has a more gradual improvement in perception, taking place over approximately 4 gates. The area of maximal change in the %Corr measure for "reconditioned" is unambiguously between the third and last gates. For "circle," the area of maximal change in the fitted curve is between gates three and four, but gates four to five would have been an equally likely area to have the maximal change.

curve may shift the area of maximal change in the fitted curve from between the third and fourth gates to between the fourth and fifth gates, for example.

If a word has a gradual slope for either the %Corr measure, the #Resp measure, or both, the areas of maximal change in the fitted curves of the two measures may be contiguous instead of identical, even if the progression of improvement in the two measures is rather similar. An example of a word in which the areas of maximal change for the two perceptual measures are contiguous because the method forced a choice of one 20 ms area as the area of most change, even though the curve is rather gradual, is shown in Figure 4.24. Even if neither perceptual measure has a gradual slope, and the locations of the areas of maximal change in both measures appear unambiguous, a lack of agreement by one gate between the two measures probably does not reflect a real difference in when progress is made toward perceiving the target segment and when toward recognizing the word, as one noisy data point can move the area of maximal perceptual change by a gate. Therefore, I believe it is justified to conclude that the 76.0% of English words and 78.4% of Japanese words which have the area of maximal change in the two perceptual measures either at the same place or at contiguous areas show progress in recognition of the word happening at the same time as progress in perception of the target segment.

I examined the responses to words where the two perceptual measures disagreed by more for each word individually. There were 23 such words in the English data and 11 in the Japanese. For each such word, I listed all the different responses given to the word at each gate, and how many subjects gave each response. I then attempted to determine, in a qualitative way, whether the area of maximal reduction in number of different responses corresponded to any change in the phonemic content of the responses given. That is, does the area of most change in the #Resp measure reflect the subjects' having perceived some phonetic information other than the information necessary to perceive the target phoneme? (The point at which the most listeners begin to give responses with the target phoneme correct, of course, is the area of maximal change for the %Corr measure.)

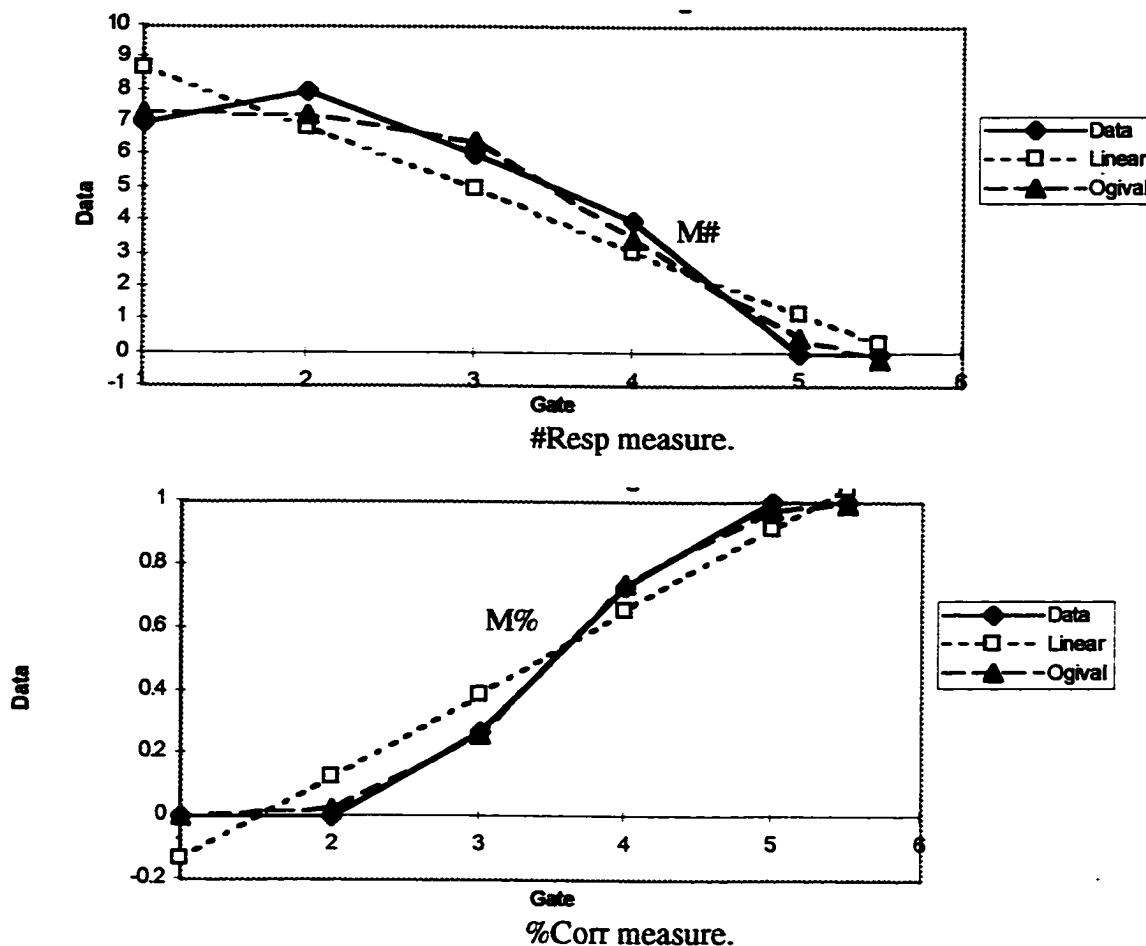


Figure 4.24. #Resp and %Corr curves for "session" /εʃ/. The area of maximal change for the number of responses data (labeled M#) is between the fourth and fifth gates, while for the percent correct data it is between the third and fourth gates (M%). Since the data for both perceptual measures has a rather gradual slope, the lack of agreement between the two measures probably does not reflect a difference in timing of perception of the /ʃ/ and recognition of the word, but rather the fact that the improvement in perception does not take place over just 20 ms.

#### 4.4.3. Cases where maximal change in #Resp represents perception of something other than the target segment

In many cases, it appears that the area of maximal change in number of responses (#Resp) does fall at a point where listeners have perceived something other than the target phoneme, or just one feature of the target phoneme. Table 4.8 shows responses to the word "Italian" /t/ as an example. Here, the area of most change in the %Corr measure (percent to get the /t/ correct) occurs between the first and second gates, although not all listeners have perceived the /t/ correctly by the second gate. The greatest change in the number of different responses, however, occurs between the fifth and sixth gates.

Table 4.8. Responses to "Italian" /t/.  
Number of subjects to give each response appears after each word.

Gate 1 (during /t/)	Gate 2 (just after /t/ closure)	Gate 3 (at /t/ release)	Gate 4 (during /t/ aspiration)	Gate 5 (during aspiration)	Gate 6 (near end of aspiration)						
age	1	breath	1	aphex <sup>18</sup>	1	it	3	attack	1	Italian	11
aha	1	hat	1	attack	1	Italian	4	etiquette	1		
air	1	heck	1	attention	1	italic	1	it	3		
frame	1	hit	3	Etcheverry	1	Italy	1	Italian	3		
hello	1	hypnotic	1	excite	1	iterate	2	italics	1		
hip	1	it	4	history	1			Italy	1		
hit	2			it	1			iterate	1		
into	1			itch	2						
pet	1			its	2						
talk	1										

Responses at the first gate are rather random. Subjects hear only approximately 34 ms of the signal at this point. At the second gate, eight of the eleven subjects give responses with the target /t/, resulting in the maximum change in percent correct (%Corr) falling here. However, responses still vary quite widely, especially at the third gate. Between the fifth and sixth gates the number of different responses drops severely, so the maximum change in number of responses (#Resp) falls here. It appears that at the sixth

<sup>18</sup> Listeners sometimes coined nonce words or gave non-word responses similar to real words. The subject who gave this response said that ['eɪfeks] was intended, so this is not a misspelling of "apex."



gate, listeners have gained considerably more information about the word than the identity of the /t/. The duration of aspiration noise may have allowed them to realize that the second syllable is stressed. The effect of the following vowel quality on the aspiration noise may even have allowed them to perceive the following vowel. If they perceived that the second syllable had to be stressed, this would eliminate all of the responses given at the previous gate except "attack, Italian, italics." The absence of "attack" and "italics" at the sixth gate could be due to chance. Correct perception of the target /t/ does not lead to as great a change in number of responses (#Resp) as the later correct perception of stress or the following vowel seems to, so the areas of maximal change of the two perceptual measures do not agree.

Table 4.9 shows a similar result for the word "mechanical" /ək/. The area of maximal change in the %Corr measure is between gates one and two, while the area of maximal change for #Resp is between gates four and five.

Table 4.9. Responses to "mechanical" /ək/.  
Number of subjects to give each response appears after each word.

Gate 1	Gate 2	Gate 3	Gate 4	Gate 5	Gate 6
make	2 McDonald's	1 MacDonald	1 McDonald	1 mechanic	9 McCarthy
McDonald's	1 McDougal	1 McDonald's	5 McDonald's	1 mechanical	1 McKinley
me	2 McGuyver	1 McGregor	1 McKay	1 microphone	1 mechanic
mechanic	2 mechanic	5 mechanic	2 mechanic	4	mechanical
mechanical	1 mechanical	2 mechanical	1 mechanical	2	mechanics
Michelangelo	1 miniature	1 Mick	1 metathesis	1	
mill	1		mitt	1	
mine	1				

Here, listeners correctly perceive the target /k/ at the second gate, except for the one response of "miniature," producing a large change in percent correct (%Corr). However, they continue to give a large number of different responses with /k/ through the fourth gate. At the fifth gate, the number of different responses drops, and although it rises again somewhat at the last gate, it appears that this drop at the fifth gate represents a real change in what was perceived. From here, the previously common responses "McDonald,

McDonald's" disappear. I believe at the fifth gate, which falls during the aspiration of the /k/, listeners perceive that the /k/ must be released into a vowel (or sonorant, as in "microphone") because of its aspiration noise. Thus, responses with word final /k/ or /k/ released into another stop are ruled out.

Table 4.10 illustrates the case of "crops" /kr/, in which the maximal change in percent to get the /r/ correct occurs between the first and second gates, but the most change in number of responses (#Resp), although it is small, is between gates four and five.

Table 4.10. Responses to "crops" /kr/.  
Number of subjects to give each response appears after each word.

Gate 1	Gate 2	Gate 3	Gate 4	Gate 5
ahead	1 crane	1 can	1 coo	1 Christ
bee	1 crap	1 Christ	1 crack	1 crash
but	1 crash	2 crook	1 cream	1 crush
can	1 creep	1 cross	1 creep	1 crutch
could	1 crew	1 crow	1 crook	2 cry
cry	1 cross	1 cruise	1 cross	1 curb
dock	1 crunchy	1 crutch	1 crouton	1 curse
fight	1 cryptic	1 cry	1 crush	1 curt
no answer	1 cuff	1 crypt	1 crutch	1 truck
tug	1 krill	1 crystal	1 purchase	1
what	1	1 quill	1	

Here, most listeners begin to perceive the target /r/ at the second gate. It is interesting that they perceive the initial /k/ at the same point: only three listeners give responses with initial /k/ at the first gate, but all do at the second gate. However, there are so many English words beginning with /kr/ that there is little change in the number of responses (#Resp), even though responses at the first gate are rather random and ten out of eleven responses at the second gate begin with /kr/. This is a case in which the cohort of possible words beginning with the transition of interest (which is word initial) is too large, leaving listeners so many possibilities even when they have perceived the second segment of the transition of interest that the number of responses (#Resp) does not decrease. Between the fourth and fifth gates, the change is also not large, but listeners may be narrowing down the quality of the following vowel at this point. At the third and fourth gates, responses

included words with /i, u, ʊ, o<sup>w</sup>, ɪ/ as the nuclear vowel as well as responses with /æ, a, ʌ, ɔ, a<sup>j</sup>/. At the fifth gate, however, the only vowels appearing in the first syllable of the responses are /a<sup>j</sup>, æ, ʌ, ɔ/. Listeners may have heard enough phonetic cues at this point to be able to perceive the following vowel as either mid central or low.

All of these are cases in which the area of most change in #Resp reflects listeners' perception of some phoneme or distinctive feature, but not of the target phoneme. The area of maximal change for the %Corr measure is located where they make the most progress toward perceiving the target phoneme, since the %Corr measure is calculated relative only to the target phoneme. Another example of this is "petition" /tɪ/, in which the maximal change in number of responses is where listeners begin giving almost exclusively responses in which the vowel of the first syllable can be devoiced or deleted ("petite, potato, petition, petunia, pituitary"). In this case, there are large drops in the number of different responses both at that point and at the area of maximal change of the %Corr measure, but an ogival curve can only have one decrease, not two separate falls with a flat area between them<sup>19</sup>.

In "custom" /kʌ/, the largest change in number of responses (#Resp) is where listeners begin to perceive the /k/ correctly, and there is very little change in the number of responses when they later perceive the vowel correctly. In "discount"<sup>20</sup> /sk/, it may be that the maximal change in the number of responses occurs where the most listeners begin giving responses with the second syllable unstressed (ruling out "discover, discovery"), but this is not completely clear. In "twelve" /tw/, there is a large change in number of responses between the penultimate and ultimate gates. At the last gate, all but one response

---

<sup>19</sup> This is one of the very rare cases in which the ogival curve may not have been the best type of curve to fit to the data. A polynomial curve, which would allow this type of double fall, could not be used, because a polynomial curve with 0 for most of its coefficients is a straight line, and thus a polynomial curve could fit even linear data.

<sup>20</sup> The stimulus was produced with first syllable stress.

has a postvocalic /l/. It may be that cues for the /l/ have spread so far that listeners can recognize the /l/ even though they have not perceived the quality of the intervening vowel yet. In "bite" /a<sup>h</sup>t/, there is a large change in the number of responses, after which almost all listeners respond either "bike" or "bite." However, "bite" responses take over from "bike" only gradually, so the maximal change in %Corr falls later. Here the maximal change in #Resp probably reflects the perception of the postvocalic obstruent as a voiceless stop, the place of which is not yet perceptible. In "soybean" /o<sup>h</sup>b/, the largest change in number of responses is where listeners perceive the /o<sup>h</sup>/ correctly. Because the initial gating point for a diphthong was placed in the middle of the steady state of the first vowel quality, listeners had not yet perceived the diphthong correctly at the first gate.

#### 4.4.4. Cases in which the cohort is too large

For the Japanese word /sodatu/ [sodatsu] 'to grow up,' the maximal change in number of responses (#Resp) appears to reflect listeners' convergence on responses with a postvocalic alveolar. This is similar to the cases of perception of a segment other than the target one discussed above. However, there is an additional factor for this word: there are so many Japanese words beginning with /so/ that there is little change in the number of responses when listeners begin to perceive the target /o/ correctly. This latter problem, of the cohort being too large even once the target segment has been perceived, is relatively common in the Japanese cases of mismatch between the two perceptual measures. /taiko/ [taiko] 'drum' and /hyoo/ [çjoo] 'list' show the same problem, as there are large numbers of Sino-Japanese compound words beginning with /tai/ or /hy/ (and even /hyo/). The underlying issue is that Japanese has a relatively small phonemic inventory and thus large numbers of words beginning with the same strings, with the added problem of productive compounding of Sino-Japanese stems. This problem also appears in English word initial

sequences, though, as in "crops" above. To solve this problem, one could have many more subjects hear each stimulus, so that the number of subjects considerably exceeds the number of common words in the cohort. While it would be desirable for other reasons as well (to help with the overall noisiness of the data) to have many more subjects in each condition, this is impractical. Alternatively, one would have to choose words which do not have a very large cohort at the point of the second segment of interest. Still, the number of words showing this excessively large cohort problem, even in the Japanese data, is not large. One should remember that the words under discussion in this section are only the ones in which the two perceptual measures did not agree, that is, a subset of the data which failed to follow hypotheses.

#### 4.4.5. Perceptual mismatch due to problems with the #Resp measure

Determining what features or phonemes listeners might be perceiving at the area of most change in the #Resp measure is a qualitative analysis, and I do not propose any method to test statistically whether there is some feature which listeners perceive at that point or not. Therefore, I will not claim that any certain percentage of the words in which the two perceptual measures do not match are due to perception of a segment or feature other than the target segment. However, for some words I was not able to identify any phoneme or feature which listeners began to perceive correctly at the area of most change in number of responses. In a few cases, I believe the lack of agreement between the two perceptual measures is solely a result of chance (noisy data), problems in the curve fitting method, or problems with the number of different responses as a measure.

There are two primary problems with the #Resp measure. First, it does not reflect whether a particular response was given by just one subject or by many, except that when a large number of subjects give the same response, the entire group of subjects cannot give as many different responses. However, this is not always adequate. In the word "nerves" /vz/, at the second gate, ten subjects responded "nerve" and one responded "nerves." By

the fourth gate, ten responded "nerves" and only one responded "nerve." However, since exactly the same two responses were given, there is no change whatsoever in the #Resp measure. Furthermore, listeners occasionally give responses which diverge widely from the segmental content of the stimuli, even when ample phonetic information is available. A listener may have coughed just as the stimulus was played, or may simply have been thinking of something else, or gotten tired as the test went on. Unfortunately, the #Resp measure will count a highly anomalous response as equal to a more typical response, even though only one listener will have given the anomalous result.

The second problem, which is related, is that the #Resp measure is strongly influenced by individual variation. Some listeners are more likely than others to give morphologically complex responses where a mono-morphemic response would be possible ("attempting" rather than "attempt" to the stimulus /ətem/). Some listeners also have larger lexicons than others, and are more likely to give low frequency words as responses. If several listeners with large lexicons or listeners who favor morphologically complex responses happen to hear the same stimuli, this can lead to a higher number of responses for the gate they heard. By chance, this will sometimes be the case, even though listeners were assigned to conditions randomly. Tables 4.11 and 4.12, showing responses for "soybean" /oʊb/ and "attempt" /em/, illustrate these problems.

Table 4.11. Responses to "soybean" /oʊb/. Number of subjects to give each response appears after each word. Maximal change in %Corr is at gates 4-5, maximal change in #Resp at gates 1-2.

Gate 1	Gate 2	Gate 3	Gate 4	Gate 5
found	1 soil	1 soil	1 sleep	1 soy
slave	1 soy	7 sordid	1 soy	4 soybean
soil	3 soybean	2 soy	7 soybean	4 soybeans
soy	2 soysauce	1 soybean	2 soysauce	2
soybean	2			
sweater	2			

Gate 6	Gate 7	Gate 8
soy	3 soy	3 solder
soybean	8 soybean	8 soy
		soybean
		soybeans
		soydenol

Although "soybean" does have a reason for the location of the maximal change in #Resp, as discussed above, what is of interest here is the increase in number of responses at the last gate. At this point, most listeners have perceived the target /b/ correctly, and the number of listeners giving responses with a /b/ does not change between the penultimate and ultimate gates. However, relative to the two preceding gates with only two responses each, the increase to five here seems quite large. "Soldier" is a somewhat anomalous response, especially since it is the first response without the correct vowel /oʊ/ since the fourth gate. The appearance of "soybeans" at this gate is chance, as there is no way that some aspect of the signal could have ruled it out at previous gates but made it more likely here. The response "soydenol"<sup>21</sup> is probably not part of most listeners' lexicons. Thus, the increase in the #Resp data at this final gate is purely a matter of chance.

<sup>21</sup> Perhaps the brand name of a drug?

Table 4.12. Responses to "attempt" /εm/. Number of subjects to give each response appears after each word. Maximal change in %Corr is at gates 4-5, maximal change in #Resp at gates 6-7.

Gate 1	Gate 2	Gate 3	Gate 4	Gate 5					
attack	1	attack	1	attack	2	attack	1	attempt	8
attempt	5	attempt	7	attempt	1	attempt	6	attempts	1
attend	2	attend	3	attend	6	attend	3	attend	1
attention	3		attention	2	attention	1	attention	1	

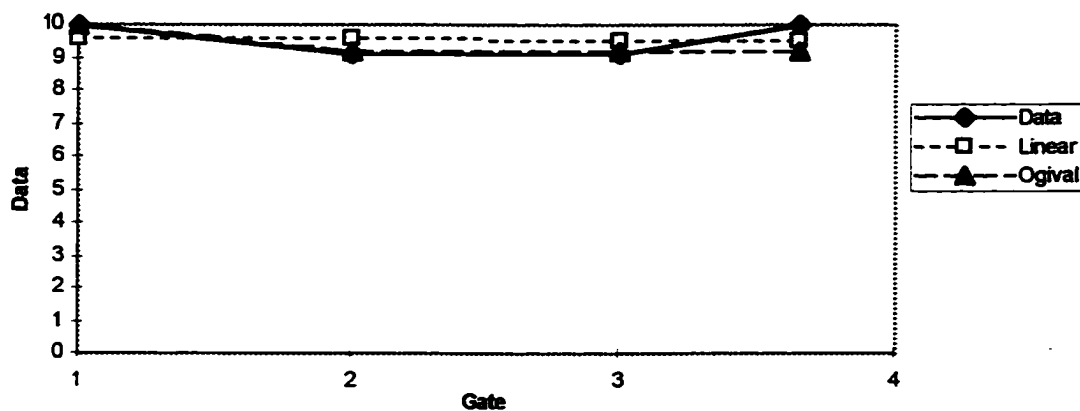
Gate 6	Gate 7		
attempt	8	attempt	10
attempted	1	attend	1
attempting	1		
attend	1		

For "attempt," there is no change in the number of responses (#Resp) at the area of maximal change in percent correct (%Corr), because the previously common response "attack" is replaced by additional forms morphologically related to "attempt" ("attempts, attempting"). The only large change in the number of responses is at the final gate, where all morphologically related forms appear to drop out, leaving only the stems "attempt, attend." However, at this point in the signal, halfway through the /m/, phonetic cues probably cannot rule out "attempts, attempted, attempting." (Some of the segments might have slightly shorter durations in the longer forms "attempts, attempting" than in "attempt," but it seems unlikely that listeners have perceived such a durational cue and realized that the word is the shorter form.) Probably, the final gate simply had listeners who prefer to give morphologically simple responses. This is a danger of using different listeners for each gate, but considering the strong arguments against the successive presentation method discussed in Section 1.4.5, it is unavoidable.

#### 4.4.6. Cases in which curve fitting is too successful

Other cases in which the two perceptual measures do not agree are the result of problems with the curve fitting, or perhaps with the curve fitting being too successful, even when the data is questionable. Figure 4.25 shows the #Resp data for the word /megumi/





Errors:  
 Linear: 0.91  
 Ogive: 0.81

Figure 4.25. #Resp data for /megumi/ 'grace.' This data is included in the calculations because the slope of the linear regression line is negative, and the ogival curve has a lower error than the linear regression line does. This is because the ogival curve fits the data very closely at the first three points, diverging only at the last point, while the regression line does not match any of the points closely. However, the slope of the linear regression line is only negative because the final gate was less than 20 ms from the penultimate gate. Had all the points been spaced equally, the slope would have been 0, and the word would have been excluded as having an anomalous slope.

'grace.' Here, the area of maximal change (decrease) in the fitted curve is between the first and second gates. However, it is obvious from the raw data that this does not really represent the area during which listeners made the most progress in narrowing down the group of words from which they were choosing. In reality, listeners made no progress toward recognizing the word during this time, remaining at a very high number of responses throughout. This is because listeners did not perceive the /e/ accurately even by the last gate, at which they were giving exclusively responses beginning with either /me/ or /mi/. There are many words beginning with those sequences, though, so even though listeners have narrowed the possibilities down to a non-low front vowel, the #Resp data does not decrease. As explained in the figure, the slope of the linear regression line of this data is very nearly 0, but since it is slightly negative, the word is included in the calculations. An ogival curve can be fit quite closely to this data, with the exception of the last point, and can give a better fit than the linear regression line, as shown in the figure.

Thus, one identifies an area which has the most change in the fitted curve, but it does not really represent a change in perception, so it is not surprising that it fails to agree with the %Corr measure. This particular problem could be avoided by setting an arbitrary minimum change in #Resp, and excluding data which does not change by at least that much. However, if a word has ten gates instead of four, a small change in #Resp (which is not reversed at a later gate as the change in Figure 4.25 was) would probably be meaningful. Because the data represents a wide variety of environments, including stimuli with widely varying cohort sizes, it is difficult to set any arbitrary criteria without excluding much meaningful data as well.

#### 4.4.7. The need for more subjects

The results presented in this section may seem discouraging, especially those regarding chance variation in the #Resp data and difficulties with the fitted curves. However, one should keep in mind that those problems provide the reasons for the failure

of predictions in a very small subset of the data, namely those words in which the two perceptual measures have their areas of maximal change in very different locations *and* examination of the individual responses does not show that the change in #Resp is a result of perceiving some segment or feature other than the target segment. There are perhaps ten such words in the English data and eight in the Japanese data<sup>22</sup>. Many of those Japanese words are the result of the cohort being too large even after the second segment of interest is perceived, as discussed above. Thus, the number of words for which the method of analyzing results failed is quite small.

Furthermore, if a very large number of listeners could be used for each stimulus (perhaps 50 instead of the 11 for English and 12 for Japanese used here), I expect that many of the remaining difficulties would disappear. A very large number of listeners would solve the problem of the cohort still being so large at the second segment of the transition of interest that the number of responses remains high throughout the transition. A larger number of listeners for each stimulus would probably also ameliorate the effect of a few listeners in one group choosing to use or not use morphologically complex words or low frequency words not in other listeners' lexicons. With a larger number of listeners, the distribution of such listeners among groups is likely to be more even. Some discrepancies between the two types of data would remain even with many listeners, such as some cases of the maximal change in #Resp reflecting perception of a segment other than the target one. Such cases, however, reflect real perceptual effects, not noise in the data.

The more important result from this section is that when progress toward perception of the target segment and progress toward recognition of the word do not happen at the same time, as one would expect them to, a common reason for this is that progress toward recognition of the word reflects perception of something other than the target segment. In

---

<sup>22</sup> As discussed above, I have at this point no quantitative way to determine whether or not there is an explanation for a particular word in terms of the number of responses change reflecting perception of some segment other than the target.

some cases, listeners have succeeded in perceiving one or more distinctive features of the target segment, but not all, and this causes them to narrow down the group of words from which they choose responses. In others, they perceive an entire segment other than the target one, perhaps a segment later in the word which has cues that spread leftward (as /l/ in "twelve" /tw/), perhaps the segment before the target one if they are not able to perceive it accurately until late in the segment (as in /o<sup>j</sup>/ of "soybean" /o<sup>j</sup>b/). It is well known that phonetic cues are distributed temporally in the speech signal, sometimes spreading rather far from the segment they supply cues for. Furthermore, different vowels, for example, require different amounts of time in order for listeners to identify them (Lang and Ohala 1996). The results presented in this section reflect these phenomena through the comparison of segment perception and spoken word recognition.

#### 4.5. Rise time of perceptual curves

##### 4.5.1. Reasons for examining the rise time

As discussed in the previous section and shown in Figure 4.23, the time span over which listeners perceive a segment correctly varies. Some words show all the improvement in perception of the target segment happening between contiguous gates, a period of 20 ms. Others show a much more gradual change. This difference could reflect systematic differences in the word list. For example, the word list includes words with the transition of interest in either the first or the second syllable of the word. This manipulation was included primarily in order to make the circumstances of spoken word recognition somewhat more natural, by not asking listeners to perceive only segments in the first syllables of words. (I also thought that the inclusion of clearly polysyllabic stimuli would encourage listeners to consider polysyllabic responses for other stimuli as well, perhaps increasing the variety of responses given.) That is, no strong predictions about differences in the behavior of first and second syllable transitions were made. However, it is possible that phonetic cues in the second syllable could be used in spoken word recognition more

quickly than phonetic cues near the beginning of the word, since the cohort of words from which to choose is smaller in the second syllable. This might be reflected as a quicker improvement, spanning fewer gates, for transitions in the second syllable.

Stress on the vowel of the transition of interest (in CV and VC transitions) was also manipulated for English. This manipulation was also used primarily in order to include a more representative variety of transitions. However, previous research using other experimental paradigms (Cutler and Butterfield 1992, Cutler and Norris 1988, McQueen et al. 1994, Cutler et al. 1996) has shown the importance of stress for English listeners' spoken word recognition and speech segmentation<sup>23</sup>. The duration of the period during which listeners make progress toward perceiving a segment might depend partly on whether the vowel of the transition is stressed or not, although it is not entirely clear what this effect might be. The position of the transition of interest relative to the fall for pitch accent in Japanese was also manipulated, and if there is any effect of English stress, it should be compared to the effect of Japanese pitch accent, as Fujimura et al. (1978) showed that English stress, but not Japanese pitch accent, affected weighting of cues for intervocalic consonants.

Furthermore, because of the foot dependent differences in speech segmentation shown by Cutler et al. for English (Cutler and Butterfield 1992, Cutler and Norris 1988, McQueen et al. 1994, Cutler et al. 1996), the position of CC transitions relative to syllable and foot boundaries was manipulated in the English word list. For CC transitions which cannot be syllable onsets (/nd, ns/ etc. but not /sk, sn/ etc.), one stimulus had the CC transition entirely within a syllable ("band"), one had it split across a syllable boundary but within the same foot ("wander"), and one had it split across a foot boundary

---

<sup>23</sup>Cutler (1986) shows that English listeners make little use of stress for spoken word recognition at early stages of the word, in the vowel of the first syllable. She relates this to the fact that there are very few word pairs in English distinguished only by stress with no vowel quality difference, so that use of stress in spoken word recognition does not provide the listener with much more information than the segmental cues do. However, the other work by Cutler and colleagues mentioned here shows that English listeners do make use of stress to help them locate word boundaries, a task for which stress does provide useful information (Cutler and Carter 1987).

("reconditioned"). Cutler et al. find differences in reaction time depending on whether stimuli are split across foot boundaries or not, so although the experimental designs are not the same, one might expect to find a difference in the speed with which phonetic cues are used to perceive a segment depending on its location relative to such boundaries. For example, if English listeners perceive speech in foot sized units, they might make less use of early cues for the /d/ in an /nd/ sequence when there is a foot boundary between the two segments. This might cause the /d/ to be perceived over a shorter window. It is not clear that the effects on reaction time Cutler and her colleagues have found will also be present for the timing of perception of segments, but since the psycholinguistic results on this subject are extensive, stemming from a variety of experimental paradigms, it is important to test for such effects in this experiment as well.

For all of these potential sources of effects (first vs. second syllable, stress/pitch accent, suprasegmental unit boundaries), one could analyze the already presented data on alignment of maximal change in the perceptual measures with  $D_{\max}$ . However, the number of gates over which listeners' perception improves can provide an additional useful measure. To analyze the data in this way, I calculated a measure of the rise time of the fitted ogival curve for the %Corr data.

#### 4.5.2. Method of calculating rise time

Rise time is a measure of how quickly the value of some parameter changes. It is usually calculated by finding the amount of the increase in the parameter (the maximum value minus the minimum), and then finding the duration from the point where the value first reaches 10% or 20% of the total amount of increase to the point where the value first reaches 90% or 80% of the total increase. The use of 10% to 90% (or 20% to 80%) instead of the entire duration of the increasing portion of the data is important, because if the data is at all noisy, the noise could affect the duration of the entire increase greatly.

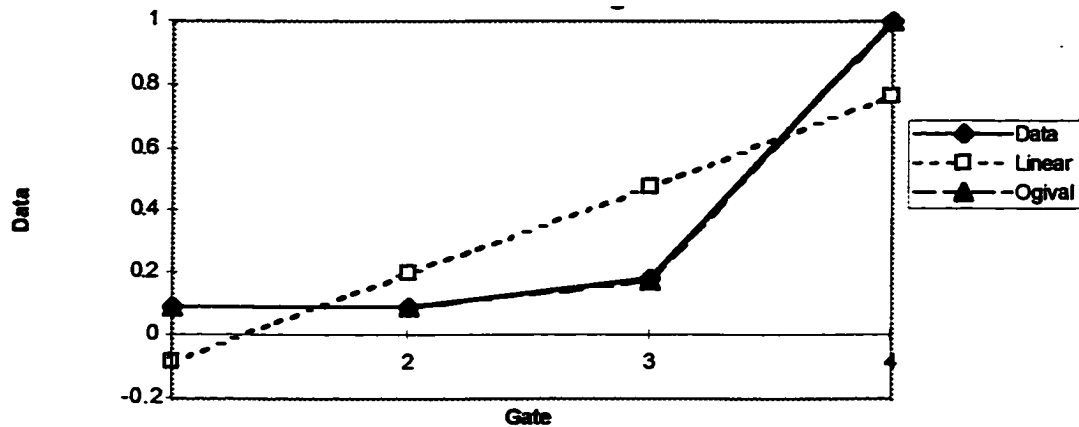
I calculated the rise time of the %Corr data by using the fitted ogival curves, for the same reasons as were given for the use of the fitted curves in general. Because the fitted curves are not noisy at all, the particular cutoff points used for the rise time calculation are of less concern than they would be in calculating rise time directly from raw data. I adapted the usual calculation of rise time slightly, by finding the gate before the fitted curve first reaches 10% of the entire increase and the gate after the curve first reaches 90% of the entire increase<sup>24</sup>. I then took the difference between those two gate numbers as the rise time for the %Corr measure. An example is shown in Figure 4.26 for a word with a fast rise time (within one gate) and a word with a slower rise time. I took the number of gates from the gate before rather than after the point at which the 10% criterion was crossed because many words have the entire change happening between contiguous gates, as in the first example in Figure 4.26, which would lead to a rise time of zero by the more usual method of using the first point after the criterion.

#### 4.5.3. Rise time results

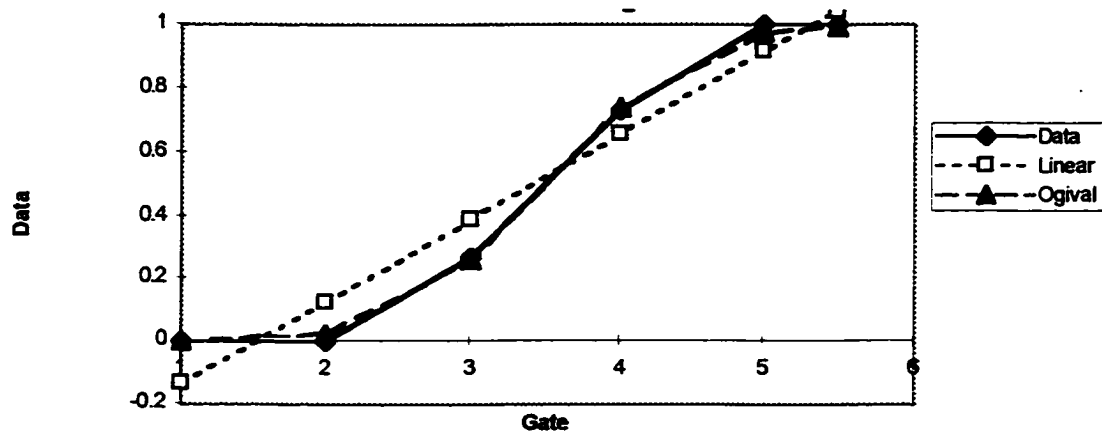
This measure of rise time was calculated for all of the %Corr data. It could also be calculated for the fall of the #Resp data, but I did not feel that this would offer much additional information relative to the %Corr rise times. The results are shown in Table 4.13.

---

<sup>24</sup> One could, instead, find the exact point where the fitted curve crosses 10% and 90% (in fractions of gates instead of in whole gates). This is possible since the fitted curve, unlike the data itself, was not measured at any point, and its value at any point (even between data points) can be calculated from its equation. This method might help to make the rise time measure somewhat more precise, and might be a good approach to adopt in future research. However, it would mean little to say that the improvement in perception happened between the points 1.23 and 4.38 on the gate number axis, for example.



"Committee" / $\tau$ /. Almost all of the increase in percent correct occurs between the third and fourth gates. The fourth gate is the first to exceed 10% of the increase, as the value at the third gate is greater than the value at the second by slightly less than 10% of the increase. Therefore, the third gate is the last before exceeding the 10% initial criterion. The fourth gate is also the first gate to exceed 90% of the increase, and so is the first point after the 90% final criterion. The rise time as calculated here is thus  $4-3=1$ , one gate.



#Resp data for "session" / $\epsilon$ /. The third gate is the first to exceed 10% of the increase, so the second is the last gate before the 10% criterion. The fifth gate is the first to exceed 90% of the increase. The rise time is thus  $5-2=3$ , three gates.

Figure 4.26. Calculation of rise times for two examples (%Corr data).



Table 4.13. Rise times of the %Corr measure, in number of gates. (Result is not an integer if the final and penultimate gates are separated by less than 20 ms and are part of the rise.) Words for which the partially correct measure was used are marked with an asterisk.

No.	Word	Trans.	Rise time	No.	Word	Trans.	Rise time
English				Japanese			
1	tip	ɪ	2.0	1	todana	[to]	1.0
2	stiff	ɪ	1.0	2	tatoe'ru	[to]	2.0
3	Tibet	ɪ	1.0*	3	ka'to	[to]	2.0
4	petition	ɪ	2.7	4	kakari'iN	[ka]	2.0
5	attic	ɪ	2.0	5	hakama'	[ka]	2.0
6	custom	kʌ	3.5	6	sya'kai	[ka]	2.0
7	skull	kʌ	5.0	7	dama'ru	[da]	1.0
8	accompany	kʌ	3.9	8	midare'ru	[da]	1.0
9	caboose	kə	1.0	9	ku'da	[da]	2.0
10	academic	kə	2.9	10	hotoke'	[ot]	2.0
11	duck	dʌ	2.0	11	himoto'	[ot]	2.0
12	citizen	ɪt	1.3	12	hakobu	[ak]	4.0
13	fitness	ɪt	2.0	13	hatake	[ak]	2.0
14	Italian	ɪt	3.0	14	ha'yaku	[ak]	2.0
15	committee	ɪt	1.0	15	kadai	[ad]	1.0
16	unity	ɪt	0.9	16	hanada'yori	[ad]	1.0*
17	bucket	ʌk	1.0	17	ka'nada	[ad]	4.0
18	mechanical	ək	2.0	18	megumi	[me]	1.7
19	indicate	ək	linear	19	tomeru	[me]	2.0
20	induction	ʌk	wrong slope	20	nemui	[ne]	linear
21	muddy	ʌd	1.7*	21	kemuri	[em]	1.8
22	cadenza	əd	2.0	22	tabemo'no	[em]	2.0
23	medicine	mɛ	1.8	23	teni'motu	[en]	0.8*
24	immense	mɛ	1.0*	24	soda'tu	[so]	2.0
25	remedy	ɛm	1.0	25	zabu'toN	[za]	2.0
26	attempt	ɛm	3.0	26	syabe'ru	[ʃa]	2.0
27	negative	nɛ	1.0	27	hokeN	[ho]	2.0
28	tenants	ɛn	1.0	28	zyosei	[os]	3.0
29	saddle	sæ	2.0	29	kazari	[az]	1.9
30	master	æs	2.0	30	basyo	[aʃ]	3.0
31	Zachary	zæ	3.0	31	gohoo	[oh]	2.0
32	asthma	æz	2.0	32	wahuku	[aɸ]	3.0
33	shell	ʃɛ	wrong slope*	33	dohyoo	[oɕ]	3.9
34	session	ɛʃ	3.0	34	harada'tu	[ra]	2.0
35	fees	fi	4.0	35	yubi'	[ju]	2.0

No.	Word	Trans.	Rise time	No.	Word	Trans.	Rise time
36	unfeeling	fi	2.0	36	kara'i	[ar]	1.7
37	leaf	if	4.0	37	huyoo	[uj]	1.6
38	relief	if	1.0	38	mawari	[aw]	1.6
39	vacuum	væ	2.0	39	tyazuke	[tʃa]	1.0
40	ravish	æv	3.4*	40	zyokyo'ozyu	[dʒo]	2.0
41	trail	re <sup>i</sup>	1.0	41	mati'	[atʃ]	4.0
42	fair	e <sup>r</sup>	3.0	42	tozi'ru	[odʒ]	3.0
43	lever	le	2.0	43	haNtai	[nt]	2.0
44	elevator	el	1.0	44	kaNdo	[nd]	wrong slope
45	yellow	je	1.0*	45	teNkiN	[ŋk]	5.0
46	watch	wa	2.0	46	kaNzeN	[nz]	2.0
47	chapel	tʃæ	wrong slope	47	seNsoo	[ns]	3.0
48	latches	ætʃ	1.8	48	keNritu	[nr]	2.0*
49	jump	dʒʌ	3.0	49	koNyaku	[nj]	4.0
50	judge	ʌdʒ	2.0	50	kiNtyoo	[ntʃ]	2.0
51	bent	nt	4.0	51	sukunai	[sk]	4.0
52	sentiment	nt	1.0	52	sikaku	[s <sup>h</sup> k]	5.0
53	reinterpret	nt	2.0	53	kitamuki	[k <sup>h</sup> t]	1.4
54	band	nd	1.0	54	kokutetu	[kt]	linear
55	wander	nd	linear	55	kyaku	[kj]	2.0
56	recondition -ed	nd	0.8	56	dakyoo	[kj]	3.0
57	axe	ks	2.0	57	hyoo	[çj]	3.0
58	hacksaw	ks	3.0	58	ryokaN	[rj]	1.0
59	unaccep- table	ks	1.0	59	mottaina'i	[t]	2.0
60	cats	ts	3.0	60	sakka	[kk]	3.0
61	Betsy	ts	2.0	61	sassoku	[ss]	2.0*
62	stop	st	3.0	62	hassya	[ʃʃ]	0.9
63	based	st	1.0	63	teNmetu	[mm]	2.0
64	pastime	st	2.0	64	aNnaizyo	[nn]	2.0*
65	skate	sk	2.0	65	tootyaku	[oo]	3.0
66	mask	sk	2.0	66	keigo	[ee]	1.0
67	discount	sk	2.0	67	syuukaN	[uu]	2.0
68	train	tr	linear	68	haori	[ao]	2.0
69	string	tr	2.0	69	siatu	[ia]	1.0
70	Detroit	tr	2.0	70	kaeri'miti	[ae]	2.0
71	crops	kr	1.0	71	taiko	[ai]	2.0
72	scrap	kr	1.0	72	koibito	[oi]	2.0

No.	Word	Trans.	Rise time	No.	Word	Trans.	Rise time
73	acrobat	kr	3.0	73	teNiN	[ēi]	2.0
74	drop	dr	1.6	74	hiN	[iŋ]	2.0
75	groan	gr	2.0	75	maNne'N- hitu	[an]	1.8
76	plain	pl	3.0	76	seNmoN	[em]	3.0
77	split	pl	1.0				
78	twelve	tw	1.0				
79	court	rt	2.0				
80	cork	rk	2.0				
81	help	lp	2.0				
82	fans	nz	2.0				
83	dance	ns	1.0				
84	fancy	ns	1.0				
85	uncon- cealed	ns	2.0				
86	snow	sn	3.0				
87	Disney	zn	2.0*				
88	farm	rm	1.0				
89	corn	m	wrong slope				
90	film	lm	wrong slope				
91	ranch	ntʃ	6.0				
92	flash	fl	2.0				
93	fragile	fr	1.0				
94	sleep	sl	2.0				
95	Iceland	sl	linear				
96	swan	sw	3.0				
97	golf	lf	1.0				
98	wharf	rf	1.0				
99	false	ls	2.0				
100	calcium	ls	3.0				
101	cultural	ʃtʃ	3.0				
102	marginal	rdʒ	2.0				
103	optical	pt	2.0				
104	pact	kt	2.0				
105	coughs	fs	3.0				
106	nerves	vz	2.0				
107	amnesty	mn	3.0				
108	garlic	rl	2.0				
109	biopsy	aʔa	1.0				

No.	Word	Trans.	Rise time
110	biography	a <sup>j</sup> a	2.0
111	biotech	a <sup>j</sup> o <sup>w</sup>	4.0
112	eon	ia	5.0
113	diagonal	a <sup>j</sup> æ	2.0
114	react	iæ	1.0
115	tiger	ta <sup>j</sup>	linear
116	bite	a <sup>j</sup> t	5.0
117	data	de <sup>j</sup>	2.0
118	fade	e <sup>j</sup> d	10.0
119	doubt	a <sup>w</sup> t	4.0
120	soybean	o <sup>j</sup> b	3.0
121	toad	to <sup>w</sup>	2.0
122	oats	o <sup>w</sup> t	5.0
123	courage	kə <sup>w</sup>	3.0
124	circle	ɔ <sup>j</sup> k	4.0
125	button	t <sup>j</sup>	2.0
126	beetle	t <sup>j</sup> l	2.0
127	apple	p <sup>j</sup> l	2.0

#### 4.5.4. Statistical analysis

I tested for the effects discussed above using the analysis of weighted means ANOVA, which allows for different numbers of subjects. The "subjects" here are the individual words, and if a particular word was more linear than ogival, for example, it cannot be used for the rise time calculation. Furthermore, not all CC transitions could be placed in environments where they crossed foot boundaries. This led to the unequal sample sizes. The sample sizes were nearly equal, though, since I tested only words which had been matched for factors other than the one being tested. For example, I did not test the rise time of every word in the experiment with the vowel of the transition of interest stressed against the relatively few words with the vowel unstressed, but tested only the words which had only the stress varied, such as "tip, Tibet" for /tɪ/ and "custom, caboose" for /kʌ/ or /kə/.

The rise time measure did not show consistent or statistically significant effects for any of the potential effects discussed above. For the manipulation of CC transition within the syllable, split across a syllable boundary, or split across a foot boundary, there was a slight tendency for rise time to be longer for CC transitions within the syllable, shorter for those within the foot but split across the syllable boundary, and shortest for those split across the foot boundary. This effect was not significant, though ( $F(2,10) < 1$ ). The difference in rise time between English words with the transition of interest in the first syllable and in the second syllable was negligible ( $F(1,20) < 1$ ). This was true whether all transitions were analyzed together, or whether CV and VC transitions were analyzed separately.

The results for stress were more interesting: for English words with CV transitions of interest, there was a nearly significant tendency for rise time to be greater for stressed vowel transitions than unstressed ones ( $F(1,6) = 4.36$ ,  $p < .09$ ). For VC transitions, however, there was a non-significant tendency for rise time to be greater if the vowel was unstressed than if it was stressed ( $F(1,8) = 3.18$ ,  $p < .12$ ). (CV and VC transitions were analyzed separately, as the effect of stress is likely to be different depending on whether the segment to be perceived is a vowel or a consonant.) Thus, vowels were perceived slightly more slowly (not necessarily later relative to  $D_{\max}$ , but just over a longer time window) when stressed, but postvocalic consonants were perceived slightly more slowly if the preceding vowel was unstressed. The sample sizes for these comparisons are small, and neither of these effects reaches statistical significance. A subsequent experiment could be designed to test this particular question with larger sample sizes.

For Japanese, the syllable position and pitch accent manipulations are inter-related, because the mora with falling pitch (the one after the accented mora) cannot be the first mora of the word<sup>25</sup>, as discussed in Chapter 2. Therefore, I carried out a three way

---

<sup>25</sup> It could be in the first syllable of the word, but would then have to be the mora nasal, for which pitch accent placement was not manipulated.

analysis, comparing transitions in the first syllable (necessarily before the accent), those in the second syllable but before the accent fall in pitch, and those in the second syllable and containing the fall in pitch (that is, second mora where the first mora is accented). I compared these three categories separately for CV transitions and VC transitions (with syllable position defined based on the vowel, since the consonant is in the following syllable). For the CV transitions, there was a slight tendency for the rise time to be shortest for transitions in the first syllable, longer for those in the second syllable but before the accent fall, and longest for those in the mora with the accent fall. However, this was not significant ( $F(2,8)=1.54, p>.05$ ). The results for the VC cases were clearly not significantly different across the three categories ( $F(2,7)<1$ ). If one groups all of the transitions which are before the accent pitch fall together and tests them against all the transitions during the pitch fall, regardless of syllable position, there is a tendency for the rise time to be greater for the transitions during the post-accent fall, but it is not significant ( $F(1,19)=2.40, p>.05$ ).

Thus, there are no significant differences in rise time for first versus second syllable position, stress, or pitch accent. Some differences for these manipulations are observable in the comparison of perceptual measures with the location of  $D_{\max}$ . For example, the tendency for postvocalic stops to be recognized early relative to  $D_{\max}$  is primarily limited to postvocalic stops after stressed vowels in English, as was discussed in Section 4.3 above. For the CC transitions, however, there appears to be no difference between those within the syllable, crossing a syllable boundary, and crossing a foot boundary with regard to when the segment is recognized relative to  $D_{\max}$ , either. Although an experiment designed to test explicitly this difference, with larger sample sizes for this particular manipulation, might find some effect, there appears thus far to be strikingly little effect of these suprasegmental boundaries on timing of perception of consonants in English. Overall, the rise time measure shows little sign of any effects of syllable position, stress, or pitch accent

in either language. However, the rise time analysis will be useful for another topic, the perception of stops in onset and coda position, discussed below.

#### 4.6. Perception of word initial stops

The usual method of evaluating the data from this experiment for segment perception, as described in the previous sections of this chapter, was to evaluate the percent of subjects to correctly perceive the second segment of the transition of interest. However, I also evaluated the percent of subjects to give responses with the first segment of the transition of interest correct for a subset of the data, namely the words with word initial stops as the first segment of the transition of interest. This was done in order to compare the timing of perception of word initial stops with that of postvocalic stops in coda position and postvocalic stops which are the onset of the following syllable. The words in the experiment which had a word initial stop as the first segment of the transition of interest were "tip, Tibet, custom, caboose, duck, train, drop, crops, groan, plain, twelve, tiger, data, toad, courage" in the English experiment and /todana/ 'closet,' /kakariiiN/ 'manager,' and /damaru/ 'to be quiet' in the Japanese experiment.

I evaluated the number of responses with the first segment correct for each word, and fit an ogival curve to the data in the same way as was done for the other data. The perception of /d/ in /damaru/, "drop," and "data" was linear, so these words were excluded from further consideration. The data, with fitted ogival curves, appears in Appendix C. I then calculated the rise time of the data for the remaining words. The results appear in Table 4.14 below.

Table 4.14. Rise time of the %Corr measure for word initial stops.  
Results are in number of gates, as in Table 4.13.

No.	Word	Stop	Rise time	No.	Word	Stop	Rise time
English				Japanese			
1	tip	t	1.0	1	todana	[t]	1.0
3	Tibet	t	2.0	4	kakari'iN	[k]	2.0
6	custom	k	1.0	7	dama'ru	[d]	linear
9	caboose	k	1.0				
11	duck	d	2.0				
68	train	t	2.0				
71	crops	k	1.0				
74	drop	d	linear				
75	groan	g	2.0				
76	plain	p	2.0				
78	twelve	t	2.0				
115	tiger	t	2.0				
117	data	d	linear				
121	toad	t	1.0				
123	courage	k	2.0				

For the English words, I compared the rise time for perception of stops in word initial position, stops in coda position, and stops in postvocalic onset position (using the words with VC transitions of interest in which the C was an onset stop). For Japanese, I tested the rise time of the word initial stops against the postvocalic onset stops. Japanese has no stops in coda position except when they are the beginning of a geminate, and the list did not include any VCC transitions (as discussed in Section 2.2.1), so these could not be tested. The results of these calculations appear in Table 4.15.

Table 4.15. Average rise time of %Corr data for stops in word-initial, postvocalic onset, and coda positions, in gates.

Stop position	English	Japanese
Word initial (onset)	1.62	1.5
Postvocalic onset	2.43	2.0
Coda	5.20	

For the English results, the difference in rise time among the three categories was statistically significant (using ANOVA of weighted means for unequal sample sizes,  $F(2,$



21)=11.75,  $p < .0005$ ). Grouping the postvocalic onsets and the coda stops together, since these are both perceived based on the cues in the VC transition in this experiment, stops perceived based on VC cues have a significantly longer rise time than the word initial stops, which are perceived based on CV cues ( $F(1,22)=8.42$ ,  $p < .01$ ). If one groups all of the stops in onset position together and tests their rise time against the coda stops, that difference is also significant. However, different cues are being used to perceive the word initial onset stops and the medial onset stops, since the words containing medial onset stops were gated out at or before the release of the stop, leaving only the VC cues, and for the final gate only, the burst. Therefore, it is unclear what conclusion one could draw from grouping the word initial onset stops and the medial onset stops together. The difference between the two stop environments for Japanese was not statistically significant ( $F(1,7) < 1$ ), but since there were only two word-initial stops for Japanese which could be included in the test, there is not enough data to draw any conclusions on this subject.

It may seem surprising that the rise time for the English coda stops is so much longer than that for postvocalic onset stops, since both sets of stops are being perceived from VC cues. This is probably an artifact of the word list. The words with coda stops are predominantly in monosyllabic words following a diphthong or long vowel, as in "fade, bite, oats." These vowels are very long, inherently because they are diphthongs, also because of word final (or utterance final) lengthening, and because vowels are longer in monosyllabic words than in polysyllabic words. Only one of the coda stops, the one with the shortest rise time, follows the vowel /i/, in "fitness." The postvocalic onset stops are predominantly after shorter vowels (/i, u/) and of course are in polysyllabic words, and not in the final syllable of the word. I believe the difference in rise time between the coda stops and the postvocalic onset stops is primarily a reflection of the fact that the vowels preceding the coda stops are longer, and hence cues to the stop can spread over a longer time window. A more thorough test of timing of perception of stops should control for vowel

quality and final lengthening. The word initial stops are unlikely to be affected by these issues, however, so the difference between them and the postvocalic stops (either onset or coda) remains valid.

There are several more analyses along these lines which might yield interesting results. For example, it might be useful to compare word initial consonants other than stops to the corresponding consonants in postvocalic position for rise time. Furthermore, one could compare the word initial onset stops to post-consonantal onset stops. However, for postconsonantal stops, the exact environment of the stop would be expected to have a strong influence on the timing of its perception. Some consonants allow cues for a following stop to spread into them, so that the stop is equivalent to a postvocalic stop, as in the case of the /t/ in "court." Other environments may allow very little spread of cues into preceding segments, as for the /t/ in "optical." For some environments, such as /t/ or /k/ after /s/ in "stop, skate," there are some cues for the stop during the fricative (McQueen 1997), but they are probably not as strong as for the /t/ in "court." A valid comparison of the rise time of word initial and post-consonantal stops would require careful investigation of the extent to which cues for a stop can spread into each possible preceding segment, in order to know which groups of words to treat as similar. Since it is clear that word-initial stops in this experiment must be perceived through cues in the CV sequence (not in the part of the signal leading into the C, since that is silence), I will for now examine only word-initial stops for purposes of comparison with post-vocalic onset and coda stops.

## 5. Conclusions

### 5.1. Dynamic perception of segments

#### 5.1.1. Distribution of information in the signal

In Section 4.1.5, I showed that between 89.5% and 97.3% of the words for each type of data and each language are better fit by an ogival curve than by a straight line. This shows that listeners make more progress toward perceiving segments and recognizing words during some parts of the signal than others. The high percentage of words which are more ogival than linear is strong evidence that information is not distributed equally throughout the signal in time, but is concentrated into some high information flow areas, while other areas carry relatively little additional information.

#### 5.1.2. Slow and fast changes and the sensitivity of the auditory system

One point which emerges from the use of the measure D to study degree of spectral change is that there are two very different types of change in the signal, namely very sudden changes and slower changes, a distinction which is also discussed by Stevens (1971). The change from silence to a voiceless stop burst is nearly instantaneous, the change from a vowel to a nasal or nasal to vowel can involve a very sudden spectral discontinuity (as in Figures 3.19 and 3.20 above), and the cessation of formants at the closure of a voiceless stop can also be quite sudden. However, many other inherently changing aspects of a speech signal, such as the movement of formants into or out of consonants, between two adjacent vowels, or during glides take place over a considerably longer time (approximately 50 ms or more). The spread of nasalization or /r/ coloring into a preceding vowel can also be considered a slow change, as the vowel gradually becomes more nasalized or the third formant gradually lowers for /r/.

These two types of changes are distinguished not only by their speed: the slower changes are change within some continuing sound, whereas the fast, near instantaneous changes are either the onset of a sound or the boundary between two distinct sounds which

are contiguous. For example, the change in formant frequencies in the transition into a stop is a change within the continuing signal of the vowel. The burst of a stop, however, is a change from silence to a sound, the burst. The change from a nasal consonant to a vowel, when there is a clear spectral discontinuity, is the change from one type of sound (the nasal) to another (the vowel), not a gradual transition within one sound. What makes this distinction important is that linguists have traditionally thought of "transitions" as including only the slower type of change, a change within a continuing sound. Thus, the previous literature on dynamic versus static cues, except Furui's (1986) work and some work by Stevens (1971, 1985), has defined the dynamic cues as those which take place within some continuing part of a sound, not as the boundaries between sounds. Strange et al. (1983) assumes that the formant transitions during a vowel constitute the dynamic cues. Kewley-Port (1983) and Kewley-Port et al. (1983) point out that the quality of burst and aspiration noise (if any) of a stop is also not static, but varies over time throughout the period from the onset of the burst to the onset of voicing. However, none of these works addresses the change from silence to the burst as a potential dynamic cue. Nor do these authors consider the onset of voicing itself as a potential dynamic cue, even though both the onset of the burst and the onset of voicing represent large changes in the signal.

In contrast to the other studies, Furui's (1986) approach is likely to emphasize instantaneous changes at boundaries of sounds over the slower changes within sounds because the measure  $D$  is most sensitive to sudden changes, as discussed in section 3.2.1. Stevens does emphasize the importance of rapid changes in the signal in several works (Stevens 1971, 1980, 1985, and to a lesser extent Stevens and Blumstein 1978, 1981, Stevens and Keyser 1989 and Stevens et al. 1986). However, some of the perceptual cues Stevens proposes, such as spectrum at consonant release, are inherently static. Others, such as the ratio of frication noise amplitude at a certain frequency to the amplitude of a vowel formant, could vary over time in natural speech, but he defines them as static and uses synthetic stimuli in which these amplitudes do not vary to test his hypothesis. He

proposes that the listener locates regions of rapid spectral change in the signal, and then extracts static cues from those regions. Furthermore, Stevens' (1971) emphasis on rapid spectral changes and exclusion of rapid amplitude changes leaves out several possible types of rapid dynamic cues. For example, he specifically excludes the change from aspiration noise to voicing as a rapidly varying cue, because it does not involve much change in where spectral peaks are located (1971). Stevens (1985) included rapid changes in amplitude as well, although he does not discuss the change from aspiration to voicing in this work.

The primary reason for proposing that dynamic cues of any sort should be important for speech perception is that the auditory system is more sensitive to changes in sounds than to continuing, steady sounds<sup>1</sup>. It is therefore important to consider whether this characteristic of the auditory system is true of near instantaneous changes between types of sounds, of slower changes within a continuing sound, or both. As discussed in Section 1.2.2, auditory nerve fibers react most strongly at the onset of a sound, and their response decreases quickly as a steady state signal continues. This phenomenon is called adaptation. Not all types of auditory nerve fibers adapt in the same way, but the group of auditory nerve fibers referred to as "high spontaneous discharge rate" comprise approximately 60% of auditory nerve fibers, and these adapt to a signal quickly, showing a strong decrease in their response over the first five to ten milliseconds after the onset of the signal (Greenberg 1997:1305-7)<sup>2</sup>.

---

<sup>1</sup> The hypothesis that speech sounds are perceived primarily through dynamic parts of the signal is meant to apply both to consonants, many of which do not have a long steady portion, and vowels, which often do. The hypothesis is that listeners make rapid progress in perceiving a vowel, for example, at its onset transition, and that hearing the following steady state results in little additional improvement. This was discussed in Section 1.2.3.2 with regard to the work of Strange et al. (1983) on vowel quality perception, and in Section 4.3.1.2 with regard to the results of this experiment.

<sup>2</sup> It is not a problem that the gating interval of the experiment is longer than the interval over which adaptation at the auditory nerve fiber level takes place. The purpose of the 20 ms gating interval is to make sure no more than one linguistically relevant change in the acoustic signal happens during a gate and to make sure the gating interval is shorter than any proposed phonological unit. The experiment is not designed to test how listeners' responses might vary as their auditory nerve fibers adapt, and it is not clear what prediction one might make on this subject.

Adaptation is one of the major reasons for thinking the auditory system is more sensitive to changes than steady signals, but this rapid adaptation clearly applies more to instantaneous changes at the onset of sounds, such as the onset of bursts, and not necessarily to changes within a sound. Descriptions of adaptation in the literature on the peripheral auditory system (in particular on auditory nerve fiber response) usually show adaptation to a simple, completely static signal such as a sine wave, from its onset. In such examples, the auditory nerve fibers fire at a high rate at the onset of the signal, but the firing rate drops off rapidly during the first 5-10 ms after the onset of the signal (Greenberg 1997:1306). This sort of rapid adaptation at the onset of a signal applies not only to the onset of a signal from silence, but also to sudden changes between two sounds. This is because auditory nerve fibers have characteristic frequencies, the frequency of sound to which each fiber responds most strongly. In the simple sine wave stimulus examples, only auditory nerve fibers with characteristic frequencies near the frequency of the sine wave respond. If two contiguous speech sounds have largely different component frequencies, as a burst or aspiration noise (primarily high frequency noise) and a vowel (strong energy below 2000 Hz) do, the high characteristic frequency auditory nerve fibers will respond to the burst and adapt to it, but at the onset of the vowel, the low characteristic frequency auditory nerve fibers will respond, and then adapt.

Because the low characteristic frequency auditory nerve fibers are not responding during a high frequency sound such as aspiration noise, in a stop-vowel sequence, the low characteristic frequency fibers have not adapted yet when the vowel begins. Thus, both the onset of the burst and the onset of voicing will show a strong initial response followed by adaptation, in different auditory nerve fibers. Therefore, one can assume that the auditory system is especially sensitive to the onset of a sound even if it does not follow silence as long as the onset of the sound is sudden and it has different component frequencies from the sound it follows. This would imply that all of the changes in speech signals described as sudden changes above (onset of bursts or voicing, sudden spectral discontinuity

between nasals and vowels, etc.) would cause adaptation at least to some extent, and therefore that the auditory system would be especially sensitive to their onsets, which usually have maxima of the measure D.

There are some reasons to think that the auditory system is also more sensitive to slower changes in the signal than to steady states, although the evidence is less clear than for the sudden changes. At the level of the auditory periphery, particularly the auditory nerve fibers, rapid adaptation would seem only to apply to the onsets of sounds, not to slow changes within a sound such as the movement of formants. However, because auditory nerve fibers respond most strongly at their characteristic frequencies, if the frequency of a formant changes enough, auditory nerve fibers of different characteristic frequencies will respond to it as its frequency changes. Delgutte 1997:511 shows that the rapid movement of the second and perhaps third formants in the /gri/ sequence of a production of "green" causes a response from successively higher characteristic frequency auditory nerve fibers over time. Thus, the auditory system might be sensitive to slower changes in formant frequencies because these changes cause a reaction among many different auditory nerve fibers as the formant frequencies change<sup>3</sup>. Greenberg (1996a) points out that when there is enough change in the frequency of a sound, as in relatively rapid changes in formants, none of the auditory nerve fibers will be responding to the signal for long enough for their response levels to decrease through adaptation. He concludes that "signals with substantial frequency modulation, such as consonantal and vocalic transitions, are likely to produce a net increase in auditory activity, both in terms of the number of neurons activated and the magnitude of excitation" (1996a:395).

There is also reason to think that the auditory system would be sensitive to slower changes at higher levels of auditory processing, namely at the level of the auditory cortex.

---

<sup>3</sup> However, auditory nerve fibers (at least the high spontaneous response ones) are less closely tuned to their characteristic frequencies at high sound pressure levels, that is for loud sounds (Greenberg 1996:378). For loud speech the movement of formants might not clearly cause a change in which auditory nerve fibers respond over time.

Greenberg points out that cells in the auditory cortex discharge much less frequently than cells in the peripheral auditory nerve fibers, usually only five to twenty times per second (1996b:4), and often only at the onsets of sounds (1996a:394). He argues in numerous papers that while the auditory periphery "performs a fine-grained analysis on a pitch-cycle-by-cycle basis" (1996a:395), the cortex performs analyses of the signal at higher levels, at intervals of approximately 40 ms and 200 ms. He points out that 200 ms is approximately the duration of an average syllable, while characteristics of the signal within 40 ms windows can provide information which distinguishes syllables from each other (cues to phonemes and distinctive features). He proposes that "the role of the cortical regions of the auditory system may be to process auditory 'events,' whose features change much more slowly than the acoustic spectrum," and relates this to the low discharge rate of the auditory cortex cells (Greenberg 1997:1319). Although research on auditory processing at the level of the auditory cortex is much less advanced than research on the peripheral level of the auditory system, these findings suggest that the auditory system may be sensitive to slower changes in the signal, such as formant transitions, at the level of the cortex. The interval of 40 ms Greenberg proposes is similar to the duration of many of the slower changes in speech, such as formant transitions.

Jamieson (1987) argues on the basis of psychoacoustic data that when a transition (particularly of formant frequencies in a CV sequence) is followed by a steady state, the auditory system is most sensitive to transitions which take about 40-60 ms, and is less sensitive to either faster or slower transitions. He shows that for sine wave stimuli with a frequency transition followed by a steady state, and with the duration of the transition varied, listeners can discriminate differences in stimuli best if the transitions last 40-60 ms. He explains the psychoacoustic basis of this preference for medium duration transitions as masking of the transition by the following steady state. He concludes that languages tend to use transitions of approximately 40-60 ms for CV transitions because when a steady



state follows, there is a psychoacoustic predisposition to perceive transitions of that duration best.

### 5.1.3. Implications of the experiment for fast and slow changes

Because the measure of degree of spectral change  $D$  is primarily sensitive to sudden changes in the speech signal, and fails to reflect slower changes like formant transitions, comparison of the area of maximal perceptual change to the location of  $D_{\max}$  is really a test of the importance of rapid changes for speech perception. Until a measure of degree of spectral change which reflects both sudden and slower changes is developed, this experimental design will test primarily the importance of dynamic cues at rapidly changing parts of the signal, especially at the boundaries between sounds. The major result of the experiment is that the number of transitions with the area of most improvement in perception ( $\%Corr^4$  measure) surrounding a  $D_{\max}$  point is significantly greater than chance. This result supports the hypothesis that dynamic cues stemming from sudden changes in the signal are more important than steady portions of the signal for speech perception. Considering that both sudden and slower dynamic cues are predicted to be more important than steady cues, and that segments which are perceived from slower dynamic cues are not likely to have maximal perceptual improvement surrounding  $D_{\max}$  because of the insensitivity of  $D$  to such cues, the result of approximately half of all words fulfilling the hypothesis is quite strong evidence for the importance of rapid dynamic cues.

Because of the insensitivity of  $D$  to slower changes in the signal, this experiment does not provide a direct test of the importance of dynamic cues stemming from slower changes, such as formant transitions, the spread of nasalization, etc. However, the experiment does provide indirect evidence for the importance of slow dynamic cues through the reasons for the categories of transitions with exceptional behavior (Section

---

<sup>4</sup> As in the previous chapter, I will abbreviate "percent correct" with "%Corr" and "number of responses" with "#Resp" when referring to the two types of perceptual data in my experiment.

4.3.1). Postvocalic stops are recognized early relative to  $D_{\max}$  because the  $D_{\max}$  point comes at the offset of formants or voicing (at the beginning of the closure for the stop), but listeners often make the most progress toward recognizing the postvocalic stop during the vowel, before reaching the stop closure. The most likely perceptual cues for the stop which occur during the vowel are formant transitions (as a cue for place of the stop) and vowel duration (as a cue for voicing of the stop). Vowel duration is not a dynamic cue, but the formant transitions are inherently changing. Thus, this is a case in which listeners are using at least some slower dynamic cues, but the hypothesis of overlap with  $D_{\max}$  is not fulfilled because  $D$  does not reflect slow changes.

The situation for vowel-vowel transitions is even more clear: the movement of formants between the two vowels is an inherently changing cue, although a slow one, but is not reflected in the measure  $D$ . Because there are no sudden changes in a VV transition, the location of  $D_{\max}$  is somewhat random, even though listeners are likely to be using the changing aspects of the signal to perceive the vowels. For the category of postvocalic sonorants, again, the cues in the part of the signal where listeners make the most progress in perception (formant movement for transitions into /r, l/ and the change to nasalization for nasals) are too slow for  $D$  to reflect, but are inherently dynamic. In the case of post-consonantal vowels, which are recognized late relative to  $D_{\max}$ , there are at least two sources of dynamic cues which are not reflected in the measure  $D$ , namely the formant transitions out of the consonant and the changes in the vowel's formants which are inherent to the vowel and would appear even if it were produced in isolation. Even English vowels which are considered to be relatively monophthongal, such as /i, æ, u/, do have some change in their formants over time. Nearey and Assmann (1986) show that the slight offglides of these vowels are important for correct perception of vowel quality. For more obviously diphthongal vowels such as /e<sup>j</sup>, o<sup>w</sup>/ as well as /a<sup>j</sup>, o<sup>j</sup>, a<sup>w</sup>/, of course, change in formants is likely to be an even more important cue. Duration is also an important cue for

many vowels, aside from the phonemic length distinction in Japanese, but it is clear that there are many inherently dynamic cues available for post-consonantal vowels which are not reflected by maxima of  $D$ .

Some segments are recognized primarily through a cue which is not dynamic, however, namely duration. In this experiment, fricatives were recognized well after the  $D_{\max}$  point, which occurs at the onset of frication. There is little sign of inherently dynamic but slow cues in the fricatives: they are often extremely steady state over a rather long duration, especially in the case of the louder fricatives, such as /s, ʃ/. Previous work on perception of fricatives has shown that duration of noise is a crucial cue (Jongman 1989, Kluender and Walsh 1992), although others suggest rise time of noise amplitude as an important cue, which is also inherently a change (in amplitude). The distinctively long segments of Japanese, of course, are also perceived as long primarily based on duration. Thus, both the fricatives and distinctively long segments have duration as their primary cue, and the fact that these often fail to have the area of maximal perceptual change surrounding a  $D_{\max}$  point is not a failing of the measure  $D$ . These are simply exceptions to the hypothesis that dynamic cues are more important than static ones<sup>5</sup>. Only one category of exceptions remains, namely the medial glides, which are recognized late relative to  $D_{\max}$ . In Section 5.6.2 below, I show that the quality of these segments (front or back) is recognized relatively early, and it is only the recognition of the glide vs. vowel distinction which is late. It may be that the quality of the glides is recognized based on dynamic cues near the  $D_{\max}$  point, but that listeners are unable to recognize that the glides are consonants until the change into the next vowel, which is also a slower dynamic cue.

In sum, there are seven categories of exceptions, in each of which there are likely explanations for why the area of maximal perceptual change fails to surround  $D_{\max}$ . Four

---

<sup>5</sup> It is not clear to me, however, that duration is a static cue, either. I have already discussed the issue of fast versus slow dynamic cues, which could be thought of as dynamic cues with short versus long duration. Duration as a cue itself is perhaps an entirely separate dimension from degree of change.

of these are types of transitions with inherently changing aspects of the signal as likely perceptual cues, which however are too slow to be reflected by the measure  $D$ . Two others are types of transitions for which duration is known to be a critical cue. That so many of the exceptional categories of transitions involve gradual changes in the signal provides indirect support for the assertion that dynamic cues, even gradual ones, are more important in speech perception than static cues. Since other studies have found many of these gradual changes to be important perceptual cues (Nearey and Assmann (1986) for formant change during relatively steady state vowels, Kewley-Port et al. (1983) for formant transitions in the environment of consonants, Fujisaki and Sekimoto (1975) for vowel-vowel formant changes), the fact that the transitions with exactly these cues are the ones that form the classes of exceptions supports the importance of slow dynamic cues.

#### 5.1.4. The meaning of alignment with $D_{\max}$

One further issue must be addressed in concluding that the results support the hypothesis of the importance of dynamic cues.  $D_{\max}$  points often fall at what would traditionally be labeled as segment boundaries (closure of stops, onset of voicing, etc.). Therefore, the fact that listeners often make the most progress in recognizing segments right at the  $D_{\max}$  point could be interpreted as meaning not that listeners make the most use of the rapidly changing parts of the signal, but rather that they only need a short sample of segments in order to perceive them, and hence perceive the segment shortly after its onset. Because a large proportion of  $D_{\max}$  points do fall at the onset of the target segment (when it has a clear boundary), this is a serious objection.

To address this objection, I evaluated the %Corr results for all words which had more than one  $D_{\max}$  point and had the maximal change in %Corr surrounding a  $D_{\max}$  point<sup>6</sup>. If alignment of change in the perceptual measure with a  $D_{\max}$  point means only that listeners can recognize a segment as soon as they hear the beginning of it, the area of most change in

---

<sup>6</sup> As usual, I excluded words where the %Corr data was more linear than ogival or had an anomalous slope.

the perceptual measure would surround the  $D_{\max}$  point at the onset of the second segment of interest, not a  $D_{\max}$  point during that segment or near the end of it. This is illustrated schematically in Figure 5.1. I examined the acoustic changes which the  $D_{\max}$  points for these words are associated with, and in particular whether the first, second, or neither  $D_{\max}$  point is associated with what would traditionally be called the onset of the segment. I then considered whether the area of maximal change in %Corr surrounded a  $D_{\max}$  point at the onset of the segment, one before the onset, or one after it.

When the transition of interest is from a medial stop to a vowel, the first  $D_{\max}$  point is at the burst and the second at the onset of voicing, which would traditionally be considered the onset of the vowel. Figure 3.3 above demonstrated such a case. Therefore, the maximal change in the perceptual measure cannot surround a  $D_{\max}$  point after the onset of the segment, so these words do not allow for a test of this question. The same is true of stop-sonorant transitions. In a fricative-vowel or fricative-sonorant transition, if there are two  $D_{\max}$  points, the first may be at a change in the quality of the frication noise or reduction in amplitude of frication noise, and the second at the onset of voicing. Also, in the word /wahuku/ [waɸuku]<sup>7</sup> 'Japanese style clothing,' there are two  $D_{\max}$  points, the first at a point where the amplitude of formants decreases and the second at the onset of frication. In all of these cases, the second of the two  $D_{\max}$  points is at the onset of the segment of interest, so one cannot test the possibility that the area of maximal perceptual change might surround a  $D_{\max}$  point later than the onset of the segment<sup>8</sup>.

For the remaining words, the proportion of words which had the area of maximal perceptual change surrounding the first  $D_{\max}$  point and the proportion with it surrounding a later  $D_{\max}$  point was calculated. The results appear in Table 5.1.

<sup>7</sup> As in previous chapters, the two segment transition of interest in a word is denoted either by printing it in bold or by transcribing it after the word itself, if the word is written only orthographically.

<sup>8</sup> For words with three  $D_{\max}$  points instead of two, I examined whether the latest  $D_{\max}$  point was at the onset of the segment of interest. Only one English word and one Japanese word had three  $D_{\max}$  points and had the area of maximal perceptual change surrounding any of them, and neither of these words had the final  $D_{\max}$  point at the onset of the segment. No words had more than three  $D_{\max}$  points.



Figure 5.1. The response of the measure  $D$  for some hypothetical two segment sequence with two peaks of  $D$  within the gated area. If the onset of the target segment (the second segment of the two segment sequence of interest) is at peak A, then there is one more peak which is after the onset of the segment (B). If listeners identify words at  $D_{\max}$  points only because they can identify the segment once they have heard a short part of it, the maximal improvement in the percent correct measure should surround peak A, not peak B, in such cases. Therefore, if a large proportion of such cases have the maximal improvement in percent correct surrounding peak B instead of peak A, this shows that it is areas of great spectral change in the signal which are perceptually important, not just beginnings of segments. However, if the onset of the target segment is at peak B, the word cannot be used to test this issue, since there is no possibility for the area of maximal perceptual improvement to surround a  $D_{\max}$  point after the onset of the segment.

Table 5.1. Results for words with more than one  $D_{\max}$  point, with regard to which  $D_{\max}$  point the area of maximal perceptual change (in %Corr) surrounds. Results are in number of words in each category.

Type of words	English	Japanese
A All words in experiment	127	76
B Number of words with more than one $D_{\max}$ point	58	34
C Number of words in row B for which the area of max. change in %Corr surrounds any $D_{\max}$ point	33	19
D Number of words in row C for which there is a $D_{\max}$ point after the onset of the segment of interest <sup>9</sup>	26	12
E Number of words in row D for which the area of max. perceptual change surrounds a $D_{\max}$ which is after the onset of the segment	18	8

When considering words which have more than one  $D_{\max}$  point, of which at least one  $D_{\max}$  point falls after the onset of the segment and the area of maximal perceptual change does surround one of the  $D_{\max}$  points, it surrounds the  $D_{\max}$  point after the onset of the segment in 18 out of 26 possible words in English (69.2%) and in 8 out of 12 possible words in Japanese (66.7%). Figure 5.2 shows an example of a word in which the maximal perceptual improvement surrounds a  $D_{\max}$  point after the onset of the segment, "committee" [ɪr].

In many of these words, the transition of interest is a transition into a stop or affricate, such as a vowel-stop, fricative-stop, nasal-stop, stop-stop, or vowel-affricate transition. Listeners often make the most progress toward recognizing the stop or affricate at the burst, not the closure<sup>10</sup>. For such transitions, the first  $D_{\max}$  point is at the closure and the second at the burst. The second segment of interest in the words "groan" /gr/, "trail" /reɪ/, and "medicine" /mɛ/ is also recognized at the second  $D_{\max}$  point<sup>11</sup>, which is after the onset of the segment. Finally, all five words in the experiment which had a flap as the

<sup>9</sup> i.e. the last  $D_{\max}$  is not at the onset of the segment of interest

<sup>10</sup> This may seem to contradict the result discussed in Section 4.3.1.1, that postvocalic stops often have the most progress toward recognition of the stop during the preceding vowel. While many postvocalic stops do show that pattern, those that do not very often have the most progress toward recognition at the burst of the stop. Which of these patterns a particular word shows may depend on the cohort of the word.

<sup>11</sup> This is the second of three for "groan."

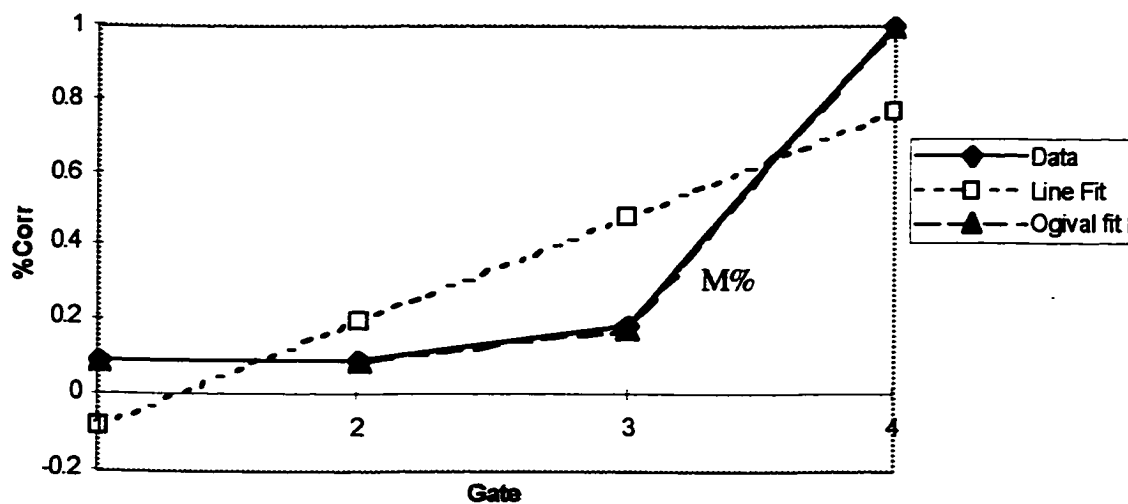
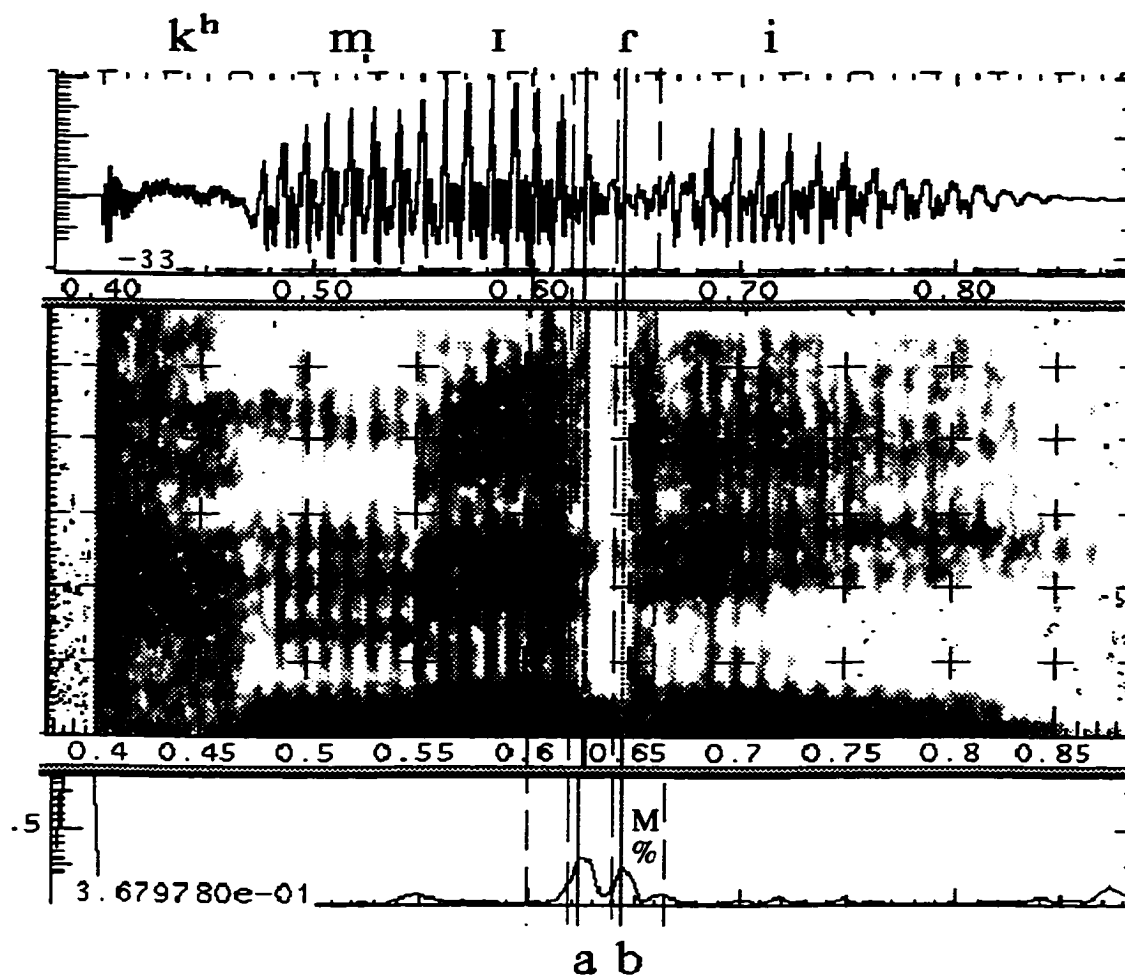


Figure 5.2. Waveform, spectrogram, D, and %Corr data for the word "committee" [ir].  $D_{\max}$  points are shown by solid vertical lines, and endpoints of gates are shown by dashed vertical lines. The first  $D_{\max}$  point ("a") is at the onset of the flap, but the area of greatest improvement in the %Corr data (M%) surrounds the second  $D_{\max}$  point ("b"), which is at the end of the flap.



second segment of the transition of interest, "citizen" /ɪt/, "unity" /ɪt/, "committee" /ɪt/, "muddy" /ʌd/, and /karai/ [karai] 'spicy,' had the area of maximal improvement in perception of the flap surrounding the latter of the two  $D_{\max}$  points<sup>12</sup>, as exemplified in Figure 5.2. In these transitions, the latter  $D_{\max}$  point comes at the end of the flap and the first at the beginning of the flap. This is a highly consistent result, and indicates that flaps are perceived through the dynamic cues at the release of the flap.

In sum, in approximately two thirds of words which have more than one  $D_{\max}$  point, one of which is after the onset of the segment of interest, and have the area of maximal perceptual change surrounding one of the  $D_{\max}$  points, it surrounds the one after the segment has already begun, not the one at the onset of the segment. This is true of both languages. This provides strong evidence that listeners are not simply perceiving segments at  $D_{\max}$  points because the  $D_{\max}$  points fall at the onsets of segments. The results from this subset of the data support my conclusion from the overall experiment that listeners make disproportionate use of changing parts of the signal, wherever in the segment the dynamic aspects of the signal may fall. The flaps, which are consistently perceived at the  $D_{\max}$  point at their releases, are an especially convincing case of this.

## 5.2. Dynamic perception and spoken word recognition

The results of the experiment are less clear with regard to the hypothesis that spoken word recognition takes place primarily through the use of dynamic cues than they are for the perception of individual segments through dynamic cues. For both languages, listeners make the most progress in narrowing in on the word at a  $D_{\max}$  point only slightly more often than would be predicted by chance, and the difference from chance is not

---

<sup>12</sup> "Muddy," however, is being evaluated on the basis of listeners getting two out of three of voicing (voiced), manner (flap), and place (alveolar) correct, since responses with flaps never exceeded 30%. Most of the improvement occurs as listeners shift to the response "mud," so this word is not evidence for how flaps are recognized.

significant. I believe that this is primarily due to problems with the number of responses (#Resp) data as a measure of progress toward spoken word recognition, and does not reflect real differences in the use of dynamic cues for the segment level and the word level of perception.

The excessive sensitivity of the #Resp measure to individual variation was discussed in Section 4.4.5 above. A response from just one subject, no matter how unlike other subjects' responses it is, counts to the same degree as a response given by a large number of subjects, and this makes the degree of error variation in the #Resp measure quite large. This is the reason for using the percent correct (%Corr) data for the majority of the analyses: if one subject at one gate gives an anomalous response, it will have little influence on the %Corr measure. The situation discussed with regard to the word "nerves" /vz/ in Section 4.4.5 presents a different problem: even a very large and sudden shift from ten of the subjects responding "nerve" to ten responding "nerves" will show no change in the #Resp at all if the remaining subject gives some other response. In a hypothetical extreme case, all of the subjects could respond with one word at a particular gate, then all of the subjects respond with a different word at a subsequent gate, and the #Resp measure would fail to reflect the change in which lexical item had been recognized.

This problem can also lead to the #Resp data being classified as having an anomalous slope, if there is no overall change in the number of responses (zero slope) or a slight increase in the number of responses. This is the case for the word "fans" /nz/: nearly all subjects respond with "fan" until the onset of frication, after which nearly all respond with "fans," but there are a few more responses different from the majority after the onset of frication than before it, so the overall trend of the data is positive, and the word is excluded as having an anomalous slope. This clearly demonstrates that there can be a large change in which word is recognized, even happening exactly at the  $D_{\max}$  point, without the #Resp measure reflecting this result.

The overall problem is that the #Resp measure, although it is intended to be a measure of spoken word recognition, is not sensitive to *which* word listeners are recognizing. The #Resp measure is really only a measure of the degree of agreement among subjects, and is only a partial measure of that, as a split of ten responses of one word and one of another will count the same as a split of five responses of one word and six of another. It shows the degree to which listeners have narrowed down the pool of possible words, but not which words remain in that pool, or how consistent listeners' judgments about those words are.

One solution to this problem with the #Resp measure is to evaluate the number of responses which have a particular segment, namely to use the %Corr measure. This reflects only one small part of spoken word recognition, however. One of the methods of analysis Grosjean (1980) uses offers another solution to this problem: he graphs the responses given against the number of subjects to give each response over time. This is similar to the method used in Tables 4.8 to 4.12 in section 4.4 above. The detailed information available in such a representation allows one to see exactly what change in subjects' perceptions led to a change in their responses, but is very difficult to evaluate quantitatively. Thus, I have only attempted to apply this method of evaluating spoken word recognition to a small subset of the data for which the results are otherwise difficult to interpret, namely those with the maximal change in the two perceptual measures located at widely separated gates. In sum, I believe the best approach to evaluating the results of this experiment for spoken word recognition is to balance the information available from the #Resp measure, the %Corr measure, and detailed analysis of how many subjects gave which responses. I have attempted to balance these three sources of information in appropriate ways throughout the analyses presented in previous sections.

The experiment does present indirect evidence that spoken words are recognized primarily through dynamic, not static cues. Recall from Section 5.1.3 that significantly more words than would be expected to by chance have the maximal change in percent

correct (%Corr) surrounding a  $D_{\max}$  point, and that the majority of exceptions appear to depend on dynamic cues which change too slowly to cause a peak in the measure D. Furthermore, in Section 4.4.2, I showed that in more than three quarters of all words, progress in spoken word recognition is closely linked in time to progress in perceiving the target segment. Even for the minority of words in which the change in the two perceptual measures takes place at widely differing points of the word, the change in number of responses (#Resp) is usually the result of listeners perceiving some segment or distinctive feature other than the target segment, or of them perceiving a particular distinctive feature of the target segment without accurately identifying the entire segment. That is, even in this small subset of the data for which spoken word recognition is not obviously related to perception of the target segment, the listeners' progress in spoken word recognition usually is the result of their perceiving some phonological feature. When it is not clearly related to their perception of some phonological feature, this is either because of the problems with the #Resp measure discussed above, or simply because of obviously noisy data (random variance).

Using many more subjects, perhaps as many as 50 for each stimulus, would probably solve the problem of random variance in the #Resp measure, and I believe would lead to the maximal change in number of responses surrounding the point of maximal spectral change more often. However, this is impractical. Since the results of this experiment show that listeners' progress in perception of segments involves dynamic cues disproportionately often, and also show that progress in spoken word recognition takes place through the perception of some segment or feature, I conclude that listeners recognize spoken words primarily through dynamic cues as well as perceiving individual segments that way.

### 5.3. Implications for the cohort model

In this section, I will discuss some aspects of the results of this experiment which have implications for models of spoken word recognition. Although there are many models of spoken word recognition, I choose to address primarily the cohort model (Marslen-Wilson and Welsh 1978, Marslen-Wilson 1987, 1990, Marslen-Wilson and Warren 1994, among other works). The words for this experiment were chosen with reference to the cohort model (to control for the change in number of possible word candidates during the transition of interest), so the results apply most clearly to that model. Since lexical neighborhood density (Pisoni et al. 1985, Luce et al. 1990) was not manipulated or controlled in the word list, there are fewer implications for a lexical neighborhood model.

#### 5.3.1. Recognition need not proceed from the first phoneme

In early versions of the cohort model, for example as described by Marslen-Wilson and Welsh (1978), the group of competitor words (the group of all words which are possible candidates for what word the listener might be hearing) was determined by finding all of the words in the lexicon which had segmental content compatible with the acoustic information in the initial 150-200 ms of the word. Whatever phonemes or distinctive features of segments the early acoustic information allowed the listener to perceive, all and only the words in the lexicon which began with those phonemes or features were members of the initial cohort. The model specified that after that point, as more acoustic information becomes available, members of the initial cohort which are not compatible with the new acoustic information drop out of consideration. For example, when a listener hears /kæ/ but has not yet perceived anything about what follows, the cohort was said to include "cat, cap, captain, catch" etc. (all words beginning with /kæ/ but no others). Top-down

information, such as syntactic and semantic context, could also affect the set of competitor words after the initial cohort had been formed.

This model has since been modified in several ways, especially in Marslen-Wilson (1987). Other models, such as TRACE (McClelland and Elman 1986) and the neighborhood activation model, or more generally the concept of the lexical neighborhood as the competitor set, have also been proposed (Pisoni et al. 1985, Luce et al. 1990, and others). The modification which is most relevant for my results is the issue of the initial cohort being determined exclusively from the acoustic information at the beginning of the word. Pisoni et al. (1985) point out that words can be recognized even if early acoustic information is produced or perceived incorrectly, and even if the very first phoneme is incorrect. Salasoo and Pisoni (1985) show that words in context can be recognized based on the end of the word's signal, even if the beginning is replaced by noise. Taft and Hambly (1986) also point out as an example that if a listener hears [grakədə'l], either because the speaker made a speech error or because the signal was ambiguous for the listener, the listener is likely to correctly recognize "crocodile" despite the mistaken first phoneme. The early version of the cohort model would fail in this situation: if a listener fails to perceive the first phoneme as the speaker intended it, the intended word would not even be a member of the initial cohort, so the listener could never recognize the word correctly.

In models based on the lexical neighborhood, the group of candidate words from which a listener recognizes one is not calculated from the first few phonemes of the word, but rather from the degree to which the stimulus is similar to any word in the lexicon, regardless of which part of the word is similar. The group of candidate words is considered to consist of all words in the lexicon which can be generated from the intended word by addition, deletion, or substitution of any one phoneme (Luce et al. 1990). This is clearly a much larger group of words than the cohort model proposes. However, the perceptual confusability of each phoneme of the intended word with each phoneme of each

candidate word is also considered in determining how likely the intended word is to be recognized. Furthermore, the tendency for listeners to recognize high frequency words more easily than low frequency words is modeled more directly in the lexical neighborhood model than in the cohort model, because both the frequency of the intended word and the frequency of all words in its lexical neighborhood enter into the calculation for the probability of the intended word being recognized. The pool of candidate words in this model includes words with the initial phoneme of the word differing from the intended word, and even words with an additional phoneme added or a phoneme deleted, so this model can account for listeners' ability to recognize words even if they fail to recognize the initial phonemes correctly.

The TRACE model also allows for correct recognition of a word even if an early phoneme in it is misperceived. This is because it allows for activation of words based on features or phonemes at any point in the signal (McClelland and Elman 1986). Marslen-Wilson (1987) modified the cohort model to allow for activation of potential candidate words on the basis of overall acoustic similarity with the stimulus, even if they would not be part of the initial cohort. In this paper, he also suggested a change to the method of including word frequency effects in the model, and the result of these two changes is to make the revised cohort model considerably more similar to both the lexical neighborhood model and TRACE.

The results of my experiment address the question of whether the pool of candidate words is calculated based on the initial segments of the word. In the experiment reported here, I used only final gating, not initial gating, so listeners should have enough acoustic information to perceive all segments of the word before the transition of interest, regardless of which gate they heard. One should note that the beginning of the stimulus word was always left unaltered, and 400 ms of silence which had been recorded with the stimulus word was left before it in the actual stimuli. Thus, there is no possibility of clicks from splicing or other artifacts altering the phonetic cues available for segments before the

transition of interest. Because the shortest gate ends approximately half way through the first segment of the transition of interest, unless the preceding segment is of a type that cannot be accurately perceived until late in the following segment (glides may be like this, as discussed in Section 5.6.2 below), listeners at all gates have adequate cues to all segments up through the segment before the transition of interest.

In gating experiments, the responses given by the entire group of listeners to hear a particular stimulus are taken to represent the pool of candidate words at that point in the signal (Grosjean 1980, Tyler and Wessels 1985, Tyler 1984). Some researchers have objected that the gating paradigm allows for excessive postperceptual processing, and does not reflect the pool of words which listeners are considering during the online process of spoken word recognition (an objection addressed by Tyler and Wessels 1985). However, most clearly online tasks, such as priming and word spotting experiments, do not allow the listener an open response format. They allow the researcher to test whether one particular word is activated during speech perception, but not what other words the listener may be considering which the researcher did not predict. The gating paradigm is therefore the most useful way to identify a large and not predetermined group of words which listeners are considering as candidates during spoken word recognition.

The results of my experiment show that the pool of candidate words often contains words which differ from the stimulus word in the early phonemes, even when listeners are allowed to hear into the second syllable of the word. I discussed such responses in Section 4.1.2.2 above with regard to deciding what counted as correct for the %Corr measure, but these responses are also relevant for this theoretical point about the pool of candidate words. Responses given in the experiment with added phonemes before the transition of interest include "affair" as a response to "fair" /eɪr/, "afraid" to the stimulus "fragile" /fr/, and "max" to the stimulus "axe" /ks/. Responses with divergent phonemes include "morph" to "wharf" /rf/, "bounce" to "doubt" /aʊt/, "thank" to "fans" /nz/, and "reduction"



to "induction" /ʌk/. In the Japanese data, such responses were less frequent, but one which both adds a phoneme and alters another at the beginning of the word is /ryohi/ [rjoçi] 'travel expenses' for /dohyoo/ [doçjoo] 'wrestling ring.' Even larger divergences appear: several subjects responded with "reading" to the stimulus "reconditioned" /nd/<sup>13</sup>. "Itch, epicure, hit, attend, until" all appear as responses to "petition" /ti/, even though listeners hear at least through the burst of the /t/. "Can, chance, fence, transport, hamster" all appear as responses to "fancy" /ns/, although only "fence" (probably the least divergent from the stimulus) was given by more than one subject.

Although the majority of responses match the stimulus for the segments before the transition of interest, these examples should be sufficient to show that words which have some acoustic similarities to the signal, but do not have the same initial segments, can enter into the pool of candidate words listeners consider during spoken word recognition. Some responses may only reflect that the listener was tired or not concentrating, but most of the examples given here have strong acoustic similarities to the stimulus words in some segments, so they are probably not just errors. In future work, I will examine the similarities between such responses and the stimuli systematically to determine what factors favor the inclusion of such words in the candidate pool. A preliminary inspection of the data indicates that at least in English, the stressed syllable or its rhyme may be the most likely to be identical to the stimulus, with divergences in unstressed syllables tolerated more often.

From this subset of the data, it is clear that the early version of the cohort model, with its focus on the initial cohort calculated from the first few segments of the word, is not the correct model of how listeners choose the group of words from which they will recognize one. The lexical neighborhood model, with its emphasis on overall degree of

---

<sup>13</sup> This may be partially the result of the cohort of /rikən/ being small, although "reconsider" is a relatively common word.

similarity to the stimulus' acoustic signal, is more nearly able to account for the divergent responses discussed here. However, the method of calculating the lexical neighborhood through addition, deletion, or substitution of any one phoneme will also fail to include some of these responses, such as "reduction" for "induction" /ʌk/, despite the similarities between the two words. Some activation models, such as TRACE, may allow for such responses, but a detailed modeling study would be necessary to be sure. A more complicated model of relative similarity between the acoustic signal and other words in the lexicon is probably necessary<sup>14</sup>.

### 5.3.2. Increases in the number of candidates over time

The early version of the cohort model (Marslen-Wilson and Welsh 1978) proposed that listeners generate the initial cohort for a stimulus, consisting of all lexical items which are consistent with the first 150 to 200 ms of the input acoustic signal, and that incorrect members of the initial cohort then drop out as more acoustic information is heard. Eventually, the only remaining member of the cohort will be the one intended by the speaker, and at this point, the listener can recognize the word, since it is the only word in the lexicon which begins with that string of phonemes. Marslen-Wilson (1987:80-81) gives as an example the case of a speaker intending the word "trespass." When the listener has perceived just /tre/, the cohort will include "trend, trends, trestle, tresses" etc. as well as "trespass." Once the listener perceives the /s/, all but "trestle, tresses, trespass" will drop out. Once the listener perceives the /p/, "trespass" will be the only possibility remaining in the cohort. In the early cohort model, reducing the cohort to a single possibility was proposed as the primary mechanism for how listeners recognize word

---

<sup>14</sup> Klatt (1979), with the LAFS model, and Johnson (1997) with an exemplar model offer similar approaches for mapping the acoustic signal directly onto the lexicon. However, their approach provides no way to determine the set of words a listener might consider as candidates during spoken word recognition unless one uses a particular token of a particular word, and has a complete database of memorized spectral representations.

boundaries. (The early cohort model requires the listener to know where the word begins in order to find its cohort, so if one could locate the end of the previous word by narrowing its cohort down to a single possibility, this would allow one to find the beginning of the next word.)

However, many researchers have pointed out that this strategy will not work for embedded words, such as "mar" or "mark" in "market" or "marquis" (Shillcock 1990:26). If the speaker intends "mar," the cohort will never be reduced to one possibility, so the early cohort model cannot allow the listener to recognize "mar," at least not without backtracking after hearing the beginning of the next word. The revised version of the cohort model (Marslen-Wilson 1987) seems to allow the listener to keep track of multiple cohorts beginning at different times, which may overlap with each other. That is, when hearing "market," the listener could be evaluating the cohort beginning at /m/ and the cohort beginning at /k/, and possibly others, simultaneously, to allow recognition of embedded words. Shillcock (1990) gives the example of "recognize speech," which in relatively fast speech can be homophonous with "wreck a nice beach," and could contain many other embedded words, such as "reckon, an, ice, eye, bee," etc.

However, even with the modification that word segmentation need not depend on the word initial cohort, no version of the cohort model makes any provision for the cohort of candidate words increasing over time<sup>15</sup>. It requires that all possible words be included in the initial cohort, and focuses on the time course of elimination of words from that cohort. This experiment and other research show, however, that an increase in number of words under consideration may be possible during the recognition of low frequency words or long words containing embedded words.

---

<sup>15</sup> The only possible way for the cohort to increase over time, and only in the revised cohort model, is if a word is similar to the stimulus signal at points after the initial few segments, but not during the initial few segments. This is the case discussed in the previous section with regard to listeners recognizing words without correctly hearing the beginning of the word. This sort of increase in the cohort is not at issue in this section, and does not affect the cohorts of embedded words.

Grosjean (1980), in his gating experiment, identified embedded words as one type of "garden path" in gating. He showed that when two syllable words, of which the beginning also forms a word, are presented out of context, most subjects will propose the shorter embedded word until there is enough acoustic information available to make that word impossible. For example, Grosjean's subjects tended to respond with "cap" to "captain," "stretch" to "stretcher," "parse" to "parsley," etc. There appear to be two possibilities for how subjects switch over to the longer response once the embedded response is ruled out. Grosjean shows the response data for "stretcher" and "captain." For "stretcher," the longer word was given by one subject even at the gates where most subjects respond "stretch." Also, isolated subjects responded with "strength" and "strudge" at gates where most responded with "stretch." Then, once "stretch" is ruled out, all subjects begin responding with "stretcher," the correct word. Although there is some variation, the overall pattern is a change from approximately three different responses, with most subjects responding "stretch," to all subjects responding "stretcher." Despite the embedded word, there is a decrease in the cohort size (as reflected by the number of responses) when the embedded word is ruled out.

However, for the word "captain," the results are different. Once most listeners have perceived the /p/, almost all respond with "cap," and a few respond with "captain" and "calculate." However, when enough acoustic information becomes available for listeners to know that /p/ cannot be the end of the word, the number of different responses increases to six for one gate, before subjects begin to agree on "captain" (Grosjean 1980:277). This is the pattern which the cohort model cannot account for. Grosjean does not discuss this difference between "captain," where the number of responses increases briefly, and "stretcher," where there is a large shift in which word is being recognized along with a small decrease in number of responses. One cannot tell how many of the embedded words in his experiment showed an increase in number of responses.

Some of the words in my experiment show similar results, with an increase in number of different responses when listeners realize that the embedded word cannot be the entire word. Examples are shown in Tables 5.2 for the word "asthma" /æz/, 5.3 for "unconcealed" /ns/, and 5.4 for /himoto/ 'origin of a fire.'

Table 5.2. Responses to "asthma" /æz/. Number of subjects to give each response appears after each word. Number of responses (#Resp) increases between gates 4 and 5, then decreases again somewhat.

Gate 1		Gate 2		Gate 3		Gate 4		Gate 5	
accident	1	act	3	acme	1	Alice	2	Alice	1
acronym	1	and	1	alibi	1	allergy	1	alley	1
act	1	animal	2	alphabet	1	apple	6	alligator	1
action	1	apparatus	1	Amsterdam	1	elevator	1	allocate	1
actually	1	apple	2	appetite	1	else	1	ask	2
apple	2	at	2	apple	6			asthma	1
ash	1							elephant	2
ass	1							elevate	1
at	1							Esther	1
ax	1								

Gate 6		Gate 7	
ask	1	as well	1
asper	1	aspen	1
aspirate	2	asphalt	1
aspiration	1	aspirate	2
aspirin	3	aspirated	1
asthma	3	aspirin	3
		asthma	2

Here, the number of responses (#Resp) increases sharply at the fifth gate, after which it falls again somewhat. The endpoint of the fifth gate is only a few milliseconds after the onset of frication for the /z/, and it is at this gate that the previously common response "apple" disappears. Subjects are not able to recognize the fricative well yet. (Section 4.3.1.5 above discussed the tendency for fricatives to be recognized well after the onset of frication.) However, they seem to be ruling out the high frequency word "apple," perhaps because of the labial or the stop. The number of responses goes up until subjects hear enough of the frication noise to perceive the alveolar fricative, reducing the number of

possible responses again. (The intended /z/ is never perceived well in this word, but the number of responses with /s/ increases sharply at the final gates.)

Table 5.3. Responses to "unconcealed" /ns/. Number of subjects to give each response appears after each word. Number of responses (#Resp) increases between gates 2 and 3.

Gate 1	Gate 2	Gate 3	Gate 4	Gate 5
unconditional 9	incandescent 1	onion 1	uncle 1	inconceivable 1
uncontrolled 1	uncomfortable 1	onions 1	unconcerned 4	inconsiderate 1
unconventional 1	unconditional 7	uncommon 1	unconscious 2	unconcerned 5
	unconditioned 1	unconcerned 4	unconsiderate 2	unconsiderate 1
	uncontrollable 1	unconscious 1	unconsistent 1	unconstitutional 1
		unconsecrated 1	unconstant 1	unconstrained 1
		unconsiderate 1		uncoordinated 1
		uncontrollable 1		

In "unconcealed" /ns/, there is an increase in the number of different responses, especially between the second and third gates. The onset of frication is near the second gate. At the first and second gates, most subjects respond with "unconditional," a relatively common word in the cohort /ʌnkən-/. At the third gate, "unconditional" seems to disappear, presumably because there is enough phonetic information to rule out the /d/. However, there are very few words in the cohort /ʌnkəns-/, and they are of low frequency. Many Latinate words beginning with "cons-" take the negative prefix "in-" instead of "un-," which may make it difficult for subjects to think of an appropriate response. (Some listeners coined words with "un-" to solve this problem, such as "unconsiderate, unconstant.") Thus, since all members of the cohort are of low frequency, the number of different responses increases when listeners cannot respond "unconditioned," even though the number of words one would predict to be in the cohort from a dictionary search decreases.

Table 5.4. Responses to /himoto/ 'origin of a fire.' Number of subjects to give each response appears after each word. Number of responses (#Resp) increases at gate 7.

Gate 1		Gate 2		Gate 3		Gate 4		Gate 5	
himo	10	himo	7	himo	8	himo	7	himo	9
himoNya	1	himono	3	himono	2	himoti	2	himono	2
himoti	1	himotoku	1	himotoku	1	himoto	1	kimoti	1
		himozii	1	himozii	1	himozii	2		

Gate 6		Gate 7		Gate 8		Gate 9	
himo	5	hi	1	himo	4	himo	4
himono	4	himo	2	himono	3	himoto	6
himoti	2	himono	4	himoti	1	himotoku	2
himotoku	1	himotoku	2	himoto	1		
		himotuki	1	himotoku	2		
		himozii	2	himotuki	1		

Glosses: /himo/ 'string,' /himoti/ 'burns long,' /himoNya/ 'ʔ<sup>16</sup>,' /himono/ 'dried fish,' /himotoku/ 'read a book,' /himozii/ 'hungry,' /kimoti/ 'feeling,' /himotuki/ 'conditional, with strings attached,' /himoto/ 'origin of a fire,' /hi/ 'fire'

In /himoto/ [çimoto] 'origin of a fire,' which is not a very common word, there is an increase in the number of different responses for the seventh and eighth gates, which are the last two gates before the burst of the /t/, both falling during the /t/ closure. Up until the sixth gate, large numbers of subjects at each gate responded with /himo/ 'string,' a relatively common word which is embedded in the stimulus word. Although there are still several responses of /himo/ at each of the later gates in the word, the number of those responses decreases during the sixth and seventh gates. However, at the seventh and eighth gates, listeners have not perceived all features of the postvocalic [t] well yet. (The postvocalic obstruents in /himotuki/, /himoti/, and /himozii/ are all phonetically affricates. It seems that at these two gates, most listeners have perceived the alveolar place of the following segment, but many have not perceived the manner correctly yet. Since the release is expected to differentiate stops from affricates, and these gates end during the stop closure, this is not surprising.) By the last gate, a majority of listeners have perceived all features of the [t], and consequently narrowed in on /himoto, himotoku/, but at the stage

<sup>16</sup> This word may be a place name, or may be a nonce word. A native speaker of Japanese could not identify it.

between when responses of /himo/ become less likely and when many listeners perceive the [t] correctly, the number of different responses increases.

In all of these cases, the target word is relatively uncommon, and there is either an embedded word or some other member of the initial cohort (as defined by cohort theory) which is more common. Not surprisingly, listeners often respond primarily with the more common word until acoustic information makes that impossible. At that point, the listeners may all switch to the target word, leading to little change in the number of different responses. Alternatively, they may switch to a variety of other words until later acoustic information allows them to narrow the responses down again, leading to a temporary increase in the #Resp data, as in the examples given here. This increase in number of responses is difficult to account for in the cohort model. One could claim that all of the later responses are also present in the initial cohort<sup>17</sup>, but that very few listeners will give these responses because more frequent members of the cohort are available. It is questionable, however, whether one can claim that a word which no listener proposes as a response during early gates of a word is truly under consideration as a candidate word.

The revised cohort model does allow for word frequency to influence the activation level of particular words (Marslen-Wilson 1987), and this could allow for modeling of these sorts of effects. The model does not, however, build in a preference for embedded words over longer words. The TRACE model does have an automatic preference for embedded words, because when a listener has heard a signal /kæp/, the signal so far is more similar to "cap" than to "captain," and will therefore activate "cap" more strongly. If the signal then continues, and additional segments follow /kæp-/, the activation of "cap" will begin to drop. It is not clear, however, whether the TRACE model would show increases in the total number of words being activated in such cases.

---

<sup>17</sup> This seems to be true for /himoto/, since all of the responses given at the gates with the largest number of different responses also appear at some earlier gate.



Although the cohort model does allow for an influence of word frequency, which might be able to model the number of words with high activation increasing when a high-frequency word is ruled out, research on the cohort model focuses almost exclusively on the narrowing of the cohort over time. It does not address the possibility that the number of words in the cohort which are *likely* to be recognized by listeners may increase over time. The assumption of this experiment that the time during which listeners make the most progress in recognizing a word would be represented by a decrease in the number of different responses was based on this model, in which the cohort is always narrowed over time. I suggest that instead of using a model in which the cohort consists of all words with sufficient resemblance to the acoustic signal, or all words with the same segments as the part of the signal so far perceived, one could define an *effective cohort*. That is, the cohort model should be modified to reflect the fact that some words which do have the same initial string of segments are not realistically in the pool of words listeners are considering at some stages, because they are of too low frequency and other words with higher frequency are available, or because they have relatively high frequency embedded words. Once higher frequency words or embedded words are ruled out by additional acoustic information, these low frequency or longer words are the only ones which are possible, so they become realistic competitors.

The fact that 13 of the 127 words in the English experiment (10.2%), and 21 of the 76 in the Japanese experiment (27.6%) have a positive slope for their #Resp data confirms the importance of the increase in number of candidate words phenomenon. The examples "unconcealed" and /himoto/ above have positive slope. Many other words in the experiment, such as "asthma" above, have a temporary increase in number of responses (#Resp), but an overall negative slope. These words are therefore included in the calculations, but the ogival curve fit to the data cannot fit the increase in #Resp, so the increase must be treated as noise. This contributes to overall error in the results. The fact that the number of responses, in words where it increases, often does so just where

listeners are beginning to perceive the target segment and therefore rule out a previously common response means that instead of finding the maximum decrease in #Resp at the  $D_{max}$  point, one may find an increase in #Resp there instead. This contributes to the large number of words which fail to have the area of maximal decrease in #Resp surrounding  $D_{max}$ . Increase in the effective cohort size, despite decrease in the dictionary defined traditional cohort, is probably in part responsible for the failure of the #Resp data to follow the hypothesis as well as the %Corr data does. It is clear that there are cases in which the number of words being seriously considered by listeners increases as more acoustic information becomes available instead of decreasing, and this should be incorporated into future models of spoken word recognition.

### 5.3.3. The size of the unit used to narrow the cohort

Researchers have suggested a wide variety of units as being important in speech perception and spoken word recognition. The issue is what size units listeners use in comparing the speech they are perceiving to lexical entries in order to recognize words. Many models of spoken word recognition use the phoneme as at least one unit of spoken word recognition (McClelland and Elman 1986). In such models, when a listener recognizes a phoneme, she can narrow down the pool of words from which she is recognizing one by comparing that phoneme to the phonemes of the lexical entries. (Using an alternative metaphor, when a listener recognizes a phoneme, activation of words which have that phoneme increases and activation of words without that phoneme decreases.) Greenberg (1996b) argues that the syllable is the unit of both speech perception and lexical representations in English, although he does not specifically claim that it is also used to link perception and lexical representations through spoken word recognition. Mehler et al. (1981) also suggest that listeners segment the speech stream into syllables before accessing the lexicon, although later results by some of the same authors showed that the perceptual effect on which this claim was based is language specific. Klatt (1979) and Johnson

(1997) both propose that lexical access can be done without any phonological units, even distinctive features, by comparing time slices of the waveform to stored acoustic representations of words. Warren and Marslen-Wilson (1987), Lahiri and Jongman (1990), and Marslen-Wilson and Warren (1994) all argue that speech is compared to lexical representations at the level of the distinctive feature, without reference to phonemes or any higher unit. Lahiri and Marslen-Wilson (1991, 1992) also take this approach, although the level of the segment is important in their definition of which distinctive features are used in perception, since they use the concept of particular segments being marked or unmarked for particular features in their model.

One of the primary arguments Marslen-Wilson and Warren offer for the view that distinctive features and not phonemes are used in spoken word recognition is that in gating experiments, listeners narrow down the cohort as soon as they perceive some feature of a segment, even if they have not perceived the entire phoneme correctly yet. This is reflected by a change in the responses listeners give: if they can perceive that a postvocalic stop is voiced, they stop giving responses with voiceless postvocalic stops, but their responses may still include postvocalic stops with a variety of places of articulation. That is, listeners do not wait until they can perceive all features of a phoneme to use that information in spoken word recognition, but rather use whatever cues are available, as soon as they are available. Marslen-Wilson and Warren (1994) conclude that models which require an intermediate stage of processing before completing spoken word recognition, such as identifying entire phonemes or syllables, are not possible.

The results of my experiment confirm the result which supports that argument, namely that listeners narrow down their responses based on individual distinctive features without waiting to perceive the entire phoneme correctly. Although I have not presented a complete analysis of the timing of perception of each feature for each target segment, three subsets of the results which I report with regard to other issues demonstrate the use of individual features in spoken word recognition. In Section 4.4.3, where I analyzed the

words for which the maximal change in the two perceptual measures does not happen at or near the same time, I showed that this is often because the maximal change in the #Resp measure reflects a change in perception of an individual feature. This is sometimes a feature of the target phoneme, sometimes a feature of a neighboring phoneme. In the word "crops" /kr/, the maximal change in #Resp seems to reflect listeners' perception of the vowel as [-high]. In "Italian" /ɪt/ and "mechanical" /ək/, the maximal change in #Resp reflects listeners' recognition that the target stops are aspirated, which allows them to perceive features of the following segment.

Secondly, many of the cases in which the target segment is never perceived well, even at the last gate, provide indirect evidence for the role of features in spoken word recognition. As described in Section 4.1.2.2, if the correct responses for a target segment never exceed 30% at any gate, the %Corr measure for that word was calculated instead by a measure of two out of three of voicing, place, and manner correct. In most such cases, this "two features correct" measure (counting place, manner, and voicing as the "features") was fit well by an ogival curve, and had a sudden increase at some gate. For example, in the word "ravish" /æv/, few listeners perceived the /v/ correctly at any gate, with only a few responses of "ravish" or "ravage." However, there was a clear shift from large numbers of responses of "rat, rap, brat" at early gates to a majority of responses of "rabbit" at late gates. This shows that listeners perceived the voicing and place of the obstruent correctly at later gates, and used this information to narrow the cohort, even though they had not perceived the manner of that obstruent yet. Similarly, in "asthma" /æz/, discussed in the previous subsection, the /z/ was never perceived correctly by more than 30% of subjects, but at late gates most listeners perceived the frication and place of that segment, and gave responses with /s/. The flap of "muddy" /ʌd/ was never perceived more than 30% correct, but by the final gate most listeners responded "mud," with the place and voicing correct, and the manner correct at the phonemic but not phonetic level. The existence of an area of

sudden increase in the percent of listeners to get two features of a segment correct, even when they do not get all three features correct, provides evidence that listeners can use individual distinctive features of a segment in spoken word recognition before they can perceive the entire phoneme.

The third subset of my results which shows use of individual distinctive features in spoken word recognition will be discussed in Section 5.7.1 below, where I show that listeners can perceive the voicing of a postvocalic obstruent correctly before the obstruent begins. The responses involved make it clear that the listeners had not perceived the place or manner of these postvocalic obstruents correctly yet by the time their judgments about voicing affected their responses.

All of these aspects of the data serve to reconfirm the results of Warren and Marslen-Wilson (1987), Lahiri and Jongman (1990), and Marslen-Wilson and Warren (1994). This does not necessarily mean, however, that there cannot be a phonemic or syllabic level of lexical representation, only that listeners can compare the input to lexical entries at the featural level, and are not restricted to comparing to the lexicon at higher levels. That is, this result only supports a featural level of lexical representation, it does not rule out other levels. Furthermore, the models without features proposed by Klatt (1979) and Johnson (1997) can probably simulate the effect of narrowing down the cohort based on individual distinctive features, since the phonetic cues to the distinctive features would be compared directly to the lexicon. The TRACE model (McClelland and Elman 1986), which Lahiri and Marslen-Wilson mention as an example of a spoken word recognition model which uses phonemes, would probably also show effects on activation of lexical items as soon as any distinctive feature is perceived, without waiting for the entire phoneme to be perceived correctly, since its distinctive feature level activates phonemes, which activate words. In sum, my results confirm the previous finding that listeners can use individual distinctive features in spoken word recognition without waiting to perceive any

higher level unit, but my experiment does not otherwise distinguish between the theories positing various sizes of units for lexical access and representation<sup>18</sup>.

#### 5.4. Effects of suprasegmental units

Suprasegmental aspects of the phonological structure, such as location of stress or pitch accent and position of the segment to be perceived in the syllable or foot, had relatively little effect on the timing of perception of segments. The one clear suprasegmental effect on perception of segments regarded stress and postvocalic stops in English. As discussed in Section 4.3.1.1, postvocalic stops are often perceived during the preceding vowel, early relative to their  $D_{\max}$  points, which fall at the onset of the stop closure. The probable explanation for this is that there are sufficient cues to place of articulation and voicing of the following stop present in the vowel for listeners to make considerable progress toward perceiving the stop during the vowel, especially if the cohort of the word disfavors other manners of articulation. However, this effect appears in the English data only when the vowel of the VC transition is stressed. Of the four words in the English experiment with VC transitions involving an unstressed vowel<sup>19</sup>, the maximal change in percent correct (%Corr) surrounds the  $D_{\max}$  point at the closure of the stop for two words, surrounds the  $D_{\max}$  point at the release for one word, and falls between the closure and release for one word. None of the VC transitions with the vowel unstressed have the area of maximal perceptual change during the vowel.

The explanation for this difference might have little to do with stress itself: early recognition of postvocalic stops appears to be more likely after longer vowels, as was discussed in Section 4.3.4.1. The English words which show this effect are often those with diphthongs, and never the VC transitions with /ɪ/ as the vowel. Furthermore, the effect is more frequent in the Japanese experiment, and occurs only with the vowel /a/.

---

<sup>18</sup> It was not designed to do so. One would need a very different type of experiment to do this.

<sup>19</sup> This excludes one unstressed VC word which was more linear than ogival.

Thus, the unstressed vowels in English may simply be too short for listeners to recognize the postvocalic stop before reaching it, partly because there are very few gates in an unstressed VC transition before the onset of the consonant.

An alternative explanation involves the results Fujimura et al. (1978) found in their cross-splicing experiment. As described in Section 1.5, Fujimura et al. found that when VC and CV transitions conflict as to the place of articulation of the C in a spliced VCV sequence, both Japanese and English listeners identify the consonant based on the CV cues. However, if the first vowel is perceived as stressed, the English listeners put more emphasis on the VC cues than if it is unstressed. The Japanese listeners show no such effect for pitch accent. Fujimura et al. related this difference to the syllable structure of Japanese, in a way which will be discussed further in Section 5.5.2 below. However, this result also fits well with the results of many psycholinguistic experiments which show that English listeners pay a great deal of attention to stressed syllables, at least for word segmentation purposes (Cutler and Butterfield 1992, Cutler and Norris 1988, McQueen et al. 1994, Cutler et al. 1996). If the early recognition of postvocalic stops only after stressed vowels in my experiment is not simply due to vowel duration differences, it would provide a confirmation of Fujimura's and colleagues' result. That is, English listeners may be able to use VC transition cues earlier and more effectively if the vowel is stressed than if it is not, simply because they pay more attention to the stressed parts of a word. In the results of my Japanese experiment, pitch accent has no effect on whether the stop in a VC transition is recognized early relative to  $D_{\max}$  or not. Stops before the accented mora and after it are about equally likely to be recognized before the  $D_{\max}$  point. This also mirrors the results of Fujimura et al., in that they found no effect of pitch accent on Japanese listeners' weighting of VC and CV cues.

The experiment by Fujimura et al. (1978) did not have the confound with vowel duration my experiment does, since all materials were recorded in Japanese, which does not reduce unaccented vowels. The stress manipulation was actually a pitch accent

manipulation, taking advantage of the fact that English listeners will interpret a Japanese VCV utterance with first mora pitch accent as having first syllable stress. It would be useful to confirm the difference in English listeners' early use of VC cues for stressed and unstressed vowels by performing a gating experiment with the duration of vowels controlled. However, the results of the current experiment do provide preliminary evidence that English listeners are able to use cues to a postvocalic stop earlier in the vowel if the vowel is stressed than if it is unstressed. This confirms yet further the importance of stress in English listeners' speech processing strategies.

One might expect the position of stress, if not pitch accent, to have further effects on the timing of listeners' use of cues to segments, given the important role of stress in English phonology. Furthermore, one might expect the timing of perception of consonant-consonant transitions to differ when they are entirely within a syllable, split across a syllable boundary, and split across a foot boundary, in consideration of the work by Cutler and colleagues (Cutler and Butterfield 1992, Cutler and Norris 1988, McQueen et al. 1994, Cutler et al. 1996) on the importance of such boundaries for word segmentation in English. Using a variety of experimental approaches, these authors have found that English listeners can recognize a word more easily if its first syllable is strong (if it has an unreduced vowel, which is often stressed). Cutler and Norris (1988) also find that a CVCC real word such as "mint" is difficult for listeners to spot if it is embedded in a CVCCVC non-word with a strong second syllable such as /mmte<sup>j</sup>v/, because listeners hypothesize a word boundary at the beginning of the strong syllable, which puts a word boundary between the final CC cluster of the real word. This is closely related to the manipulation of boundary location relative to CC clusters used in the present experiment.

However, as discussed in Section 4.5, there are no significant effects of stress or pitch accent location, of the transition of interest being located in the first or second syllable of the word, or of location of a CC transition relative to suprasegmental boundaries. This lack of significance is clear when one evaluates the data in terms of rise time of the %Corr



measure (the speed with which listeners go from not perceiving a segment to perceiving it). With the exception of the stress difference for the postvocalic stop effect just discussed, there is also no effect of any of these suprasegmental manipulations on the likelihood of a segment being identified at its  $D_{\max}$  point.

The lack of effects for these suprasegmental manipulations is ambiguous. It may be that, as important as the location of stress and of syllable boundaries is for word segmentation, as shown by Cutler's extensive work, these suprasegmental factors have little influence on the timing of how listeners use phonetic cues to perceive segments. However, before concluding that a certain effect does not exist, one needs to perform a very sensitive experiment in order to avoid the danger of missing a real effect just because there were not enough subjects in the experiment<sup>20</sup>. In this experiment, the focus was on testing a wide variety of transitions in a wide variety of environments, in order to test the theory of dynamic cues for speech perception in relatively representative conditions. Therefore, only a few pairs of words with each manipulation were included. An experiment designed to test any one of these suprasegmental effects separately, with many more pairs of words, might find effects on the length of time over which cues are used (rise time) or on the timing of use of cues relative to  $D_{\max}$ .

### 5.5. Language specific effects of phonology on perception

Because the composition of the word lists for the two languages is strongly influenced by their phonological systems, I have avoided comparing the overall results for the two languages, such as the percentage of words to have the area of maximal perceptual change surrounding a  $D_{\max}$  point. However, as explained in Section 4.3, it is possible to compare the two languages by examining subsets of the words which are more closely matched for environment or type of effect. Therefore, for words which showed each of the

---

<sup>20</sup> For a discussion of the issue of power, especially in an experiment designed to show the lack of an effect, see Keppel (1991).

effects discussed in Section 4.3.1 (early recognition of postvocalic stops and sonorants, late recognition of vowels and fricatives), I compared the degree to which these segments were perceived late or early in the two languages. Unfortunately, this limits the cross-linguistic comparison to small subsets of the data, and the small numbers of words included in each category make it difficult to find significant cross-linguistic differences. In several cases, I suspect that the difference between the two languages reflects a real difference, but it is not statistically significant in this data.

I hope that future experiments directed specifically at these smaller questions can use a larger number of words which are more carefully controlled for the particular factor involved, and can thus establish the validity of these cross-linguistic differences. The current experiment was designed primarily to test the use of dynamic cues in both languages. The main purpose of including a wide variety of environments was to perform a test of dynamic cues under more realistic and thorough conditions than had been done in the previous literature. By including such a wide variety of environments, though, several effects which apply only to subsets of the data were found. Many of these could not have been found using only stop-vowel-stop stimuli, for example. However, the word lists could not include large numbers of words for each environment. The cross-linguistic comparison of these environment specific effects in this experiment should be considered exploratory research for these reasons.

#### 5.5.1. Effect of phoneme inventory

One non-significant but interesting cross-linguistic difference is the tendency for English vowels to be recognized later relative to the onset of the vowel than Japanese vowels are. I believe this effect is the result of differences in the phoneme inventory of the two languages, namely in the size of the vowel system. Japanese has exactly five vowel

phonemes, although each can be phonemically long or short<sup>21</sup>. The length distinction is made almost entirely by duration, not through a combination of durational and vowel quality differences (Jenkins et al. 1997, Uchida 1997, Hubbard, to appear). The dialect of English which the speaker and most of the listeners for the English experiment speak, however, has eleven vowel phonemes which can appear in stressed syllables, excluding the diphthongs /æ<sup>j</sup>, a<sup>w</sup>, o<sup>j</sup>/. Although duration is an important cue for many vowels, there are far greater vowel quality differences than in the Japanese short/long pairs (Hubbard, to appear). It seems likely that English listeners require a longer sample of a vowel in order to identify it correctly than Japanese listeners do, simply because English listeners have more vowels to choose from. Therefore, when listeners hear a word gated out shortly after the onset of the vowel, Japanese listeners may be able to identify the vowel correctly from the short portion of the vowel's signal they heard, while English listeners cannot. In particular, English listeners are likely to need a relatively long sample of the signal in order to distinguish the vowels with the most similar quality, such as /e/ and /æ/, /a/ and /ʌ/, etc.

Although the effect of later vowel recognition in English than in Japanese is not significant in this experiment, there is confirmation for it in other studies. Lang and Ohala (1996), in a gating study of California English CVC and CV nonsense syllables, found considerable confusion between several vowels 50 milliseconds after the onset of the vowel. For some consonantal environments (particularly after /h/), there was substantial confusion between most vowels 40 ms after the onset of the vowel. Even in consonantal environments which did not cause such strong confusions (after voiced stops), some tense/lax pairs of vowels were often confused as much as 100 ms after the onset of the vowel. However, Furui (1986), in his gating study of Japanese, found that final gated vowels were usually identified correctly very soon after the  $D_{\max}$  point, which was at the

---

<sup>21</sup> /aa/ occurs in only a few lexical items, however, and /ii/ appears almost exclusively crossing morpheme boundaries. /ee, oo, uu/ are extremely common in Sino-Japanese morphemes. Thus, while all five vowels can be distinctively long, not all long vowels have the same status in the language.

onset of the vowel. Furui's presentation of his results does not make it clear which vowels, in which contexts, may have been confused for how long into the vowel, but it does seem clear that the Japanese vowels of his study were identified accurately earlier after onset of the vowel than most of the English vowels in Lang and Ohala's (1996) study were.

Furui's results on perception of Japanese vowels seem to disagree with the result in the current experiment, in that Furui finds the Japanese vowels being identified accurately very shortly after the  $D_{\max}$  point, whereas in the current study, approximately one third of the Japanese CV transitions have the area of maximal improvement in perception of the vowel after the gate surrounding the  $D_{\max}$  point. This is despite the fact that Furui uses a shorter gating interval (10 ms), making the window around the  $D_{\max}$  point shorter. However, it is not possible to tell from Furui's article at exactly what point the vowel was perceived in individual syllables. Only the average relationship of  $D_{\max}$  and the percent correct for vowel identification is available. Furthermore, Furui used four listeners, each of whom responded to all gates of each syllable (successive presentation), and each of whom responded to the entire set of stimuli five times. These listeners were also trained extensively, apparently on the same stimuli, for two days before taking the test. Thus, the listeners had had considerable practice with the stimuli by the time they took the test, and the five repetitions of the test by each listener also provide a large amount of data to average over. The use of a closed class test with nonsense word stimuli also reduces the variation. Since the listeners were so familiar with the stimuli and provided multiple judgments for each one, it is not surprising that they were able to identify the vowels more quickly and consistently than listeners in the current experiment could.

In sum, the difference in results for gated vowels between Furui's (1986) Japanese study and Lang and Ohala's (1996) English study, in combination with the current experiment's non-significant difference in delay for vowel identification time, imply that Japanese listeners are able to identify vowels accurately sooner after the onset of the vowel

than English listeners are. This is probably because the English vowel system distinguishes more than twice as many vowels through vowel quality differences as Japanese does. The fact that Japanese has much less reduction of vowels than English may also be a factor. Furthermore, English vowels which are basically monophthongal probably have more change in vowel quality over time than Japanese vowels do, and Nearey and Assmann (1986) have shown this slight diphthongization to be an important cue in English vowel perception. This cue would be removed by gating. Thus, there are several probable reasons for English listeners' relatively late identification of vowels, based partly in the phonemic inventory of the two languages<sup>22</sup>.

Costa et al. (in press), using Dutch and Spanish, also find influences of the size of phoneme inventory on how quickly listeners can perceive sounds. Their study is quite different from the current one: they use a target spotting experiment, and they find that a larger number of phonemes in the consonant inventory makes it more difficult for listeners to factor out consonant-vowel coarticulation when listening for a vowel in a CV sequence, for example. However, their finding that a large phoneme inventory slows down reactions to segments is related to the result of slower vowel perception in English than in Japanese presented here, and confirms the result that size of phoneme inventory affects speech perception.

#### 5.5.2. Effect of syllable structure constraints

A clearer language specific effect appears in the early recognition of postvocalic stops: maximal progress toward perceiving postvocalic stops is made significantly more before the  $D_{\max}$  point at the stop closure in English than in Japanese. In words where

---

<sup>22</sup> The suggestion that listeners may need a longer sample of a vowel to perceive it correctly if they have a large vowel inventory might seem counter to the results on auditory processing presented above showing the importance of the initial 5-10 ms of a signal, during which adaptation takes place. However, auditory nerve fibers do not completely cease responding once adaptation has taken place, the response rate only decreases. The importance of beginnings of sounds (because of adaptation) does not mean that languages cannot make distinctions based on duration, for example. However, change in formants during even steady state English vowels is probably also a factor.

postvocalic stops are recognized early, the area of maximal improvement in %Corr falls an average of 51 ms, or three gates, before the  $D_{\max}$  point in English. In Japanese, it falls an average of just 15 ms before the  $D_{\max}$  point, at the last gate before it. I believe this shows an influence of the syllable structure of a language on listeners' weighting of cues for perceiving certain contrasts.<sup>23</sup>

As explained in Section 1.5 and also discussed in Section 5.4, Fujimura et al. (1978) and Kakehi et al. (1996) both find in splicing experiments that if the VC and CV transitional cues to a consonant in a VCV sequence are made to conflict by splicing together a VC and a CV with different places of articulation, listeners identify the consonant based on the cues in the CV portion of the signal. This is true of Japanese, English, and Dutch listeners. Results such as these have provided a strong argument that cues to place are stronger in a CV transition than in a VC transition. However, under some conditions English and Dutch listeners rely more heavily on the weaker VC cues than usual, and both groups of researchers find that Japanese listeners do not do this. In the experiment by Fujimura et al. (1978), if English listeners perceived the first vowel as stressed, they were more likely to identify the consonant based on the VC transition. Japanese listeners, however, identified the consonant as having the place of the CV part of the signal, regardless of the location of pitch accent. In the experiment by Kakehi et al. (1996), if the burst of the stop and the onset of the following vowel were gated out, Dutch listeners were more able than Japanese listeners to make use of the VC cues, especially when durational cues to the consonant were also altered or removed.

Both groups of researchers mention that Japanese does not make any place distinctions in coda position. Japanese has an extremely restricted syllable structure, with

---

<sup>23</sup> In Section 5.4, I considered the possibility that English listeners' early recognition of the stop if the vowel is stressed, but not if it is unstressed, could be a result of vowel duration differences dependent on stress and not of stress itself. Here, the issue of vowel duration is less problematic, because only the words which demonstrated early recognition of the stop are considered. This restricts the vowels under consideration to the relatively long stressed English vowels and the relatively long vowel /a/ in Japanese, so all of the words being compared have relatively long vowel duration.

only two types of entities which can appear in coda position. These are the mora nasal, which assimilates in place to the following segment, and the beginning of a geminate obstruent<sup>24</sup>, which of course also has the same place of articulation as the rest of the geminate. Assimilation of coda consonants in Japanese is shown in (1).

(1) Possible coda consonants for Japanese

/kaNpai/	[kampai]	'a toast'
/haNtai/	[hantai]	'opposite'
/haNko/	[han̩ko]	'name seal'
*[hamtai] etc.		
/haQpa/	[happa]	'leaves'
/kaQta/	[katta]	'bought'
/saQka/	[sakka]	'author'
*[sakta] etc.		

Because the only two possible coda consonants both assimilate in place to the following onset, Japanese listeners can always recover information about place of articulation from the following segments, the onset and vowel of the next syllable. (The mora nasal /N/ can occur in word final position, but no place distinctions are made for the word final nasal. /Q/ cannot appear word finally.) In a VCCV (including VNNV) sequence, the cues to the place of the CC (or NN) in the CV transition are stronger than the VC cues. The VC and CV cues will always match in Japanese, so Japanese listeners are never forced to perceive place from the weaker VC cues. Fujimura et al. (1978) and to some extent Kakehi et al. (1996) propose this as the reason why Japanese listeners are less likely to make use of VC transition information than English and Dutch listeners are, even when CV transition information is removed or made less relevant.

I believe this is also the reason behind the effect of early perception of postvocalic stops being stronger in English than in Japanese in my data. Since listeners hearing Japanese can always recover place information from the stronger CV cues, they do not make as much use of the weaker VC cues as English listeners do, and hence do not

---

<sup>24</sup> There is a further restriction on codas: geminate obstruents must be voiceless, except in a very few recent borrowings such as /baggu/ 'bag' and /beddo/ 'bed.'

recognize stops as early in the preceding vowel as English listeners do. This result provides confirmation of the cross-linguistic effect Fujimura et al. and Kakehi et al. found through an additional experimental paradigm, namely gating rather than splicing. Furthermore, the splicing experiments show a difference in how listeners of the two languages weight the various cues when all of the cues are available, but this gating experiment shows that the difference is also in the timing of listeners' use of the cues. English listeners are not only more likely to use the weaker VC cues than Japanese listeners are, as shown by the splicing experiments, they are also able to use them earlier. The Fujimura et al. and Kakehi et al. results, in combination with my own, all show that facts about the phonology of a language, even to the level of constraints on syllable structure, influence the strategies listeners adopt in perceiving individual segments of the language.

#### 5.6. Perceptual motivations for phonological universals or alternations

In the previous section, I discussed some results of this experiment which show an influence of the phonology of a language on how listeners perceive it. Results of the experiment suggest that there is an influence in the opposite direction, as well: how individual segments are perceived affects the phonology of the language. The overall purpose of the experiment is to test the use of dynamic cues by determining whether the areas of the signal during which there is rapid improvement in perception fall at areas of great change in spectral information. Identifying the parts of the signal during which listeners most quickly become able to perceive segments, and particularly quantifying the speed with which they do this through the rise time measure (Sections 4.5 and 4.6), offer a way to investigate the perceptual motivations behind some phonological patterns. In this section I will discuss how three common phonological patterns are motivated by facts about the rate of information transfer in speech perception.



### 5.6.1. The preference for onsets over codas

It is a well known phonological (near) universal that the onset position of a syllable is favored, or has a special status, which coda position does not. This status of onsets is seen in several aspects of the phonology of many languages, such as in syllabification rules, in the number of contrasts allowed, and in the frequency of syllables with certain structures. In syllabification, the preference for onsets over codas is known as the principle of onset maximization, which states that in a VCCV sequence, if the CC is a possible syllable onset cluster in the language (as judged by its legality in word initial position), the string will be syllabified as V.CCV, not VC.CV or VCC.V (Blevins 1995). The same applies to VCV and VCCCV sequences. The syllabification with the onset maximized is usually supported by evidence from stress rules which are conditioned by syllable weight, alternations conditioned by coda vs. onset position, etc. This is not a universal rule: there are quite a few cases, discussed by Blevins (1995), in which a sequence is syllabified as VC.CV even though the CC is a possible onset cluster of the language. However, the pattern is extremely common.

The second argument for the favored status of onsets rather than codas, number of contrasts allowed, was demonstrated in (1) above. It is extremely common for languages to have more distinctions allowed in onset position than in coda position (Blevins 1995). Japanese, for example, allows many more phonemes to appear in the onset of a syllable than in the coda (only two), and does not allow place (or voicing) distinctions in coda position at all, although three places of articulation are distinguished in onset position. With regard to preference for certain syllable structures, all languages allow CV syllables, but some do not allow VC or CVC syllables. There are also many phonological processes which appear to avoid creating onsetless syllables (all of these generalizations are documented in Blevins 1995<sup>25</sup>). Thus, there are several well known and highly cross-

---

<sup>25</sup> The well known facts about syllable structure mentioned in this paragraph have, of course, appeared in a great many publications. Blevins (1995) is cited because she provides a useful summary of all these points.

linguistic phenomena which show a special status for consonants in onset position rather than in coda position.

Steriade (1997) investigates the neutralization of voicing distinctions in many languages, a pattern which is often analyzed as neutralization in coda position, with voicing distinctions maintained in onset position. She lists the perceptual cues to voicing distinctions, and evaluates a variety of environments for which of the potential perceptual cues are available in each environment. She ranks the possible environments for obstruents, for example "before another obstruent," "after a sonorant and at the end of the word," "between two sonorants," etc., for how many of the perceptual cues to voicing are present in each environment. She then shows that there is an implicational hierarchy based on the availability of perceptual cues: if a language has a voicing distinction in a certain environment, it also has a voicing distinction in all environments with more cues available. If it neutralizes its voicing distinction in a certain environment, it also neutralizes it in all environments with fewer cues available. She further shows that although many of the voicing neutralizations she discusses have been described as neutralization in coda position, some of them in fact cannot be adequately described by reference to syllable structure, and require reference to availability of perceptual cues. Thus, she shows that perceptual factors are the cause of some patterns which have usually been analyzed as effects of syllable structure.

Stevens discusses in several publications (Stevens 1971, 1980, 1985, Stevens and Blumstein 1981, Stevens and Keyser 1989) the importance of fast changes in the speech signal, particularly those taking less than approximately 30 milliseconds<sup>26</sup>, as regions of the signal which carry a large amount of information. He suggests that some distinctions can be perceived during these short time windows of rapid spectral change, and that these are the distinctions which languages prefer to use to distinguish their consonant inventories.

---

<sup>26</sup> The exact maximum duration of such cues is not known. Stevens (1971) defines the rapid spectral changes as those taking less than 50 ms, while in other work (1985) he emphasizes regions of 10-30 ms.

These distinctions are the ones most commonly found in the world's languages. Furthermore, a language which does not use very many distinctive features for its consonants (and therefore has a small consonant inventory) is likely to use only the distinctions which can be perceived over a short time window. Presumably, only languages with larger numbers of distinctions would use the ones which take longer to perceive, and they would also have the more quickly perceived distinctions in their systems as well. These ideas are discussed further in Lang and Ohala (1996). An example is the fact that even languages with small inventories use the feature [ $\pm$ continuant] and distinguish stops from some other manner of articulation. Continuancy versus the lack of it can be perceived over a very short time window. Secondary articulations, such as distinctive palatalization, are likely to take longer to perceive, and these are usually only found in languages with large consonant inventories (Stevens and Keyser 1989, Stevens et al. 1986, and Lang and Ohala 1996).

Stevens discusses the perceptual saliency of various distinctions, but does not discuss how the environment a segment is in might influence its perceptual saliency, since he applies these ideas to entire distinctions regardless of environment<sup>27</sup>. Considering Steriade's work on the differential perceptual saliency of the voicing distinction in various environments, it seems likely that other distinctions would also be more salient in some environments than in others. Perhaps some segments can be perceived during a very short time window (an important factor for perceptual saliency according to Stevens) in some environments, but not in others. The results of my experiment show that at least English listeners perceive word initial stops over a significantly shorter time window than either stops in coda position or stops which are postvocalic but in onset position, as discussed in Section 4.6. When listeners perceive a stop based on its CV cues (in this experiment, where it is word initial), the entire improvement in perception, even reaching 100% correct,

---

<sup>27</sup> He does discuss how combinations of features can enhance or detract from the saliency of a feature which is overall relatively salient, however (Stevens and Keyser 1989).

often happens between the first and second gates. This is within less than 40 ms after the release of the stop. A postvocalic stop, especially one after a long vowel, often shows a very gradual improvement in perception of the stop occurring through a large portion of the vowel. Thus, it appears that stops can be perceived from their CV cues over a much shorter time window than when they are perceived from their VC cues.

CV cues are considered to be stronger than VC cues for most distinctions<sup>28</sup> for other reasons as well, such as the presence of the burst at release into a vowel. Greenberg (1998) finds less reduction of syllable onsets than of either nuclei or codas in connected, natural speech. The results found by Fujimura et al. (1978) also confirm perceptually that listeners rely more heavily on CV cues than VC ones<sup>29</sup>. If a stop is in onset position, it will have the stronger and more rapid CV cues available, unless it is part of an onset cluster such as /tr-/ and is not the final member of that cluster. Even if a stop is a non-final member of an onset cluster, in English it will have to be released into some sonorant segment, and will have most of the same cues available as in a CV sequence. If a stop is in coda position, it has only the weaker and slower VC cues. Furthermore, if Stevens is correct that cues which occur over a short time window are perceptually more helpful, the faster perception of stops from CV cues than from VC cues found in this experiment adds an additional reason for considering CV cues to be more important.

Both the shorter time window of CV cues and the other factors which make them stronger than VC cues provide a perceptual motivation for the phonological universal of favored status for onsets over codas. If consonants are in onset position, they will have the

---

<sup>28</sup> However, Steriade (1997) points out that some distinctions, such as preaspiration or retroflexion, have strong cues near the closure of the consonant rather than the release.

<sup>29</sup> Fujimura et al. (1978) also played the spliced VCV syllables to listeners backwards, with the bursts removed, so that the CV and VC cues in the signal were reversed. Listeners still relied on the transition out of the vowel more heavily than the transition into it, even though the transition out of the vowel had been produced as a VC, not a CV sequence. One could conclude that there are no physical differences between CV and VC cues. However, it could also be that naturally produced CV cues are stronger/faster than VC cues, and listeners therefore learn to focus on the transition out of a vowel, and apply this learned strategy even under experimental manipulations of the cues. In either case, it is clear that listeners do focus on the cues in a CV transition rather than those in a VC transition. Also, stop bursts should provide rapid cues in a CV transition, and these were removed from the stimuli Fujimura et al. used.

faster and stronger cues available for them, but if in coda position, they will not. These results further our understanding of the perceptual motivations underlying patterns of syllable structure.<sup>30</sup> This suggestion of a perceptual foundation for the common patterns of syllable structure is not intended to mean that there cannot be other reasons for phonological patterns involving syllable structure as well. It is only meant to show that perceptual factors contribute to the abstract phonological patterns which are analyzed through syllable structure.

### 5.6.2. Alternations between glides and high vowel nuclei

It is fairly common cross-linguistically to find alternations between glides and the high vowels most similar to them, particularly between [j] and [i] or between [w] and [u]. While this is not a phonological universal, it is a common pattern, as discussed by Blevins (1995) with regard to syllabification and by Hayes (1989) with regard to compensatory lengthening. The analysis of such alternations usually involves linking a glide to a mora or delinking a high vowel from a mora.

Poser (1988) discusses such a case in Japanese, involving the irregular non-past form of the verb /iw-/ 'to say,' shown in (2).

(2)	'say'	'meet'	
	iw-	aw-	underlying stem
	iwanai	awanai	negative non-past
	itte	atte	TE form
	iinagara	ainagara	'while V-ing'
	ieba	aeba	conditional
	ioo	ao	hortative
	<b>yuu</b>	au	positive non-past

Here, the negative form of the verb, as well as the fact that the rest of the paradigm matches the forms of other stems ending in /w/, make it clear that the underlying stem of the verb is

---

<sup>30</sup> Greenberg (1996a) also discusses the importance of onsets in speech perception, but he is primarily referring to the importance of the onset of individual sounds, not to onsets as a syllable position. This is, however, related to this issue, in that the auditory motivation for Stevens' suggestion that the faster cues are the more important lies in the sensitivity of the auditory system to rapid changes in the signal, as discussed in Section 5.1.2 above.

*/iw-/*. Historically, */w/* disappeared except before */a/*, so the deletion of the */w/* in all forms but the negative is a regular aspect of this paradigm. The non-past form is irregular, though: instead of */i-u/* (which is how it is written in the syllabary), it surfaces as */yuu/*. Poser (1988) analyzes this form through a glide formation rule which changes the */i/* into */y/*, and subsequent compensatory lengthening. He provides a few additional examples of */i/* alternating with */y/* in Japanese, although none of them occurs in more than a few words.

The results of my experiment show some confusions which mirror typical glide-high vowel alternations. For the stimulus "eon" */ia/*, there were numerous responses with initial palatal glides, as shown in Table 5.5, along with responses with the correct vowel sequence (i.e. "eon, eons") and nearly correct VV sequences (e.g. "Ian").

Table 5.5. Responses to the stimulus "eon" with initial palatal glides, and number of subjects giving that response at each gate.

Response	Number of subjects giving this response at each gate number										
	1	2	3	4	5	6	7	8	9	10	11
yes	1	1	1								
you	1	3			2						
year		1									
yeast			2		1						
Yiddish			1								
yellow			1				1			1	1
young			1				1	1	4	3	3
yearn			1								
yield				1	2	1	2		1		
yarn							1				
yummy							1				
yelp								1			
yell								1			
yum								1			
yeah										1	
yawn								1			2

There are a variety of responses with an initial palatal glide instead of a two syllable VV sequence, and even at the final gates of the word, approximately half the subjects give responses with a glide. The last gate for this word is half way through the vowel */a/*, and yet even at that point, subjects do not seem to be sure whether the initial segment was a glide or a vowel. The subjects have no difficulty in perceiving the quality of the initial

segment: from the third gate onward, all responses which do not begin with /j/ begin with the vowel /i/. Even at the second gate, three responses begin with /i/, although the first gate is shortly after the beginning of the word, so the first and second gates do not allow listeners to hear very much of the signal. Thus, the high front quality of the segment becomes clear during approximately the initial 50 ms of the signal, but the status of the segment as a glide or a syllable nucleus remains unclear even into the following segment.

The two words in the Japanese experiment with medial glides as the target segment showed similar effects. The word /huyoo/ [ɸujoo] 'unnecessary' had especially interesting responses: in addition to responses which had, as the speaker intended, the syllable nucleus /u/ and the glide /y/, listeners also gave the responses in Table 5.6.

Table 5.6. Responses to the stimulus /huyoo/ [ɸujoo] 'unnecessary' with syllable structures other than the intended one.

Response	# subjects per gate			
	1	2	3	4
/fireNtse/ 'Florence'	1			
/figiasukeeto/ 'figure skating'		1		
/fizi/ 'Fiji'			1	
/fittonesu/ 'fitness'			1	
/hikkosi/ 'moving'		1		
/hinomaru/ 'the sun flag'		1		
/huiiti/ 'not in agreement'		1	2	
/hui/ 'unexpected'		2	2	1
/huiuti/ 'surprise attack'			1	

For this word, there were also numerous responses of /huyu/ 'winter' and /huyoo/, with the intended syllable structure. The responses shown in Table 5.6 are quite interesting, though. This stimulus had three types of responses with syllable structures other than the intended one. The first group of responses listed in the table are all borrowings in which the source word begins with [f] and a high front vowel. In Japanese phonology, outside of recent loanwords, the only vowel which can follow the labial fricative is /u/, as the labial fricative is an allophone of /h/ which appears before /u/. Recent borrowings from source words with [fi] or [fi] are usually borrowed as [ɸ<sup>w</sup>i]. It is somewhat unclear what

phonemes this sequence, which is restricted to recent borrowings, represents. I have chosen to transcribe these words phonemically as /fi/. Some speakers, especially younger speakers who speak English well, pronounce this as [ϕi], but many speakers still pronounce these words with a labial offglide to the fricative. I believe that the subjects who gave these loanwords as responses perceived the initial labial fricative, perceived the intended vowel /u/ as a glide, and perceived the intended glide /y/ as the nucleus of the syllable. In order to make the /u/ into a glide following another consonant, they had to resort to this special class of recent borrowings. No responses with initial /fi/ appear anywhere else in the experiment.

The meaning of the second group of responses listed in the table is less certain, since there are only two of this type of response, but they may also represent a misperception of glides and high vowels. In Japanese, /h/ is realized as [ç] before /i/, so the responses /hikkosi, hinomaru/ begin with a palatal fricative. It may be that the listeners who gave these responses perceived the frication, but not its place, and perceived the intended glide /y/ as the vowel of the first syllable, missing the intended /u/ entirely. By making /i/ the vowel, they chose the palatal fricative. The final group of responses, /hui, huiuti/, represent a large number of subjects. The fricative here is the bilabial allophone of /h/, [ϕ], conditioned by the vowel /u/. These listeners seem to have perceived the quality of both the intended vowel /u/ and the intended glide /i/, but made both into syllable nuclei<sup>31</sup>, adding an extra syllable to the beginning of the word. As in "eon," responses of this type continue into the last gate of the word, which in this case is late in the glide /y/. Again, after the initial gate, listeners have little difficulty perceiving the quality of both the /u/ and the /y/, but there is considerable confusion as to which is a glide and which is a syllable nucleus throughout the gated area.

---

<sup>31</sup> Neither /u/ nor /i/ is a possible diphthong of Japanese.



The word /mawari/ [mawari] 'surroundings' shows similar results. Many subjects gave responses with the correct /aw/, such as /mawari/, /mawaru/ 'to turn around (intrans.)', /mawasu/ 'to turn around (trans.)', /mawarimiti/ 'detour,' /mawasi/ 'Sumo wrestler's belt.' However, there were several responses in which they perceived the quality of the /w/, but did not perceive it as a glide, shown in Table 5.7. There were also many responses with an unrelated consonant following the /a/, in which listeners seem not to have perceived the quality of the /w/ at all.

Table 5.7. Responses to the stimulus /mawari/ [mawari] 'surroundings' with syllable structures other than the intended one.

Response	# subjects per gate				
	1	2	3	4	5
/mausiro/ 'directly behind'			1	1	2
/mausupiisu/ 'mouthpiece'			1		1
/mausu/ 'mouse'				1	1
/maoo/ 'Satan'				1	
/maoki/ gloss uncertain <sup>32</sup>					1

In these responses, subjects perceived the quality of the /u/, but treated it as a second vowel<sup>33</sup>. A historical change in Japanese made all tautomorphic sequences of /au/ into /oo/, so in order to give responses with /a/ followed by a vowel of approximately the quality of /u/, subjects had to use either a response with a morpheme boundary between the two vowels (/ma-usiro/), recent borrowings (the second and third responses), or use the sequence /ao/ instead. In this word, unlike the others discussed here, there are no such responses at the early gates, although there are a considerable number at the later gates. There were also very few responses with the correct /maw-/ at the first two gates. Listeners seem not to have perceived the quality of the /w/ until approximately the third

<sup>32</sup> A native speaker of Japanese could not identify this word, and it does not appear in dictionaries.

<sup>33</sup> /au/ is not usually considered to be a diphthong in Japanese, because it cannot occur except across a morpheme boundary in native words. However, Vance (1987) says it may be a diphthong in recent borrowings.

gate, and not to have perceived it as a glide very well even at the last gate, which is near the end of the /w/.

In the English word "eon," one might think that the large number of responses with palatal glides stems from the low frequency of "eon," the status of "Ian" as a proper name, and the lack of other familiar words beginning with /ia/ or similar vowels. However, for both the Japanese words discussed here, there are responses with the intended syllable structure which are of much higher frequency than the responses shown in the tables here. For /huyoo/, which is itself relatively frequent, /huyu/ 'winter' would be an alternative with very high frequency. /mawari/ is also of much higher frequency than most of the responses in Table 5.7. Therefore, the responses with confusions between glides and vowels (as syllable nuclei) are not an artifact of low frequency target words.

These results show that although listeners usually have little difficulty perceiving the high front or high back quality of both glides and vowels, they do have difficulty in perceiving whether that quality is a syllable nucleus or a glide, that is, whether the segment is moraic or not. The three words discussed here are all of the cases in the experiment which involve high vowels or glides in an environment where the phonotactics of the rest of the stimulus allows for this sort of confusion about syllable structure. For example, the /i/ of the /iæ/ sequence in "react" cannot be confused with a palatal glide because /rj/ is not a legal onset in English<sup>34</sup>.

The fact that the confusion between glides and high vowels in these words continues to the end of the gated area in each case is particularly important. Returning to Stevens' (1980) claim that distinctions which can be perceived over a short time window are easier for listeners to perceive and are therefore the most likely to be used in languages, it is clear that the distinction between a glide and a high vowel, when the phonotactics

---

<sup>34</sup> In the Japanese word /siatu/ [ʃiatsu] 'acupressure,' there is no way to make the /i/ into an offglide, but there are several responses beginning with /siya-/ [ʃija] and many beginning with /sya-/ [ʃa], even at the last gate. This may be another case of the same phenomenon.

allows for a contrast between the two, cannot be perceived during a short window. Rather, listeners' confusion between the two is far-reaching. The contrast between glides and high vowels is not rare cross-linguistically, as one might think that Stevens' suggestions would predict. Even Japanese, with its relatively small consonant and vowel inventories, distinguishes between vowels and glides in some environments, as shown by the responses above. However, glides and high vowels do alternate with each other in many languages, and I believe this is the result of the long lasting perceptual confusion we see in these responses. When a language does use a distinction which listeners need a long time window in order to perceive, the long lasting confusion between the segments may result in alternations between them. The long duration of the window needed to perceive the glide/high vowel distinction provides a perceptual motivation for the cross-linguistically common alternation between them.

### 5.6.3. Dissimilation

Ohala (1981b, 1986, 1989) proposes an explanation for dissimilation based on listeners' misperceptions of the signal. He points out that the types of segments which, cross-linguistically, tend to dissimilate, are those segments for which perceptual cues spread a long way from the segment. For example, dissimilations involving retroflex sounds are well known, as in the historical development of /fɛbjuəri/ (instead of /-bru-/, as the orthography would indicate) and "pilgrim" from Latin "peregrinus," and the modern nonstandard English pronunciation /lɑːbəri/ for "library." Retroflexion affects the signal at a great distance from the /r/ itself. Secondary articulations such as labialization also dissimilate relatively often, and the lowered second formant caused by labialization can spread over a long duration. Distinctions which have only cues which are highly localized, such as the manner of articulation 'stop,' are unlikely to dissimilate.

Ohala argues that when a speaker intends to produce a word which contains two occurrences of a feature with far reaching cues, the listener may misparse those cues, attributing them entirely to one occurrence of the feature or segment. Ohala terms this perceptual "hypercorrection." The listener then concludes that there is only one occurrence of the feature in the word. If the listener adopts this misperception as her own representation for the word, and then produces that form, this is the beginning of a potential dissimilatory sound change. When a word contains two occurrences of a feature which does not have far reaching cues, for example two stops, the listener will have no difficulty attributing the localized cues to the two segments for which they were intended, and no dissimilation will occur.

There is considerable similarity between the far reaching cues important for Ohala's explanation of which features dissimilate and the distinctions which do not fit Stevens' description of distinctions perceptible over a short time window. One would expect those contrasts which are not perceptually salient according to Stevens' short time window definition of salient features to be exactly the ones whose cues spread far from the segment, and are likely to dissimilate. Therefore, in the current experiment, features which often dissimilate should be perceived far from the  $D_{\max}$  point, if that is at the onset of the segment, and should be perceived over a long time window (should have a slow rise time).

To test this idea, I examined the rise time and the likelihood of being recognized far from  $D_{\max}$  for two types of segments, postvocalic /r/ and /l/ in English and post-consonantal /y/ in Japanese. /r/ and /l/ are frequently subject to dissimilation<sup>35</sup>, both historically, as in the well known Latin example, and synchronically, as in modern English dialectal variation, the "library" example mentioned above. (Ohala (1981b, 1986) gives several more examples of dissimilation of retroflexion or "r" sounds in Sanskrit and Provençal.)

---

<sup>35</sup> Ohala (1981b) lists laterality, as well as nasalization, as features which may not be likely to dissimilate. He argues that while both of these features do have some cues which spread far into surrounding vowels, their major cues are spectral discontinuities, which cannot spread. A more complete study of this topic would involve choosing which features to test for perception over a long time window carefully.

In Japanese, however, the flap /ɾ/ does not have such far reaching cues as the English retroflex /ɻ/ does, and there is no lateral. However, Japanese /y/ after another consonant, such as /ky/, /hy/, etc., is often analyzed as distinctive palatalization rather than as a separate glide segment. Secondary articulations such as palatalization are a possible target of dissimilation, and are a primary example of a non-salient feature in Stevens' definition. Therefore, both of these types of segments are predicted to be recognized far from  $D_{\max}$  and to have a slow rise time for the %Corr measure.

For the Japanese post-consonantal glides, I included all words with Cy as the transition of interest, including the word /koNyaku/ [koɲjaku] 'engagement,' where the palatal glide follows the mora nasal and would never be analyzed as a secondary articulation. It nevertheless has the potential for its cues to spread. The average rise time of the %Corr for the palatal glide in these words is 2.6 gates. As was explained in Section 4.3.1.4, the palatal glide in these transitions is often recognized before the  $D_{\max}$  point, because its cues can spread into the release, aspiration, or frication noise of the preceding segment. It is not clear to what group of words one should compare the average rise time of segments which are predicted to dissimilate: many types of segments have a relatively long rise time for other reasons. Postvocalic stops have a long rise time, probably because cues for their place of articulation spread into the preceding vowel and the duration of the vowel provides a cue to their voicing. The feature of stops which is predicted not to spread, and thus not to cause dissimilation, is their "stoppedness" (Ohala 1981b), however, not their place. The rise time measure reflects listeners' ability to perceive all the features of the segment, not just the ones we may be interested in.

However, one can safely compare the rise time of the segments which are predicted to dissimilate, and therefore to have a slow rise time, to the rise time of the word initial stops discussed in Section 4.6, since these have already been shown to be perceived based on cues in a short time window and to have short rise times. The average rise time for the Japanese word initial stops is 1.5 gates. Since only a few words in the experiment have a

rise time of more than 3 gates for their segment of interest, the difference between the average of 1.5 gates for the word initial stops and 2.6 gates for the palatal glides is relatively large. Because there are so few Japanese words in the experiment with word initial stops, and because this analysis is quite preliminary, I have not performed any statistical tests for this comparison.

The results for the English words with postvocalic /l, r/ are less clear. There are only two such words, "fair" and "elevator." They have rise times of 3 gates and 1 gate, respectively. The average rise time for the word initial stops in English is 1.62. Clearly, far more words need to be investigated to make this point clear, but the perceptual improvement for "fair" and "elevator" does not seem to take place very slowly. However, both words are recognized early relative to their  $D_{\max}$  points, and this was attributed to the spreading of their cues into the preceding vowel.

The application of the results of this experiment to dissimilation is preliminary, and the results for English are unclear. However, among the distinctions Japanese makes, palatalization (whether it is phonemically a secondary articulation or a post-consonantal glide) is one of the ones universally most often subject to dissimilation. The most progress toward recognizing the palatal glide in post-consonantal position in Japanese is often early relative to the  $D_{\max}$  point, and these segments also have a relatively long rise time. This is the behavior expected for the types of segments prone to dissimilation. Therefore, the results of this gating experiment for Japanese tentatively help to confirm the perceptual motivation Ohala proposes for dissimilation.

### 5.7. Integrating speech perception and formal phonology

Formal theories of phonology have concentrated overwhelmingly on the production of speech, almost to the exclusion of its perception. All of the major recent theoretical approaches to phonology, from early generative phonology to current work in Optimality Theory, are theories of how to convert the underlying representation to the surface form, or

an approximation of it<sup>36</sup>. These theories do not address how a listener, who hears a surface form, knows which underlying representation the surface form belongs to. The task in spoken word recognition is to work back from a surface form to determine which lexical entry, which underlying representation, that surface form represents. Some aspects of spoken word recognition, such as normalization for speaker specific variation, probably do not require the listener to use knowledge about the phonology of the language (although Johnson's (1997) exemplar based model calls the separation of speaker normalization from other aspects of spoken word recognition into question). However, other aspects of matching a surface form to a lexical entry do require the listener to have, and use, knowledge about the phonology of the language. How does an English listener know that a voiceless unaspirated [t] in one environment is to be associated with the underlying form /t/ and in another with /d/? This, and much more complicated knowledge about the grammar of the language, is required. If working from the surface form back to the lexical entry requires the listener to use the grammar of the language, then formal theories of phonology should have some way to model this.

Very few researchers have attempted to do that, but some have attempted to bring formal phonology and speech perception together in other ways. Lahiri and Marslen-Wilson (1991, 1992) use formal theories of phonology to shape theories of speech perception. They propose that speech is perceived by comparing it to a radically underspecified underlying representation (UR), not by comparing it to a surface representation. Steriade (1997) takes a different approach, using experimental findings about speech perception to inform formal theories of phonology. She does this by arguing that the relative availability of perceptual cues in different environments explains the

---

<sup>36</sup> Some recent work in OT, notably that by Flemming (1995), advocates doing away with underlying representations entirely, and evaluating the well-formedness of a surface form by comparing it to all the other surface forms of the same morpheme, not by comparing it to any unique underlying form of the morpheme. This approach, while fascinating for its implications of what speakers know, is still a model of production, not a model of perception. It models what forms are acceptable for a speaker of a language to produce, not how a listener understands what morpheme a speaker intended.

conditions for some phonological patterns better than models of syllable structure do. Specifically, she shows that the environment for some voicing neutralizations must make reference to availability of perceptual cues, and cannot be described in terms of syllable coda position, as was discussed in Section 5.6.1 above. Flemming (1995) takes a similar approach to formal modeling of phonological universals of phoneme inventories. Finally, Smolensky (1996) does attempt to model young children's speech perception in an OT framework, but his example is quite preliminary, and this method will not be discussed here.

#### 5.7.1. Perception by comparison to the UR

Lahiri and Marslen-Wilson (1991, 1992) argue that because a given lexical entry can have many different surface forms depending on its environment and on speech rate, listeners could not possibly perform spoken word recognition by comparing the signal they hear to a surface form, since they could not know which of the many surface forms to use. Instead, they propose that listeners extract distinctive features from the signal, and compare these distinctive features to the underlying representations of lexical entries. (From that state, presumably, the cohort model of spoken word recognition takes over in explaining how listeners identify the correct lexical entry.) Furthermore, they claim that the underlying representation listeners use for this purpose is a radically underspecified one. They believe this means that listeners can make use of non-distinctive phonetic cues only if they are for marked features, not if they are for unmarked features. Jongman et al. (1992) also argue for the use of underlying phonological representations in speech perception, although with somewhat different implications.

Lahiri and Marslen-Wilson (1991, 1992) present results from a gating study of English and Bengali on perception of distinctively and non-distinctively nasalized vowels in the two languages to support this proposal, as was discussed briefly in Section 1.4.2. Focusing on the English portion of their experiment, their claim is that since English does



not have distinctively nasalized vowels, and [+nasal] is the marked value for consonants, English listeners can use the non-distinctive cue of vowel nasalization to perceive that a following consonant is nasal, since that is the marked value of the feature. However, they predict that English listeners cannot use lack of nasalization on a vowel to rule out a following nasal consonant as a possibility, since [-nasal] is the unmarked value of the feature. They present experimental results to support this hypothesis, but their conclusions rest largely on the interpretation of the minority responses, given by only five to fifteen percent of the subjects, and they do not present statistical tests for most of these results. Ohala and Ohala (1995) replicate this experiment and re-analyze both the predictions and conclusions at length, so I will not discuss the details of this study here. (Ohala (1992) also discusses these conclusions.)

The results from my experiment with regard to perception of the voicing of postvocalic obstruents, and listeners' use of that information in spoken word recognition, have implications for Lahiri and Marslen-Wilson's proposal about radical underspecification. I analyzed the responses listeners in my experiment gave to stimuli with postvocalic obstruents to see how accurately listeners' responses at various gates matched the voicing of the postvocalic obstruent of the stimulus. This is certainly not the first test of listeners' ability to perceive the voicing of a postvocalic obstruent based on cues occurring during the vowel (see Raphael 1972, 1981, Mermelstein 1978, Walsh and Parker 1983, Nearey 1997), but the methods used in my experiment are very similar to those used by Lahiri and Marslen-Wilson to test their theory of perception by comparison to radically underspecified UR for vowel nasalization. Therefore, the evaluation of perception of postvocalic obstruent voicing in my experiment provides a good test of Lahiri and Marslen-Wilson's proposal on a different distinctive feature from the one they tested.

In English, vowels are considerably longer before a voiced obstruent than before a voiceless obstruent (Peterson and Lehiste 1960, Chen 1970). (Many languages have this durational difference, and Kluender et al. 1988 suggest an auditory motivation for it.)

Lahiri and Marslen-Wilson's proposal that listeners can only use non-distinctive cues to perceive a marked feature, not to rule out an unmarked feature, means that English listeners should be able to use the non-distinctive cue of a lengthened vowel to perceive a following voiced obstruent, but they should not be able to use a short vowel to perceive that the following obstruent is voiceless. This is because this vowel length difference is not distinctive in English, and voicing is considered marked for obstruents (Chomsky and Halle 1968:406). (The vowel length difference here is the difference in duration between the vowels of "bad" and "bat," for example, not the contrastive difference between /i/ and /ɪ/, although that is often also referred to as a vowel length difference in English. While the vowel duration difference in "bad" and "bat" may serve as one of the more important perceptual cues to the identity of the following stop, few speakers of standard dialects of American English would feel that these are two different vowels, so the durational difference will be considered a non-distinctive cue to the distinction in the following consonants, whether that distinction involves actual vibration of the vocal cords or not<sup>37</sup>.)

Lahiri and Marslen-Wilson (1991) are quite clear about this prediction, although they do not discuss postvocalic obstruent voicing specifically. They state that they are assuming radical underspecification, in which "the feature array for a given segment will not contain a specification for any feature, distinctive or not, that has the unmarked value. Consequently, the only specifications in the underlying representation, on this account, are those for features which are (a) distinctive, and (b) have the marked (or non-default) value" (1991:253). There can be no doubt that they would posit underlying representations of English postvocalic obstruents in which only [+voice], and not [-voice], is specified. They must therefore predict that listeners cannot use a short vowel to rule out a following voiced obstruent, and that when listeners hear a short vowel, they will choose randomly (within

---

<sup>37</sup> In some dialects, such as some New York dialects, the quality of these vowels may be quite different. The statements about perceptual cues and phonological status of these words may not apply to all dialects of English.

constraints of the cohort) between voiced and voiceless following obstruents, or perhaps choose a response with no obstruent at all. It is important to note that Lahiri and Marslen-Wilson do not predict that listeners, in the absence of cues to the marked value of a feature, will choose the default unmarked value. Rather, they predict that in the absence of cues to the marked value of a feature, listeners will not be able to make any judgment about that feature, and will fail to rule out the marked value based on the lack of cues for it.

I have not presented an analysis of perception of each feature of each target segment over time for my entire experiment, but I did analyze the perception of voicing in postvocalic obstruents in order to address this question. I evaluated the responses to these words for the voicing of the postvocalic obstruent of the response at three stages: the first gate of the word, which is located near the middle of the pre-obstruent vowel, the last gate before the obstruent begins, and the final gate, which is located in the middle of a fricative or just after the burst of a stop. I will present the investigation of the English words first. I excluded some words with postvocalic obstruents for several reasons: "citizen, committee, muddy" (for their /t/ and /d/ phonemes) were excluded because their phonemic postvocalic stops are realized as flaps. The primary cue to voicing of postvocalic obstruents is assumed to be duration of the preceding vowel, and dialects of American English vary as to whether there is any difference in vowel duration before flaps derived from /t/ and /d/ (cf. "latter, ladder" and "writer, rider"). Since many listeners, and possibly the speaker, would not have any durational difference, these words do not allow for a test of the hypothesis.

I further excluded "fitness" (for the /t/) and "induction" (for the /k/) because there was only one gate before the closure of the stop in these words. In order to compare the voicing judgments earlier in the vowel to those immediately before the onset of the obstruent, there must be at least two gates ending before the onset of the obstruent. I also excluded "soybean" (/b/) because its cohort places extreme limits on the responses listeners can give ("soy, soil, soybean(s), soymilk, soysauce") and is likely to affect their choice of postvocalic segment more than for other words. Finally, I excluded the two words with

transitions into /f/, "leaf, relief," because many listeners perceived these as "leave, relieve" even at the end of the /f/. The speaker's production of these words may have been idiosyncratic. The remaining eight words with postvocalic voiceless obstruents and three words with postvocalic voiced stops were included in the test.

An example of the responses given at the three gates which were examined, and the evaluation of the percent of responses with a voiced or voiceless obstruent, appears in (3). The number of subjects giving each response is shown in parentheses after the response.

(3) Responses to:	"latches" /ætʃ/ #	"fade" /eɪd/ #
at first gate (middle of æ, early in eɪ)	lap (3)	face (7)
	Latin (2)	fate (3)
	laugh (1)	fake (1)
	laughter (1)	
	lack (1)	
	lactose (1)	
	loud (1)	
	box (1)	
	<u>91% voiceless</u>	<u>100% voiceless</u>
at last gate before closure	laugh (4)	fade (8)
	last (2)	phase/faze (2)
	lap (1)	faith (1)
	lapse (1)	
	blast (1)	
	glasses (1)	
	box (1)	
		<u>100% voiceless</u>
at last gate (just after stop burst, or during frication noise)	latch (4)	fade (11)
	latched (1)	
	latchkey (1)	
	match (1)	
	lattissimus dorsi (1)	
	Latvia (1)	
	Lattimer (1)	
	latter (1)	
	<u>82% voiceless<sup>38</sup></u>	<u>100% voiced</u>

<sup>38</sup> "Lattimer, latter" are not counted as voiceless responses because their /t/ is realized as a flap. Listeners may have recognized the postvocalic obstruent as voiceless, identified it as the phoneme /t/, then given responses in which /t/ is not realized as voiceless, but I am not willing to assume that based on 11 listeners. If a large number of listeners heard this stimulus, and many gave responses with /t/ realized as a flap, but none gave responses with /d/ realized as a flap, then one could be sure these responses should be counted as voiceless.

For the word "latch," the voicing of the voiceless postvocalic obstruent is perceived correctly by most listeners at all three time points. The voiced obstruent in "fade," however, is perceived as voiceless by all listeners early in the vowel, but most listeners perceive its voicing correctly by the last gate before the closure of the /d/. All listeners perceive it as voiced once they have heard the /d/.

The average percentages for all the English words included in this comparison appear in Table 5.8.

Table 5.8. Average percentage of listeners giving responses with postvocalic voiceless obstruents, voiced obstruents, or no obstruent to English stimuli with postvocalic voiceless or voiced obstruents at the first gate (in the middle of the preceding vowel), the last gate before the onset of the obstruent, and the final gate (after the burst of a stop or halfway through the frication of a fricative or affricate).

#### VOICELESS STIMULI

Gate	Voiceless responses	Voiced responses	No obstruent responses
First	0.72	0.06	0.23
Last pre-obstruent	0.89	0.06	0.06
Final	0.97	0.00	0.03

#### VOICED STIMULI

Gate	Voiceless responses	Voiced responses	No obstruent responses
First	0.79	0.21	0.00
Last pre-obstruent	0.27	0.73	0.00
Final	0.00	1.00	0.00

Overall, listeners tended to give responses with a postvocalic voiceless obstruent when the vowel was relatively short regardless of what the stimulus was. That is, for the voiceless stimuli at both pre-obstruent gates and for the voiced stimuli at the first gate, they gave a majority of voiceless responses. However, when the vowel was longer, as in the voiced stimuli at the last gate before the obstruent, they switched to a majority of responses with voiced obstruents. The shift for the voiced stimuli from 79% voiceless responses when the vowel is cut off early to 73% voiced responses when it has approximately its natural duration is striking. At the final gate, by which point listeners had heard part of the obstruent itself, both types of obstruents were perceived rather accurately.

I tested the difference between the percentage of listeners giving voiced responses at the last pre-obstruent gate for the voiceless and voiced stimuli using an ANOVA with weighted means to correct for the unequal number of words in the two groups. The difference for the two types of stimuli was significant ( $F(1,9)=22.8, p<.005$ ). This shows that English listeners are able to use the cue of preceding vowel duration (or whatever other cues might be available during the vowel, all of which would be considered non-distinctive) to perceive the voicing of both voiced and voiceless obstruents, even though [-voice] is the unmarked value for obstruents. This result conflicts with Lahiri and Marslen-Wilson's proposal.

I also investigated the perception of voicing for the Japanese words with postvocalic obstruents. The voiceless obstruents had approximately the same results as in English, with a large percentage of listeners giving voiceless responses at all three points in time. However, the voiced obstruents did not show the shift which was present in the English data between the first gate and the last pre-obstruent gate. One Japanese word, /*kanada*/ 'Canada,' did have a higher percentage of voiced responses at the last gate before the /*d*/ closure than earlier in the vowel, but the increase was relatively slight. There were a majority of voiced responses for this word at all three time points, probably because the corresponding word with a voiceless obstruent, /*kanata*/ 'yonder,' is somewhat archaic and of low frequency. The other Japanese words with postvocalic voiced obstruents showed no increase in the percentage of voiced responses between the first gate and the last gate before the obstruent at all.

I suspect the lack of this effect in Japanese reflects a smaller difference in vowel duration before voiced and voiceless obstruents than is present in English. English has phonologized this difference, making it considerably larger than in many other languages (Chen 1970). Even in English, the difference in vowel duration is much larger for tautosyllabic vowel-obstruent sequences than when the obstruent is the onset of the next syllable. In Japanese, the obstruent must always be the onset of the next syllable, since

geminate were not under consideration in this comparison. In investigations of duration compensation with regard to the hypothesis of mora timing, researchers have shown that Japanese vowels are longer before voiced obstruents than before voiceless ones (Port et al. 1980, Homma 1981, Beckman 1982), but this difference is much smaller than in English. Thus, the durational difference which in English is likely to be the strongest cue during the vowel for voicing of the following obstruent is not as large in Japanese. Listeners appear to be at most marginally able to use this potential cue to distinguish obstruent voicing in Japanese.

In sum, English listeners are able to distinguish the voicing of a postvocalic obstruent based on non-distinctive cues during the preceding vowel, whether the voicing of the obstruent is marked or unmarked. This provides evidence against Lahiri and Marslen-Wilson's (1991, 1992) proposal that speech is perceived by comparing it to a radically underspecified underlying representation.

#### 5.7.2. Perception as the basis for formal constraints

As was discussed in Section 5.6.1, Steriade (1997) re-analyzes a wide variety of cases of voicing neutralization and other laryngeal neutralizations which have usually been analyzed as conditioned by syllable structure. She shows that many of these patterns can be better accounted for as neutralizing a distinction in environments which supply fewer perceptual cues for it, and maintaining the distinction where more perceptual cues are available. Thus, she finds that these neutralizations are not directly a result of syllable structure, but rather that the apparent effects of syllable structure are motivated by the relative availability of perceptual cues in particular environments. She proposes that information about which environments have sufficient perceptual cues available is incorporated in the speaker's grammar of the language. She models this in Optimality Theory through the ranking of a constraint which preserves voicing specifications (or other laryngeal specifications) and several constraints forbidding the voicing distinction in

particular environments. Individual languages differ as to the position of the "preserve [voice]" constraint relative to the particular constraints forbidding voicing, and thus neutralize voicing in different environments.

Flemming (1995) takes a similar approach to modeling the choice of contrasts in languages, for example, how many height distinctions a language makes in its vowel inventory and which vowels are used to make those distinctions. He uses constraints saying that a segment must be similar to the corresponding segment in other forms of the morpheme and groups of constraints requiring 1) that vowels differ from each other in their first formant by at least a certain amount, and 2) that at least a certain number of vowels be distinguished by their first formant values. Flemming also analyzes constraints on sequences of sounds, such as the fact that distinctive labialization is not possible after labials in many languages (the type of sequential constraint investigated by Kawasaki (1982) and discussed in Section 1.2.4). For this sort of issue, he discusses whether cues for the place of consonants are present during the transition into or out of the consonant. He uses constraints similar to those which determine number of vowel contrasts, specifying that contrasting segments or sequences must have a certain degree of difference for a particular feature, for example the second formant. /p/ and /p<sup>w</sup>/, for example, would not differ sufficiently on F2 at stop release, and a contrast between them would therefore violate that constraint. While Flemming's (1995) analyses make little use of syllable structure, they do provide a further example of modeling perceptual constraints on phonological patterns through constraints in OT.

As discussed in Section 5.6.1, the results of my experiment show that CV cues allow listeners to perceive stops over a shorter time window than VC cues do. If one follows Stevens in the view that cues which allow a distinction to be perceived quickly are better than those which require a longer time window, then it is better to have a stop in a position where it will have CV cues than in a position where it will only have VC cues. Since stops in onset position have at least CV cues, and stops in coda position have only



VC cues, this provides a perceptual motivation for the cross-linguistic preference for licensing more contrasts in onset position than in coda position and for the overall preference for CV syllables over VC syllables. This is similar to Steriade's analysis of laryngeal neutralizations and Flemming's (1995) analysis of sequential constraints in that I suggest that contrasts are allowed where more or better perceptual cues are available for them. Thus, my results provide a parallel case to the laryngeal neutralizations Steriade investigates, and thereby support her approach of using information about speech perception to inform the systems of constraints used to account for phonological patterns.

### 5.8. Overall conclusions

- Areas of perceptual and acoustic importance (high concentration of information)

In this experiment, I have tested listeners' use of dynamic cues through a large scale gating study of Japanese and English, using a much more representative variety of stimuli than previous work on dynamic cues has used. For each word, I have identified an area in time which is important for perception, the area during which listeners' perception of a segment or recognition of a word improves the most. The fact that a large proportion of the words have an area during which perception improves rapidly, while it improves less rapidly at other times, shows that information is not distributed equally throughout the signal, but rather is concentrated in certain salient parts of the signal. Using Furui's measure of degree of spectral change  $D$ , I have located points in the acoustic signal which are hypothesized to be the most perceptually salient because of the high degree of change in the signal at those points. By comparing the location of the important regions in time as determined by the perceptual measures to the location of the points hypothesized to be salient because of their acoustic characteristics, I have tested the hypothesis that listeners perceive speech primarily through rapidly changing portions of the signal.

- Most speech sounds are perceived primarily through dynamic cues

The results of the experiment show that listeners are more likely than would be expected by chance to rapidly become able to perceive a segment at the point in the signal with the most acoustic change. These results especially reflect the perceptual use of very fast changes in the signal, as those are the changes to which the measure  $D$  is the most sensitive. The segments which do not follow the hypothesis that the area of maximal improvement in perception should surround a point of maximal spectral change fall into several classes. For each of these classes, there are reasons based in what we know about perceptual cues for the specific type of segment or based on identifiable problems with the measure  $D$  which explain the failure of the class of segments to follow the hypothesis. Most of these classes of exceptions, in fact, provide indirect evidence that slower changes in the signal are also important in speech perception.

- Basis of dynamic cues in the auditory system

I relate the fact that many segments are perceived at a point of rapid change in the signal (the words with maximal perceptual change surrounding  $D_{\max}$ ) to properties of the peripheral auditory system, namely to the phenomenon of adaptation at the level of the auditory nerve fibers. Furthermore, I relate the evidence for the importance of slower changes to recent results on auditory processing at the level of the auditory cortex. Although the interpretation of the results is complex, the experiment does provide evidence that the sensitivity of the auditory system to changes rather than steady state signals, both at the peripheral level and at the auditory cortex, is reflected in the timing of listeners' perception of speech sounds.

- Spoken word recognition

The results for spoken word recognition are less direct, but the fact that listeners' progress in narrowing in on what word they hear is closely aligned in time with progress in

segment or feature perception, and that segment perception takes place at regions with dynamic cues, supports the hypothesis that dynamic regions of the signal are also the most important for spoken word recognition. The use of the open response method for the experiment also allows for the collection of other information about spoken word recognition, and I find that the results of this experiment confirm results by other researchers on aspects of spoken word recognition. I also suggest a modification to current theories of spoken word recognition based on the phenomenon of increase instead of decrease in the number of candidate words with additional acoustic information.

- Influence of phonology of the language on perception

I find that the phonological system of a language influences how listeners perceive segments of the language in several ways. The size of the phoneme inventory has some effect: listeners need to hear more of a vowel to identify it correctly if their language has many vowels than if it has only a few. More interestingly, the differences in syllable structure constraints of Japanese and English are found to influence how listeners of these two languages weight the cues in VC versus CV transitions: Japanese listeners are slower to make use of information about the place of a consonant in a VC transition than English listeners are because Japanese makes no place distinctions in coda position. This is in accord with previous findings by Fujimura et al. (1978) and Kakehi et al. (1996). I find relatively little influence of suprasegmental units on the timing of perception, except that stress may induce English listeners to make more use of VC transitional cues than they otherwise would.

- Perceptual explanations for common phonological patterns

The results of this experiment show that facts about the distribution of information in the signal, and about the timing of listeners' use of dynamic aspects of the signal, can provide perceptual motivations for well known phonological phenomena such as the

universal preference for onsets over codas, the common alternation between glides and high vowels, and perhaps dissimilation. I will briefly review the explanations proposed for each of these.

- **Dissimilation**

I show that at least one type of segment which is known to dissimilate often, the post-consonantal palatal glide, requires a long time window for listeners to perceive it. One type of distinction which does not participate in dissimilation, the manner "stop," is perceived over a very short time window. This matches Ohala's claim that segments which dissimilate are those with perceptual cues which spread over a long time in the signal.

- **Alternations between glides and high vowels**

Similarly for glide/high vowel alternations, I show that while the quality of a glide or high vowel is perceived quickly, listeners can only perceive the moraic status of the segment (whether it is a vocalic nucleus or a glide) very gradually. Where phonotactic environment allows a segment to be either a glide or a high vowel, listeners' responses are split between responses with glides and those with high vowels throughout the relevant segment or even into the following segment. I suggest that distinctions which have continuing perceptual confusion over a long time window, such as the glide/high vowel distinction, are likely to participate in common cross-linguistic alternations.

- **Cross-linguistic preference for onsets**

Finally, I address cross-linguistic patterns regarding syllable onsets and codas. I show that listeners' perception of stops from CV cues takes place over a very short time, while their perception of stops from VC cues is much more gradual. I relate this to Stevens' assertion that languages are more likely to make use of distinctions which can be perceived over a short time window, and Steriade's evidence that languages allow more

distinctions in positions where many perceptual cues are available for them. I argue that the cross-linguistic tendency to allow more contrasts in syllable onset position than in coda position results from the fact that consonants in onset position have the rapid CV cues, while consonants in coda position have only the slower VC cues.

- Importance of perceptual studies for explaining phonological patterns

I believe that experimental studies of the mechanisms of speech perception can offer us a better understanding of the motivations for some well known and cross-linguistically common phonological patterns. In this dissertation, I have tried to contribute to our knowledge of how factors in speech perception influence phonological structure through a study of the timing of flow of information, and specifically of the hypothesis that parts of the signal with rapid acoustic change are used disproportionately in perception. I offer explanations for several phonological patterns based in listeners' use of rapidly changing parts of the signal. I believe the search for perceptual motivations which underlie the effects of syllable structure has particularly strong potential for future research of this type.

## References

- Altmann, Gerry T.M. (ed.) 1990. *Cognitive models of speech processing: Psycholinguistic and computational perspectives*. Cambridge: MIT Press.
- Amano, Shigeaki. 1997. Rime cognate: A new lexical competitor set for spoken word recognition. Poster presentation at the 134th meeting of the Acoustical Society of America, San Diego, December 1-5, 1997. *Journal of the acoustical society of America* 102 (5, part 2).3135.
- Beckman, Mary. 1982. Segment duration and the 'mora' in Japanese. *Phonetica* 39.113-135.
- Bladon, Anthony. 1987. Extending the search for a psychophysical basis for dynamic phonetic patterns. *The psychophysics of speech perception*, ed. by M.E.H. Schouten, 258-263. Dordrecht: Martinus Nijhoff Publishers.
- Blevins, Juliette. 1995. The syllable in phonological theory. *A Handbook of Phonological Theory*, ed. by John A. Goldsmith, 206-244. Cambridge: Blackwell.
- Breen, J.W. 1997. JDIC: Japanese English Electronic Dictionary, v. 2.6. <http://www.rdt.monash.edu.au/~jwb/japanese.html#edict-proj>.
- Carnegie Mellon University. 1995. *The Carnegie Mellon Pronouncing Dictionary*, v. cmudict.0.4. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>.
- Chen, Matthew. 1970. Vowel length variation as a function of the voicing of the consonant environment. *Phonetica* 22.129-159.
- Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English*. New York: Harper and Row.
- Costa, Albert, Anne Cutler, and Nuria Sebastián-Gallés. In press. Effects of phoneme repertoire on phoneme decision. *Perception and psychophysics*.

- Cotton, Suzanne and François Grosjean. 1984. The gating paradigm: A comparison of successive and individual presentation formats. *Perception and psychophysics* 35.41-48.
- Cutler, Anne. 1986. Forbear is a homophone: Lexical prosody does not constrain lexical access. *Language and speech* 29.201-220.
- Cutler, Anne, and Sally Butterfield. 1992. Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of memory and language* 31.218-236.
- Cutler, A., and D.M. Carter. 1987. The predominance of strong initial syllables in the English vocabulary. *Computer speech & language* 2.133-142.
- Cutler, A., J. Mehler, D.G. Norris, and J. Segui. 1986. The syllable's differing role in the segmentation of French and English. *Journal of memory and language* 25.385-400.
- Cutler, A., J. Mehler, D.G. Norris, and J. Segui. 1992. The monolingual nature of speech segmentation by bilinguals. *Cognitive psychology* 24.381-410.
- Cutler, A., and D.G. Norris. 1988. The role of strong syllables in segmentation for lexical access. *Journal of experimental psychology: Human perception & performance* 14.113-121.
- Cutler, Anne, Dennis Norris, and James McQueen. 1996. Lexical access in continuous speech: Language-specific realisations of a universal model. *Phonological structure and language processing*, ed. by Takashi Otake and Anne Cutler, 227-242. Berlin: Mouton de Gruyter.
- Cutler, Anne, and Takashi Otake. 1994. Mora or phoneme? Further evidence for language-specific listening. *Journal of memory and language* 33.824-844.
- Cutler, Anne, and Takashi Otake. To appear. Pitch accent in spoken word recognition in Japanese.

- Delgutte, B. 1980. Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. *Journal of the acoustical society of America* 68.843-857.
- Delgutte, Bertrand. 1997. Auditory neural processing of speech. *The Handbook of Phonetic Sciences*, ed. by William J. Hardcastle and John Laver, 507-538. Cambridge: Blackwell.
- Efremova, I.B., K. Fintoft, and H. Ormestad. 1963. Intelligibility of tonic accents. *Phonetica* 10.203-212.
- Flemming, Edward S. 1995. Auditory representations in phonology. Ph.D. dissertation, UCLA.
- Frauenfelder, Uli H., and Lorraine Komisarjevsky Tyler, eds. 1987. Spoken word recognition. Cambridge: MIT Press.
- Frederiksen, John R. 1967. Cognitive factors in the recognition of ambiguous auditory and visual stimuli. (Monograph.) *Journal of personality and social psychology* 7.
- Fujimura, Osamu, M. J. Macchi, and L.A. Streeter. 1978. Perception of stop consonants with conflicting transitional cues: A cross-linguistic study. *Language and speech* 21.337-346.
- Fujisaki, Hiroya, and Sotaro Sekimoto. 1975. Perception of the time-varying resonance frequencies in speech and non-speech stimuli. *Structure and process in speech perception: Proceedings of the symposium on dynamic aspects of speech perception*, ed. by A. Cohen and S.G. Nooteboom, 269-282. Eindhoven, Netherlands, August 4-6, 1975. New York: Springer-Verlag.
- Furui, Sadaoki. 1986. On the role of spectral transition for speech perception. *Journal of the acoustical society of America* 80.1016-1025.
- Gesuato, Sara. 1996. Perception of alveolar and velar allophones of English /l/ in word-initial and word-final positions. *Proceedings of the 22nd annual meeting of the Berkeley Linguistics Society*, ed. by Jan Johnson, Matthew L. Juge, and Jeri L. Moxley, 116-131. February 16-19, 1996.



- Greenberg, Steven. 1994. Speech processing: Auditory models. *The encyclopedia of language and linguistics*, ed. by R.E. Asher, 4206-4227. Oxford: Pergamon.
- Greenberg, Steven. 1996a. Auditory processing of speech. *Principles of experimental phonetics*, ed. by Norman J. Lass, 362-407. St. Louis: Mosby.
- Greenberg, Steven. 1996b. Understanding speech understanding: Towards a unified theory of speech perception. *Proceedings of the ESCA tutorial and advanced research workshop on the auditory basis of speech perception*, Keele, England, 1-8.
- Greenberg, Steven. 1997. Auditory function. *Encyclopedia of acoustics*, ed. by Malcolm J. Crocker, 1301-1323. New York: John Wiley.
- Greenberg, Steven. 1998. Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. *Proceedings of the ESCA workshop on modeling pronunciation variation for automatic speech recognition*, Kexrode, May 3-6 1998.
- Grosjean, François. 1980. Spoken word recognition processes and the gating paradigm. *Perception and psychophysics* 28.267-283.
- Grosjean, François, and James Paul Gee. 1987. Prosodic structure and spoken word recognition. *Cognition* 25.135-155.
- Hayes, B. 1989. Compensatory lengthening in moraic phonology. *Linguistic inquiry* 20-2.253-306.
- Homma, Yayoi. 1981. Durational relationship between Japanese stops and vowels. *Journal of phonetics* 9.273-281.
- Hubbard, Kathleen. To appear. Quantity and quality: A cross-linguistic survey of vowel length patterns.
- Jamieson, Donald G. 1987. Studies of possible psychoacoustic factors underlying speech perception. *The psychophysics of speech perception*, ed. by M.E.H. Schouten, 220-230. Dordrecht: Martinus Nijhoff Publishers.

- Jenkins, James J., Winifred Strange, Kanae Nishi, Brett H. Fitzgerald, Sonja A. Trent, and David H. Thornton. 1997. Acoustic comparison of the effects of coarticulation on the production of Japanese and American English vowels. Poster presentation at the 134th meeting of the Acoustical Society of America, San Diego, December 1-5, 1997. *Journal of the acoustical society of America* 102 (5, part 2).3134.
- Johnson, Keith. 1997. The auditory/perceptual basis for speech segmentation. *Ohio State University working papers in linguistics* 50.101-113.
- Jongman, Allard. 1989. Duration of frication noise required for identification of English fricatives. *Journal of the acoustical society of America* 85.1718-1725.
- Jongman, Allard, Joan A. Sereno, Marianne Raaijmakers, and Aditi Lahiri. 1992. The phonological representation of [voice] in speech perception. *Language and speech* 35.137-152.
- Jonides, John, and Steven Yantis. 1988. Uniqueness of abrupt visual onset in capturing attention. *Perception and psychophysics* 43.346-354.
- Takehi, Kazuhiko, Kazumi Kato, and Makio Kashino. 1996. Phoneme/syllable perception and the temporal structure of speech. *Phonological structure and language processing*, ed. by T. Otake and A. Cutler, 126-143. Berlin: Mouton de Gruyter.
- Kawasaki, Haruko. 1982. An acoustical basis for universal constraints on sound sequences. Ph.D. dissertation, University of California, Berkeley.
- Keppel, Geoffrey. 1991. *Design and analysis: A researcher's handbook*. 3rd. ed. Englewood Cliffs, NJ: Prentice Hall.
- Kewley-Port, Diane. 1983. Time-varying features as correlates of place of articulation in stop consonants. *Journal of the acoustical society of America* 73.322-335.

- Kewley-Port, Diane, David B. Pisoni, and Michael Studdert-Kennedy. 1983. Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *Journal of the acoustical society of America* 73.1779-1793.
- Klatt, Dennis H. 1979. Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of phonetics* 7.279-312.
- Kluender, Keith R., and Margaret A. Walsh. 1992. Amplitude rise time and the perception of the voiceless affricate/fricative distinction. *Perception and psychophysics* 51.328-333.
- Kluender, Keith R., Randy L. Diehl, and Beverly A. Wright. 1988. Vowel-length differences before voiced and voiceless consonants: an auditory explanation. *Journal of phonetics* 16.153-169.
- Lahiri, Aditi, and Allard Jongman. 1990. Intermediate level of analysis: Features or segments? *Journal of phonetics* 18.435-443.
- Lahiri, Aditi, and William Marslen-Wilson. 1991. The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition* 38.245-294.
- Lahiri, Aditi, and William Marslen-Wilson. 1992. Lexical processing and phonological representation. *Papers in laboratory phonology II: Gesture, segment, prosody*, ed. by Gerard J. Docherty and D. Robert Ladd, 229-254.
- Lang, Carrie, and John J. Ohala. 1996. Temporal cues for vowels and universals of vowel inventories. *Proceedings of the Fourth International Conference on Spoken Language Processing*, October 3-6, 1996, Philadelphia.
- Lindblom, Björn. 1984. Can the models of evolutionary biology be applied to phonetic problems? *Proceedings of the Tenth International Congress of Phonetic Sciences*, Utrecht, 1983, ed. by M.P.R. Van den Broecke and A.Cohen, 67-81. Dordrecht: Foris Publications.
- Luce, Paul A., David B. Pisoni, and Steven D. Goldinger. 1990. Similarity neighborhoods of spoken words. *Cognitive models of speech processing:*

- Psycholinguistic and computational perspectives, ed. by Gerry T.M. Altmann, 122-147. Cambridge: MIT Press.
- Marslen-Wilson, William D. 1987. Functional parallelism in spoken word-recognition. *Cognition* 25.71-102.
- Marslen-Wilson, William D. (ed). 1989. *Lexical representation and process*. Cambridge: MIT Press.
- Marslen-Wilson, William. 1990. Activation, competition, and frequency in lexical access. *Cognitive models of speech processing: Psycholinguistic and computational perspectives*, ed. by Gerry T.M. Altmann, 148-172. Cambridge: MIT Press.
- Marslen-Wilson, William, and Paul Warren. 1994. Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological review* 101.653-675.
- Marslen-Wilson, William D., and Alan Welsh. 1978. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive psychology* 10.29-63.
- McClelland, James L., and Jeffrey L. Elman. 1986. The TRACE model of speech perception. *Cognitive psychology* 18.1-86.
- McQueen, James M. 1997. Transitional probability, not lexical knowledge, influences compensation. Paper given at the 134th meeting of the Acoustical Society of America, San Diego, December 1-5, 1997. *Journal of the acoustical society of America* 102 (5, part 2).3134.
- McQueen, J.M., D.G. Norris, and A. Cutler. 1994. Competition in spoken word recognition: Spotting words in other words. *Journal of experimental psychology: Learning, memory and cognition* 20.621-638.
- Mehler, Jacques, Jean Yves Dommergues, Uli Frauenfelder, and Juan Segui. 1981. The syllable's role in speech segmentation. *Journal of verbal learning and verbal behavior* 20.298-305.

- Mermelstein, Paul. 1978. On the relationship between vowel and consonant identification when cued by the same acoustic information. *Perception and psychophysics* 23.331-336.
- Nearey, Terrance M. 1997. Speech perception as pattern recognition. *Journal of the acoustical society of america* 101.3241-3254.
- Nearey, Terrance M., and Peter F. Assmann. 1986. Modeling the role of inherent spectral change in vowel identification. *Journal of the acoustical society of America* 80.1297-1308.
- Nihon Hoosoo Kyookai (NHK). 1985. *Nihongo hatsuon akusento jiten* [Japanese pronunciation and accent dictionary]. Tokyo: Nihon Hoosoo Shuppan Kyookai.
- Ohala, John J. 1975. Phonetic explanations for nasal sound patterns. *Nasalfest: Papers from a symposium on nasals and nasalization*, ed. by C.A. Ferguson, L.M. Hyman, and J.J. Ohala, 289-316. Stanford: Language Universals Project.
- Ohala, John J. 1981a. Articulatory constraints on the cognitive representation of speech. *The cognitive representation of speech*, ed. by Terry Myers, John Laver, and John Anderson, 111-122. Amsterdam: North Holland Publishing Company.
- Ohala, John J. 1981b. The listener as a source of sound change. *Papers from the parasession on language and behavior*, Chicago Linguistic Society, ed. by Carrie Masek, Roberta Hendrick, and Mary Frances Miller, 178-203. May 1-2, 19891, Chicago, Illinois.
- Ohala, John J. 1983. The origin of sound patterns in vocal tract constraints. *The production of speech*, ed. by Peter F. MacNeilage, 189-216. New York: Springer Verlag.
- Ohala, John J. 1986. Phonological evidence for top-down processing in speech perception. *Invariance and variability in speech processes*, ed. by Joseph S. Perkell and Dennis H. Klatt, 386-401. Hillsdale, NJ: Lawrence Erlbaum.

- Ohala, John J. 1989. Sound change is drawn from a pool of synchronic variation. *Language change: Contributions to the study of its causes*, ed. by Leiv Egil Breivik and Ernst Håkon Jahr, 173-198. Berlin: Mouton de Gruyter.
- Ohala, John J. 1992. Comments on chapter 9. *Papers in laboratory phonology II: Gesture, segment, prosody*, ed. by Gerard J. Docherty and D. Robert Ladd, 255-257.
- Ohala, John J., and Haruko Kawasaki-Fukumori. 1996. Alternatives to the sonority hierarchy for explaining segmental sequential constraints. *Studies for Einar Haugen*, ed. by S. Eliasson and E.H. Jahr. Berlin: Mouton de Gruyter.
- Ohala, John J. and Manjari Ohala. 1995. Speech perception and lexical representation: The role of vowel nasalization in Hindi and English. *Phonology and phonetic evidence, papers in laboratory phonology IV*, ed. by Bruce Connell and Amalia Arvaniti, 41-60. Cambridge: Cambridge University Press.
- Ohde, R.N., and M.T. Ochs. 1996. The effect of segment duration on the perceptual integration of nasals for adult and child speech. *Journal of the acoustical society of America* 100.2486-2499.
- Öhman, S.E. 1966. Perception of segments of VCCV utterances. *Journal of the acoustical society of America* 40.979-988.
- Otake, T., G. Hatano, A. Cutler, and J. Mehler. 1993. Mora or syllable? Speech segmentation in Japanese. *Journal of memory and language* 32.258-278.
- Peterson, Gordon E., and Ilse Lehiste. 1960. Duration of syllable nuclei in English. *Journal of the acoustical society of America* 32.693-703.
- Pierrehumbert, Janet B., and Mary E. Beckman. 1988. *Japanese tone structure*. Cambridge: MIT Press.
- Pisoni, David B., H.C. Nusbaum, P.A. Luce, and L.M. Slowiczek. 1985. Speech perception, word recognition, and the structure of the lexicon. *Speech communication* 4.75-95.

- Plomp, R. 1964. The ear as a frequency analyzer. *Journal of the acoustical society of America* 36.1628-1636.
- Pols, L.C.W., and M.E.H. Schouten. 1978. Identification of deleted consonants. *Journal of the acoustical society of America* 64.1333-1337.
- Port, Robert F., Salman Al-Ani, and Shosaku Maeda. 1980. Temporal compensation and universal phonetics. *Phonetica* 37.235-252.
- Poser, William J. 1988. Glide formation and compensatory lengthening in Japanese. *Linguistic inquiry* 19.494-503.
- Raphael, Lawrence J. 1972. Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *Journal of the acoustical society of America* 51.1296-1303.
- Raphael, Lawrence J. 1981. Durations and contexts as cues to word-final cognate opposition in English. *Phonetica* 38.126-147.
- Repp, Bruno H. 1980. Accessing phonetic information during perceptual integration of temporally distributed cues. *Journal of phonetics* 8.185-194.
- Roengpitya, Rungpat. In Press. A perceptual experiment on Thai consonant types and tones. *Proceedings of the joint meeting of the 16th International Congress on Acoustics and the 135th meeting of the Acoustical Society of America, June 20-26, 1998, Seattle.*
- Salasoo, Aita, and David B. Pisoni. 1985. Interaction of knowledge sources in spoken word identification. *Journal of memory and language* 24.210-231.
- Shillcock, Richard. 1990. Lexical hypotheses in continuous speech. *Cognitive models of speech processing: Psycholinguistic and computational perspectives*, ed. by Gerry T.M. Altmann, 24-49. Cambridge: MIT Press.
- Smolensky, Paul. 1996. On the comprehension/production dilemma in child language. *Linguistic inquiry* 27.

- Steriade, Donca. 1997. Phonetics in phonology: The case of laryngeal neutralization. Unpublished manuscript, UCLA, June 1997.
- Stevens, Kenneth N. 1971. The role of rapid spectrum changes in the production and perception of speech. Form and substance. *Phonetic and linguistic papers*. Presented to Eli Fischer-Jørgensen, ed. by L.L. Hammerich, Roman Jakobson, and Eberhard Zwirner, 95-101. Kobenhavn, Akademisk Forlag.
- Stevens, K.N. 1980. Discussion. *Proceedings of the Ninth International Congress of Phonetic Sciences, Copenhagen, 1979*, 3.185-186.
- Stevens, Kenneth N. 1985. Evidence for the role of acoustic boundaries in the perception of speech sounds. *Phonetic linguistics: Essays in Honor of Peter Ladefoged*, ed. by Victoria A. Fromkin, 243-255. Orlando: Academic Press.
- Stevens, K.N., and S.E. Blumstein. 1978. Invariant cues for place of articulation in stop consonants. *Journal of the acoustical society of America* 64.1358-1368.
- Stevens, K.N., and S.E. Blumstein. 1981. The search for invariant acoustic correlates of phonetic features. *Perspectives on the study of speech*, ed. by P.D. Eimas and J. Miller, 1-38. Hillsdale, NJ: Erlbaum.
- Stevens, Kenneth N., and Samuel Jay Keyser. 1989. Primary features and their enhancement in consonants. *Language* 65.81-106.
- Stevens, Kenneth N., Samuel Jay Keyser, and Haruko Kawasaki. 1986. Toward a phonetic and phonological theory of redundant features. *Invariance and variability in speech processes*, ed. by Joseph S. Perkell and Dennis H. Klatt, 426-463. Hillsdale, NJ: Lawrence Erlbaum.
- Strange, Winifred, James J. Jenkins, and Thomas L. Johnson. 1983. Dynamic specification of coarticulated vowels. *Journal of the acoustical society of America* 74.695-705.
- Sugito, Miyoko. 1995. *Osaka Tokyo akusento onsei jiten [Osaka Tokyo accent phonetic dictionary]*. CD ROM. Maruzen.



- Taft, Marcus, and Gail Hambly. 1986. Exploring the cohort model of spoken word recognition. *Cognition* 22.259-282.
- 't Hart, J., and A. Cohen. 1964. Gating techniques as an aid in speech analysis. *Language and speech* 7.22-39.
- Tillmann, Hans-Günther. 1980. *Phonetik: Lautsprachliche Zeichen, Sprachsignale und lautsprachlicher Kommunikationsprozeß*. Stuttgart: Klett-Cotta.
- Tyler, Lorraine K. 1984. The structure of the initial cohort: Evidence from gating. *Perception and psychophysics* 36.417-427.
- Tyler, Lorraine K., and Jeanine Wessels. 1985. Is gating an on-line task? Evidence from naming latency data. *Perception and psychophysics* 38.217-222.
- Uchida, Teruhisa. 1997. Categorical perception of Japanese moraic phonemes. Poster presentation at the 134th meeting of the Acoustical Society of America, San Diego, December 1-5, 1997. *Journal of the acoustical society of America* 102 (5, part 2).3094.
- Vance, Timothy J. 1987. *An introduction to Japanese phonology*. Albany: State University of New York Press.
- Walsh, Thomas, and Frank Parker. 1983. Vowel length and vowel transition: cues to [±voice] in post-vocalic stops. *Journal of phonetics* 11.407-412.
- Warren, Paul, and William Marslen-Wilson. 1987. Continuous uptake of acoustic cues in spoken word recognition. *Perception and psychophysics* 41.262-275.
- Warren, Paul, and William Marslen-Wilson. 1988. Cues to lexical choice: Discriminating place and voice. *Perception and psychophysics* 43.21-30.
- Yantis, Steven, and John Jonides. 1984. Abrupt visual onsets and selective attention: Evidence from visual search. *Journal of experimental psychology: Human perception and performance* 10.601-620.



Appendix A: Subjects' language backgrounds<sup>1</sup>**English experiment** (total: 154 subjects)

Number of subjects is shown to the right of the category

**Monolingual English subjects:**

No other language during childhood	73 <sup>2</sup>
Minimal exposure to another language during childhood (includes Spanish, Yiddish, Hebrew, French, Avestan, Polish, Arabic, and Mandarin)	11

**Bilingual subjects (acquired other language first, now English dominant)<sup>3</sup>:**

Cantonese/English	9
Mandarin/English	8
Korean/English	4
Vietnamese/English	3
Taiwanese/English	2
Farsi/English	1
Finnish/English	1
Hebrew/English	1
Japanese/English	1
Chow-Jew (Chinese)/English	1
Portuguese/English	1
Russian/English	1
Tagalog/English	1
Thai/English	1
Urdu/English	1

**Bilingual subjects (acquired English and the other language simultaneously, now English dominant):**

Tagalog/English	2
Spanish/English	2
Cantonese/English	2
Italian/English	1
Turkish/English	1
Mandarin/English	1
Telugu/English	1
Vietnamese/English	1

**Balanced bilingual subjects:**

Spanish and English learned simultaneously	1
Spanish learned first, now Spanish/English balanced	3
Russian learned first, now Russian/English balanced	1
Mandarin learned first, now Mandarin/English balanced	1

---

<sup>1</sup> Only languages to which subjects were exposed as children are listed.

<sup>2</sup> Includes one speaker each of British, Brooklyn, and Boston English.

<sup>3</sup> Both this and the next category include subjects who are now passive bilinguals in the language other than English.

**Trilingual subjects****All three languages learned from infancy, now English dominant:**

Cantonese/Mandarin/English all learned simultaneously, now English dominant	1
Mandarin/Taiwanese/English all learned simultaneously, now English dominant	1
English/Gujerati/Hindi all learned simultaneously, now English dominant	1
English/Hindi/Punjabi all learned simultaneously, now English dominant	1

**Two languages learned from infancy, now English dominant:**

Kmer and Cantonese learned first, now English dominant	1
Mandarin and Taiwanese learned first, now English dominant	2

**One language learned from infancy, now English dominant, varying degrees of exposure and current ability in another language:**

Korean learned first, now English dominant, also speaks/spoke Japanese	1
Ilokano learned first, now English dominant, also speaks/spoke Tagalog	1
Hindi learned first, now English dominant, also speaks/spoke Punjabi	1
Polish learned first, now English dominant, also speaks/spoke German	1
Mandarin learned first, now English dominant, also speaks/spoke Taiwanese	2

**Now balanced bilingual with varying degrees of exposure and current ability in another language:**

Mandarin learned first, now M./English balanced, also speaks/spoke Cantonese	1
Laotian learned first, now Laotian/English balanced, also speaks/spoke Thai	1
Croatian and English learned first, now English/Spanish balanced	1

Japanese experiment (total: 120 subjects, all monolingual or nearly monolingual Japanese speakers except for languages learned as young adults)

Number of subjects who lived in only one dialect area as children:

Tokyo	22
Aichi	28
Kanagawa	6
Yokohama	6
Gifu	5
Fukuoka	4
Saitama	4
Chiba	3
Ibaraki	3
Hyoogo	2
Mie	2
Osaka	2
Shizuoka	2
Tochigi	2
Fukushima	1
Hiroshima	1
Hokkaido	1
Ishikawa	1
Kagawa	1
Kagoshima	1
Miyazaki	1
Nagano	1
Niigata	1
Wakayama	1
no response	1

Number of subjects who lived in more than one dialect area as children:

Hyoogo, Tokyo	2
Aichi, Toyama, Aomori	1
Chiba, Hyoogo	1
Fukuoka, Osaka	1
Hiroshima, Takamatsu, Nagoya	1
Hyoogo, Kanagawa, Niigata, Chiba, Aichi, Saitama	1
Hyoogo, Tokyo, Hokkaido	1
Ibaraki, Tokyo	1
Iwate, Osaka	1
Mie, Tokyo	1
Nagoya, Shizuoka	1
Niigata, Hokkaido, Tokyo	1
Osaka, Chiba, Ibaraki	1
Osaka, Kyoto, Shiga	1
Shizuoka, Osaka	1
Tokyo, Osaka	1
Tokyo, Sendai	1

Number of subjects who lived outside Japan before age 15, and ages during which they lived there:

US	3	ages 0-2.5, 5-6, and 9-12
England	1	age 5-6
Switzerland	1	age 9-12

## Appendix B: Graphs of results

The following pages show graphs of the data for each word for both the number of responses measure (#Resp) and the percent correct measure (%Corr). Each graph shows the data points (solid line with diamond markers), the best linear fit (dotted line with unfilled square markers), and the fitted ogival curve (dashed line with triangular markers). When the data was excluded because of an anomalous slope (not negative for #Resp or not positive for %Corr), its graph shows only the linear fit, as no ogival curve was fit to such data. Data which was better fit by a line than by an ogival curve does show both the linear fit and the unsuccessfully fitted curve.

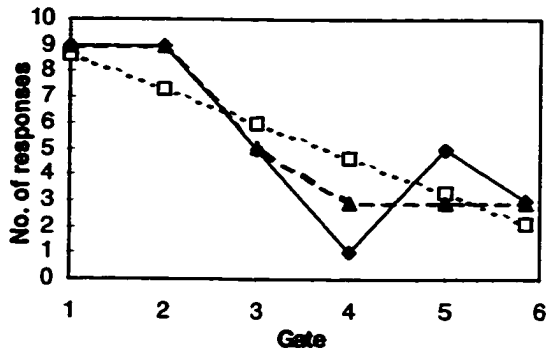
Above each graph is the number and name of the word (numbers keyed to the tables in Chapter 4), and the transition of interest. The number after the capital L is the least squares error of the linear fit, the number after the capital O is the least squares error of the ogival fit, and the numbers after the small m are the area of maximal slope of the fitted curve, in gate numbers.

For each graph, the x-axis represents gate number (time of endpoint of gate), and the y-axis is either the number of different responses given to a stimulus (#Resp) converted to a 0-10 scale or the percent of responses with the second segment of the transition correct (%Corr) converted to a 0-1 scale, as described in Chapter 4.

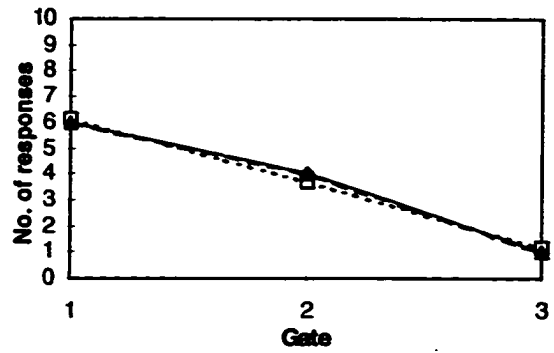
The graphs appear in the following order:

1. English number of responses measure (from page 340)
2. English percent correct measure (from page 356)
3. Japanese number of responses measure (from page 372)
4. Japanese percent correct measure (from page 382)

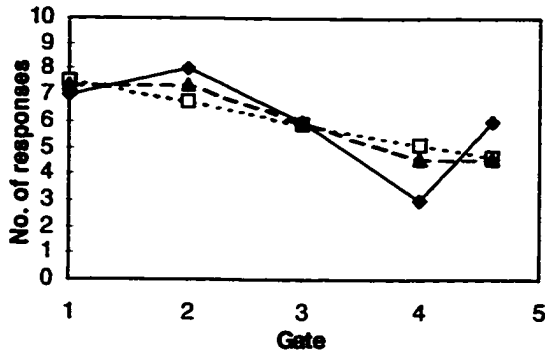
1 tip /tɪ/ L: 4.554 O: 2.834 m: 2-3



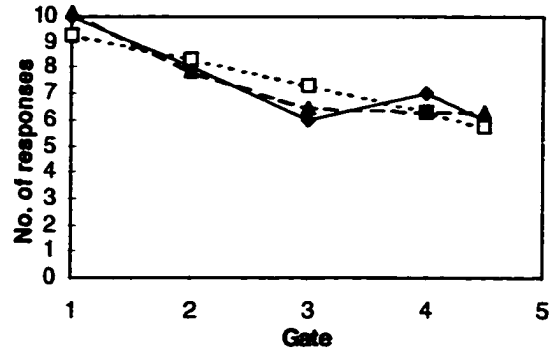
5 attic /tɪ/ L: 0.408 O: 0.060 m: 2-3



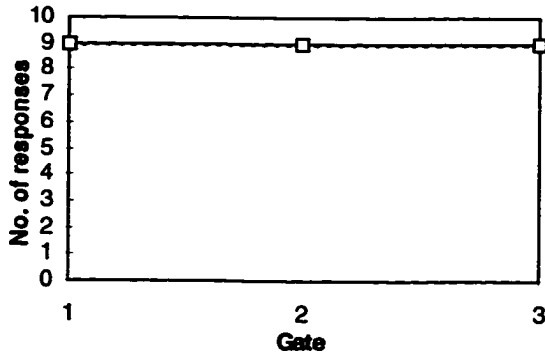
2 stiff /tɪ/ L: 2.873 O: 2.238 m: 3-4



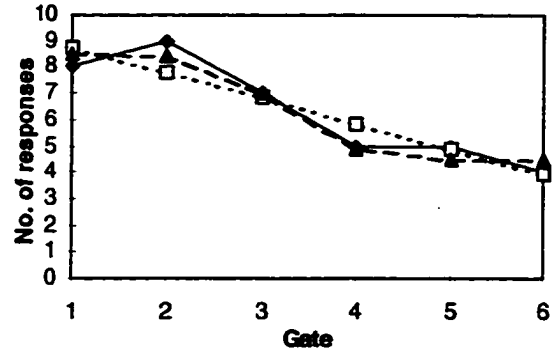
6 custom /kʌ/ L: 1.673 O: 0.913 m: 1-2



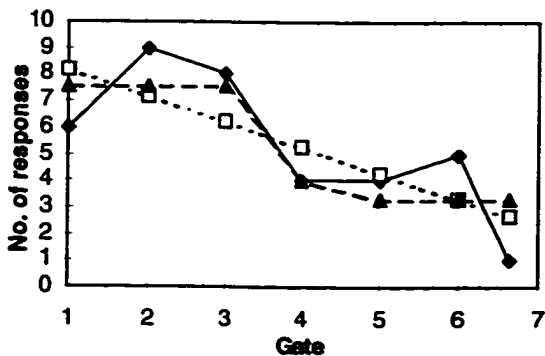
3 Tibet /tɪ/ L: 0.000



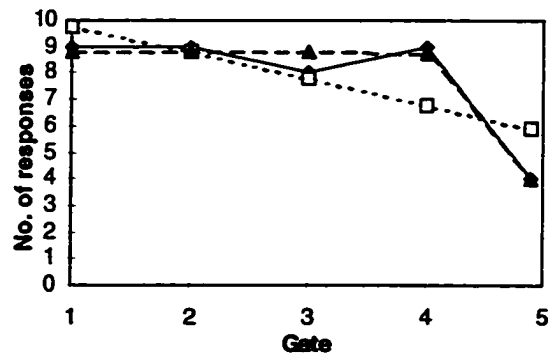
7 skull /kʌ/ L: 1.679 O: 1.053 m: 3-4



4 petition /tɪ/ L: 4.292 O: 3.654 m: 3-4

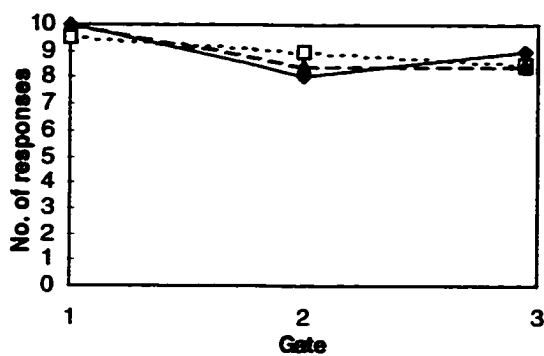


8 accompany /kʌ/ L: 3.028 O: 0.890 m: 4-5

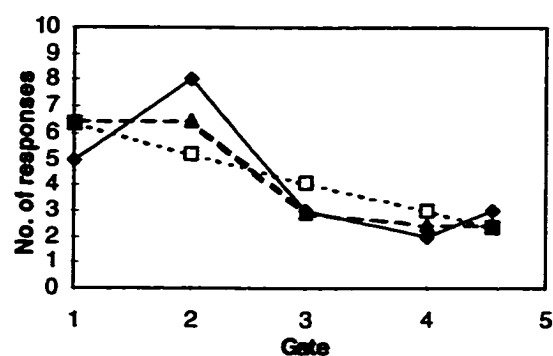




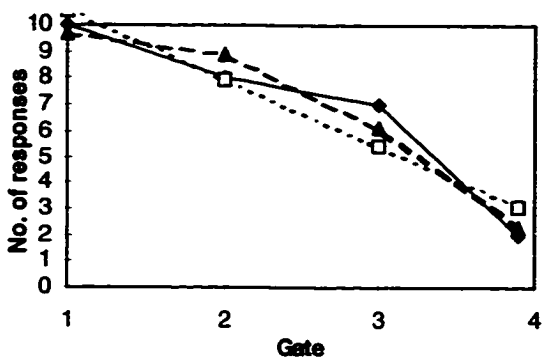
9 caboose /kə/ L: 1.214 O: 0.710 m: 1-2



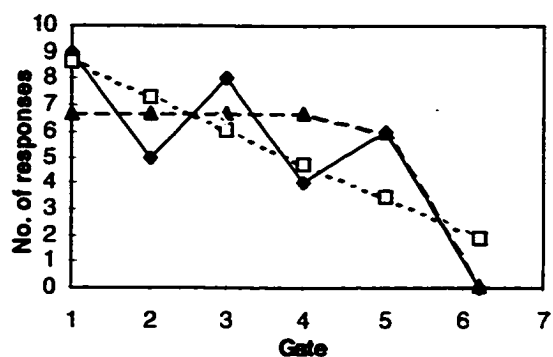
13 fitness /t/ L: 3.478 O: 2.245 m: 2-3



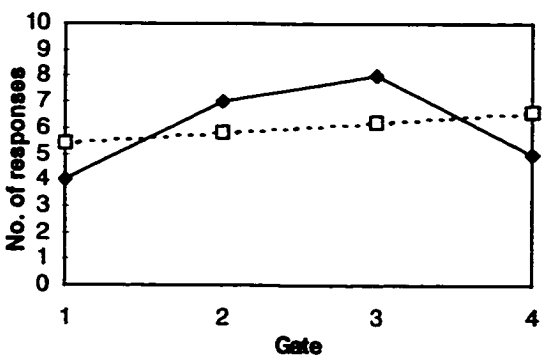
10 academic /kə/ L: 2.008 O: 1.291 m: 3-4



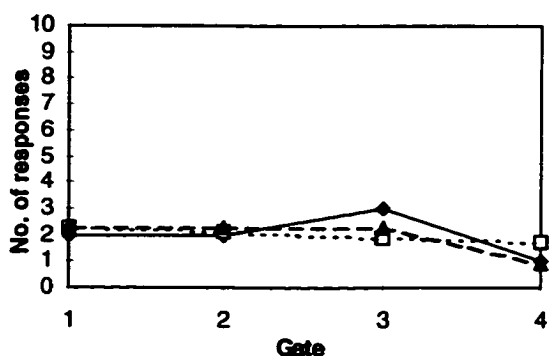
14 Italian /t/ L: 4.481 O: 4.140 m: 5-6



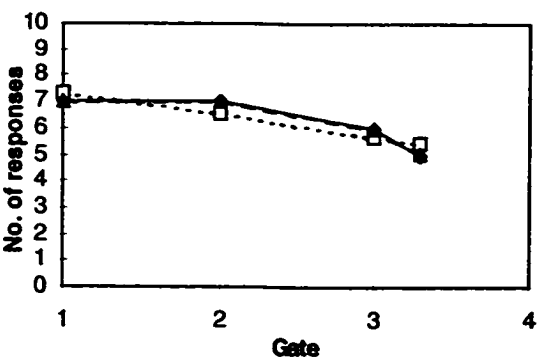
11 duck /dʌ/ L: 3.033



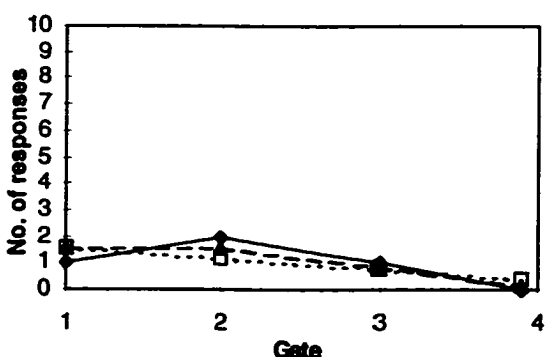
15 committee /t/ L: 1.342 O: 0.822 m: 3-4



12 citizen /t/ L: 0.801 O: 0.050 m: 3-4



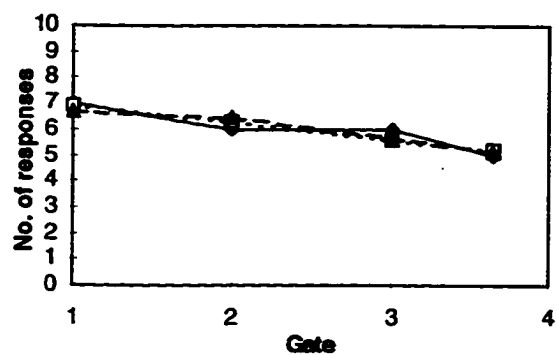
16 unity /t/ L: 1.110 O: 0.740 m: 3-4



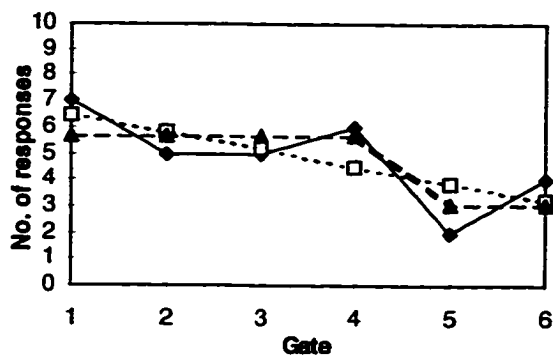
17 bucket /ʌk/ L: 2.683 O: 1.002 m: 4-5



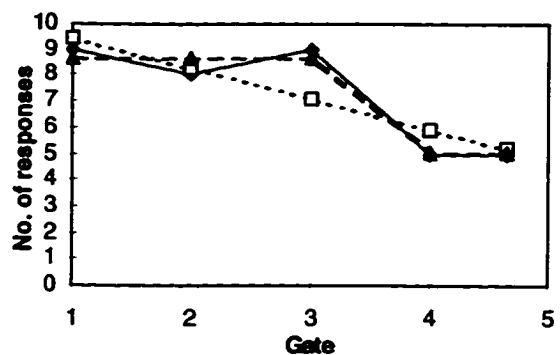
21 muddy /ʌd/ L: 0.512 O: 0.640 m: 3-4



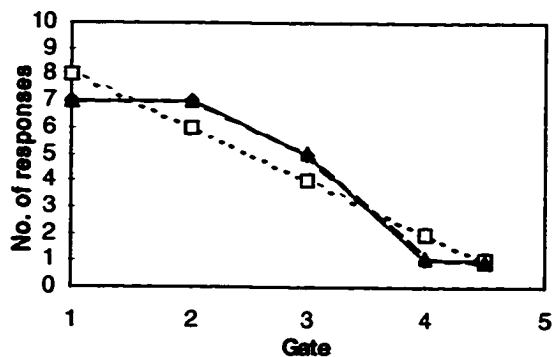
18 mechanical /ək/ L: 2.697 O: 2.184 m: 4-5



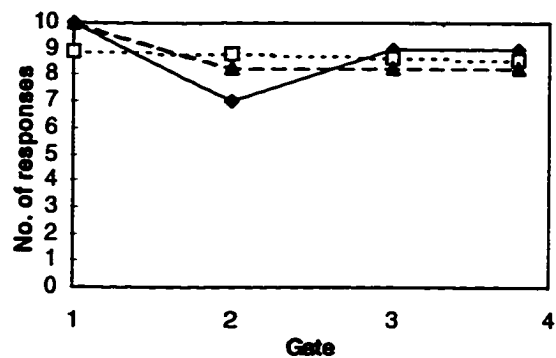
22 cadenza /əd/ L: 2.184 O: 0.822 m: 3-4



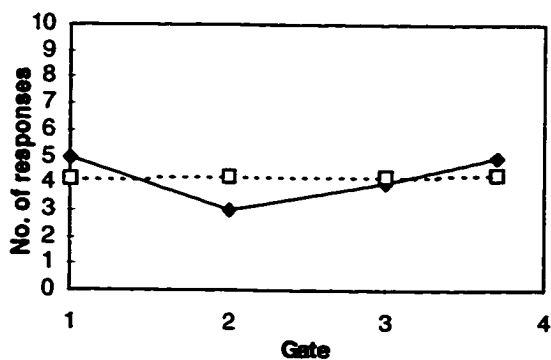
19 indicate /ək/ L: 2.000 O: 0.141 m: 3-4



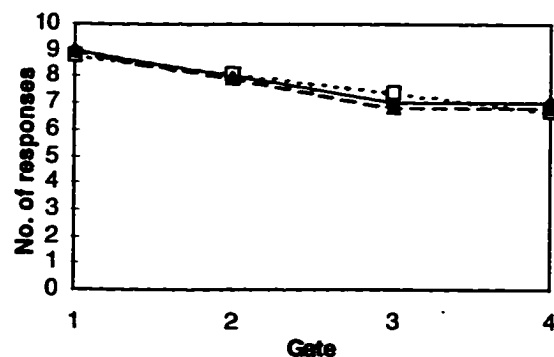
23 medicine /mɛ/ L: 2.164 O: 1.634 m: 1-2



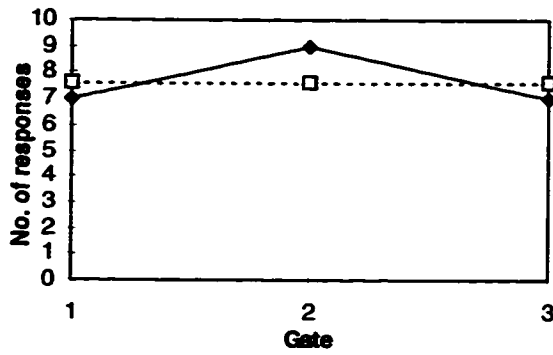
20 induction /ʌk/ L: 1.653



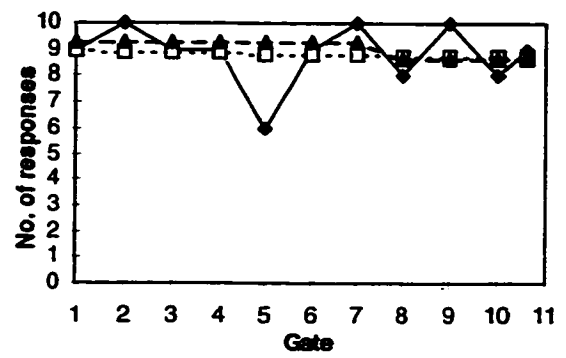
24 immense /mɛ/ L: 0.548 O: 0.175 m: 2-3



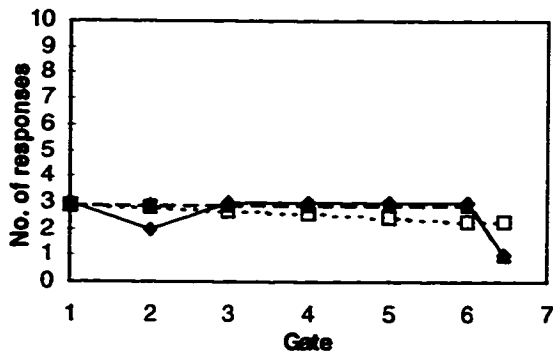
25 remedy /ɛm/ L: 1.633



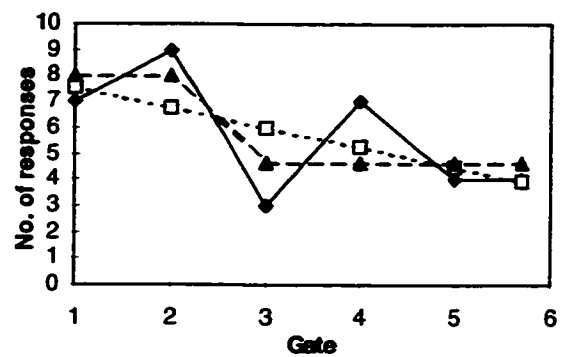
29 saddle /sæ/ L: 3.681 O: 3.873 m: 7-8



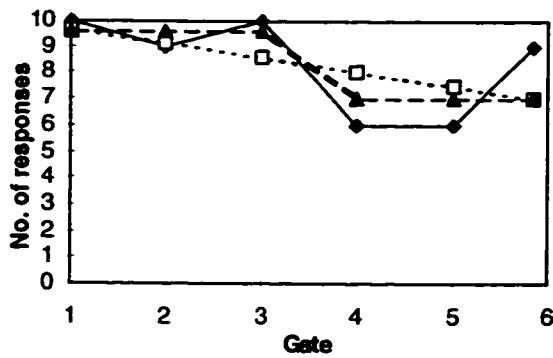
26 attempt /ɛm/ L: 1.822 O: 0.920 m: 6-7



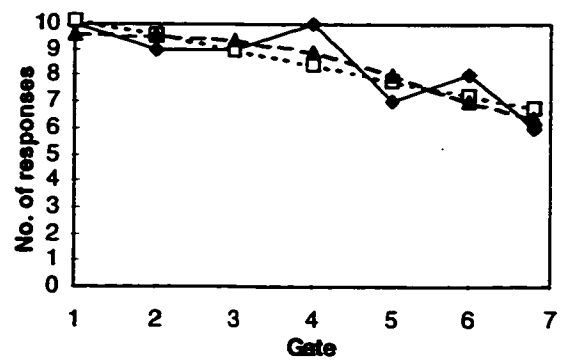
30 master /æz/ L: 4.197 O: 3.324 m: 2-3



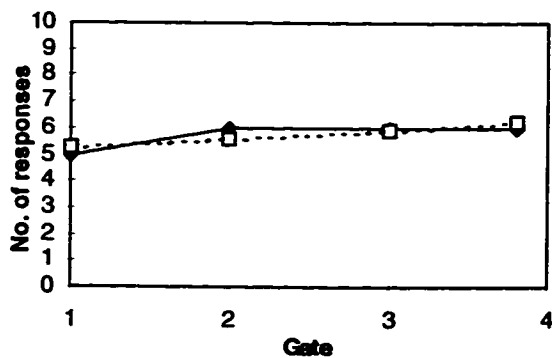
27 negative /nɛ/ L: 3.521 O: 2.586 m: 3-4



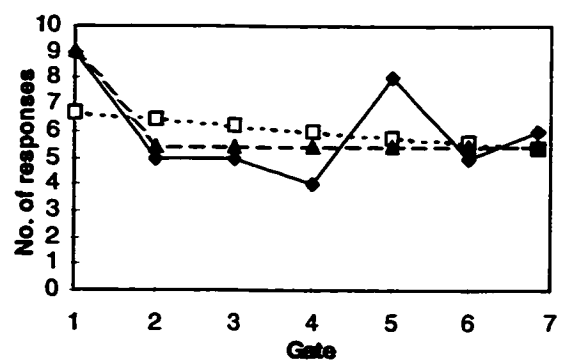
31 Zachary /zæ/ L: 2.178 O: 2.003 m: 5-6



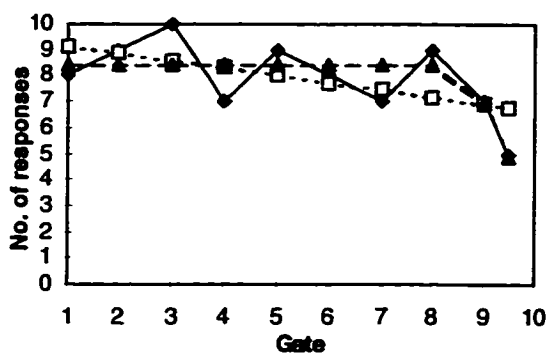
28 tenants /ɛn/ L: 0.525



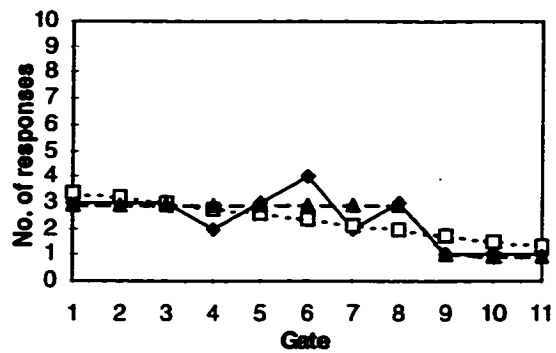
32 asthma /æz/ L: 4.321 O: 3.085 m: 1-2



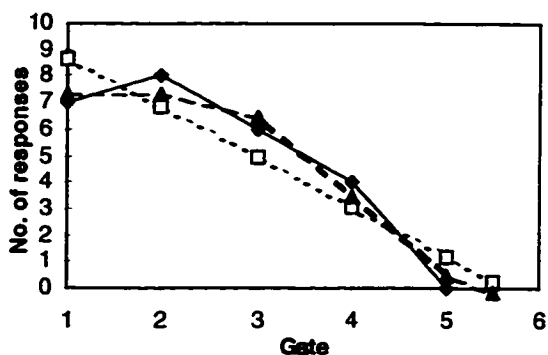
33 shell /ʃɛ/ L: 3.562 O: 2.810 m: 9-10



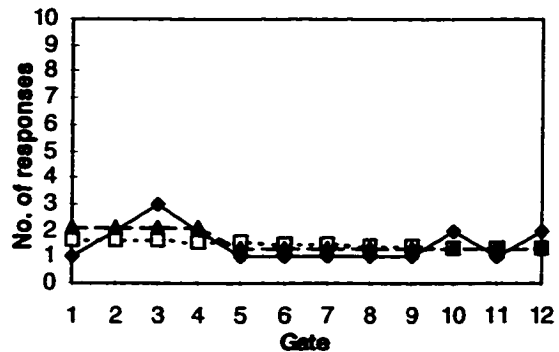
37 leaf /iʃ/ L: 2.395 O: 1.698 m: 8-9



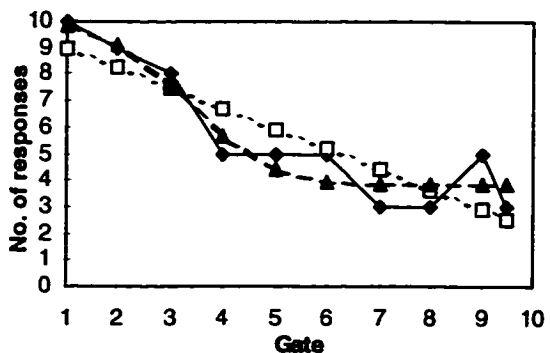
34 session /ɛʃ/ L: 2.782 O: 1.169 m: 4-5



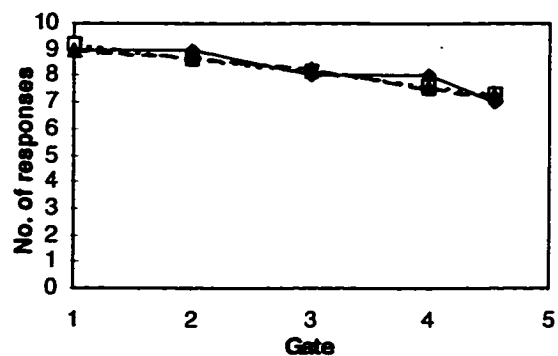
38 relief /iʃ/ L: 2.197 O: 1.892 m: 4-5



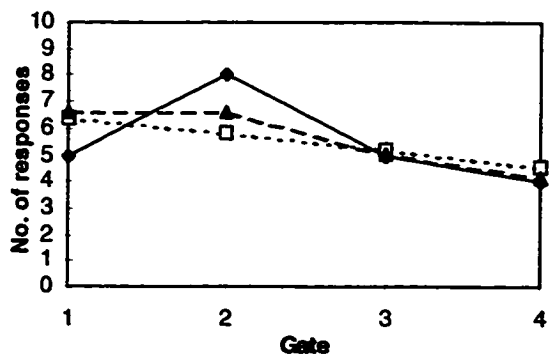
35 fees /fi/ L: 3.583 O: 2.374 m: 3-4



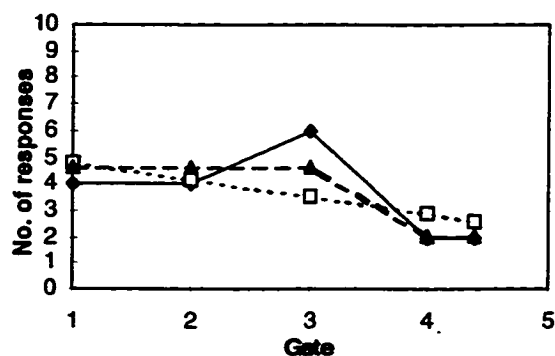
39 vacuum /væ/ L: 0.649 O: 0.606 m: 3-4



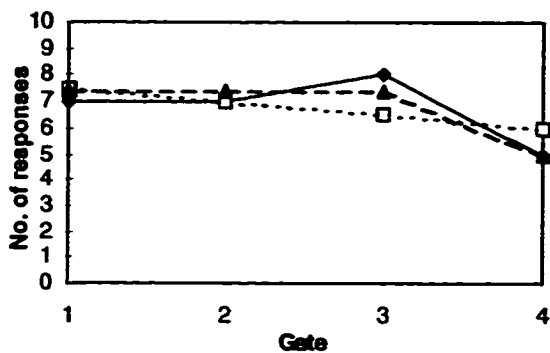
36 unfeeling /fi/ L: 2.683 O: 2.133 m: 2-3



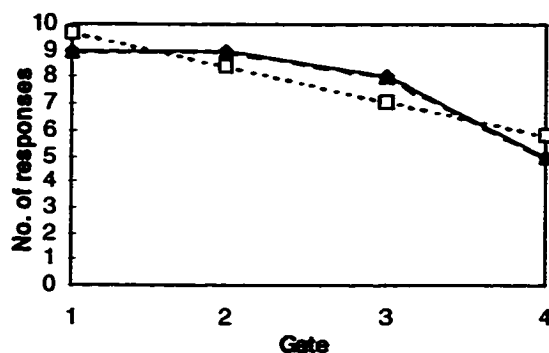
40 ravish /æv/ L: 2.825 O: 1.636 m: 3-4



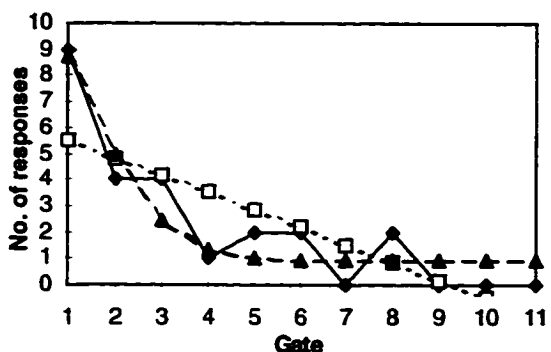
41 trail /re<sup>i</sup>/ L: 1.871 O: 0.822 m: 3-4



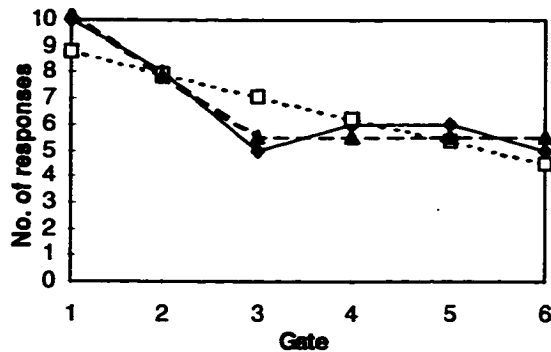
45 yellow /je/ L: 1.517 O: 0.063 m: 3-4



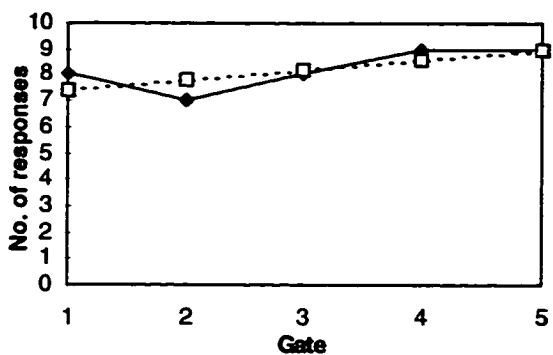
42 fair /e<sup>i</sup>r/ L: 5.019 O: 3.234 m: 1-2



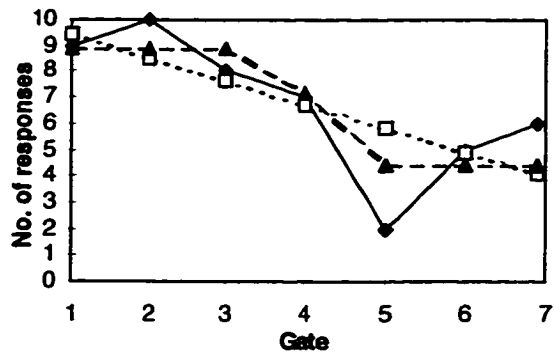
46 watch /wa/ L: 2.545 O: 1.031 m: 2-3



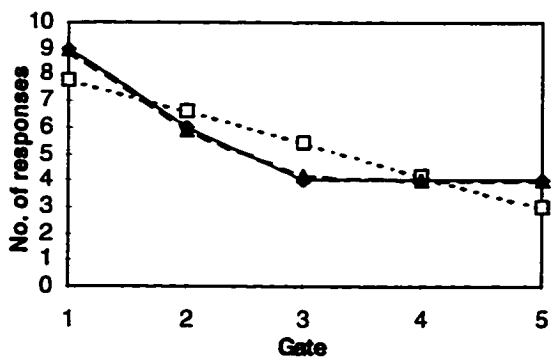
43 lever /ε/ L: 1.095



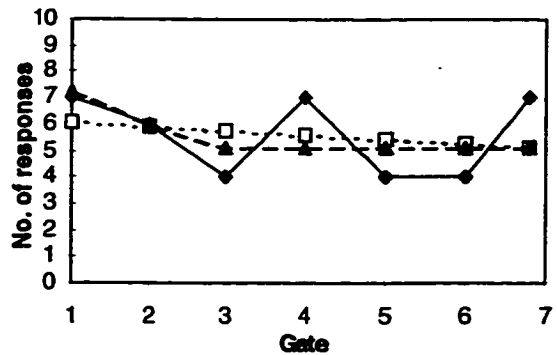
47 chapel /tʃæ/ L: 4.556 O: 3.287 m: 4-5



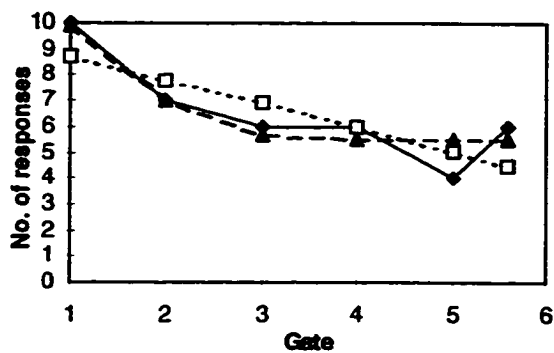
44 elevator /εl/ L: 2.191 O: 0.194 m: 1-2



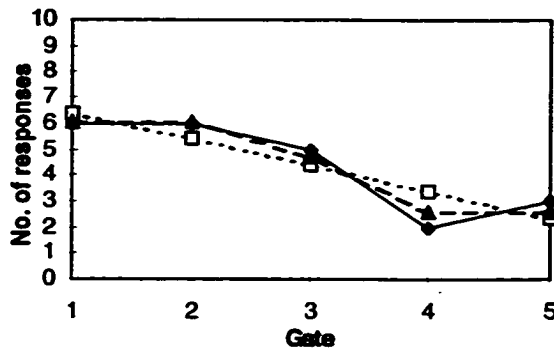
48 latches /ætʃ/ L: 3.610 O: 3.310 m: 1-2



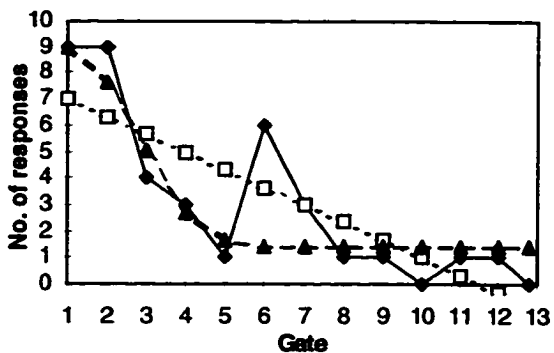
49 jump /dʒʌ/ L: 2.534 O: 1.689 m: 1-2



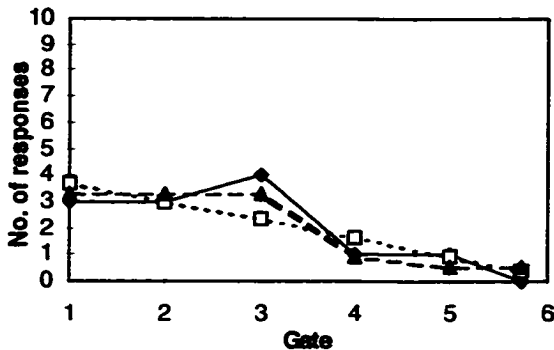
53 reinterpret /nt/ L: 1.789 O: 0.796 m: 3-4



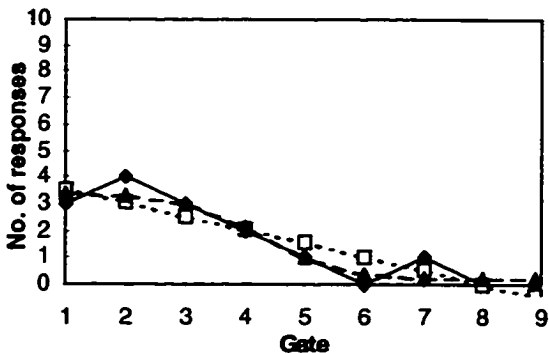
50 judge /ʌdʒ/ L: 6.375 O: 5.618 m: 2-3



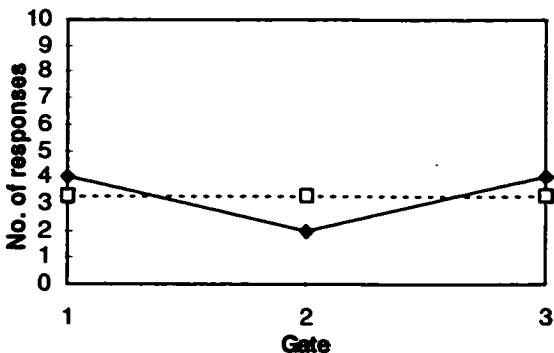
54 band /nd/ L: 1.972 O: 1.087 m: 3-4



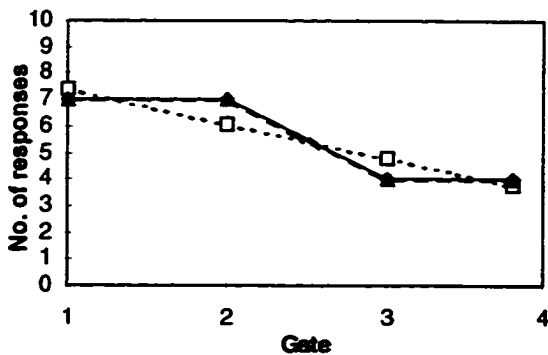
51 bent /nt/ L: 1.783 O: 1.208 m: 4-5



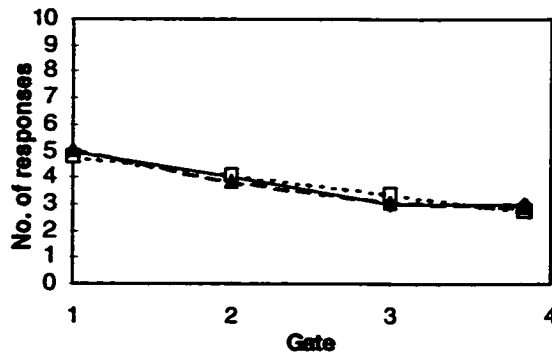
55 wander /nd/ L: 1.633



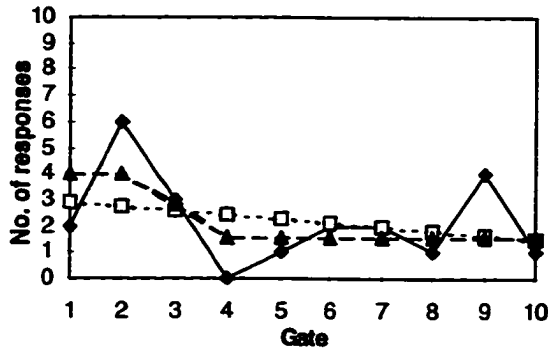
52 sentiment /nt/ L: 1.291 O: 0.004 m: 2-3



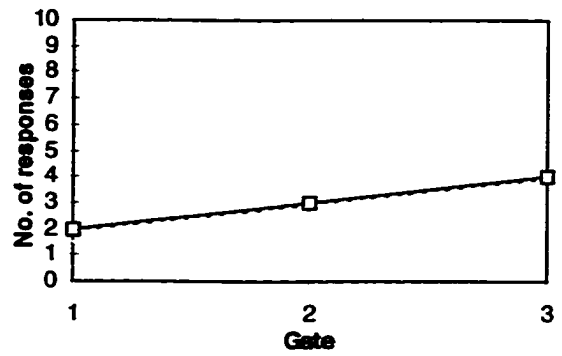
56 reconditioned /nd/ L: 0.487 O: 0.197 m: 1-2



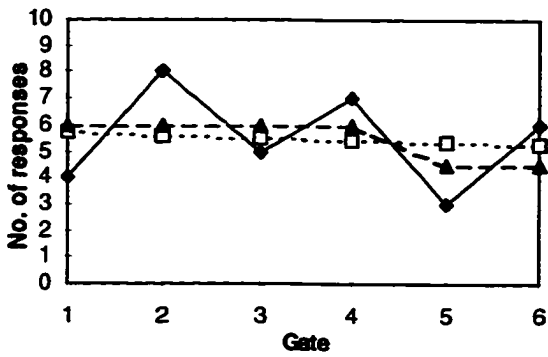
57 axe /ks/ L: 5.055 O: 4.214 m: 3-4



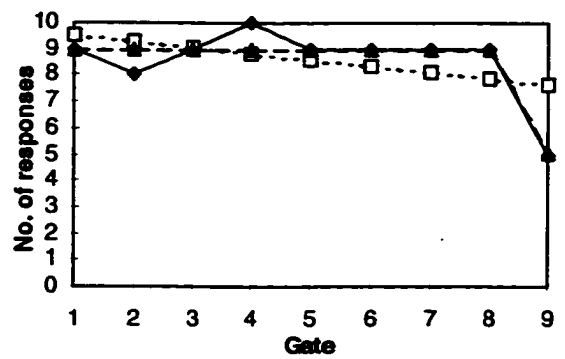
61 Betsy /ts/ L: 0.000



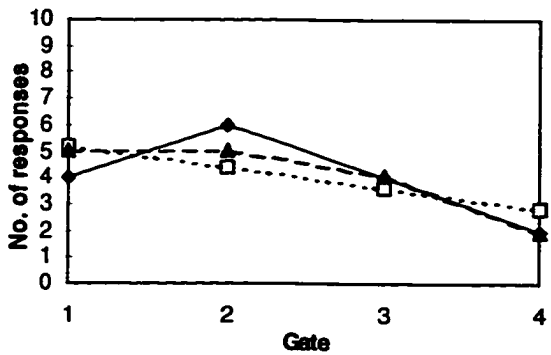
58 hacksaw /ks/ L: 4.168 O: 3.808 m: 4-5



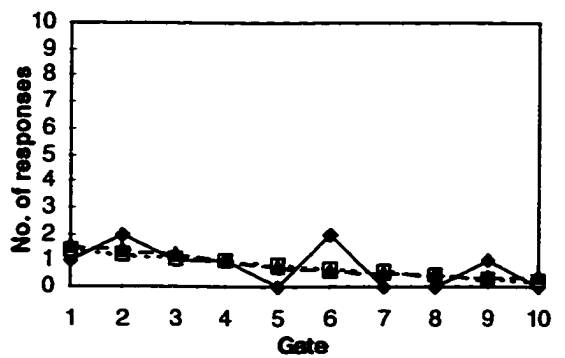
62 stop /st/ L: 3.599 O: 1.416 m: 8-9



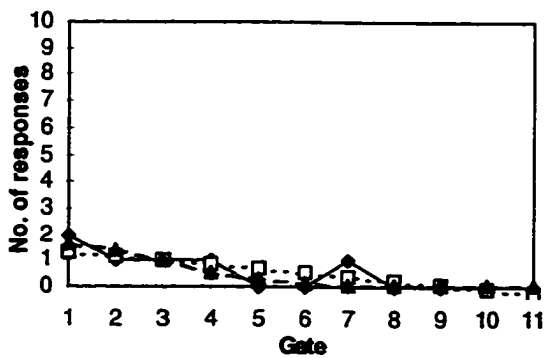
59 unacceptable /ks/ L: 2.191 O: 1.433 m: 3-4



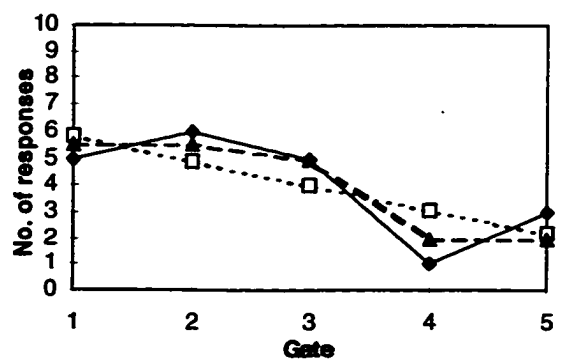
63 based /st/ L: 2.033 O: 2.048 m: 3-4



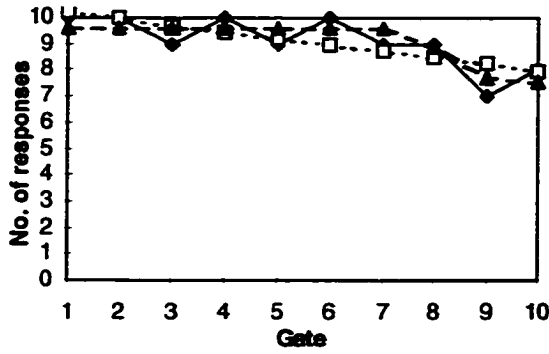
60 cats /ts/ L: 1.335 O: 1.173 m: 3-4



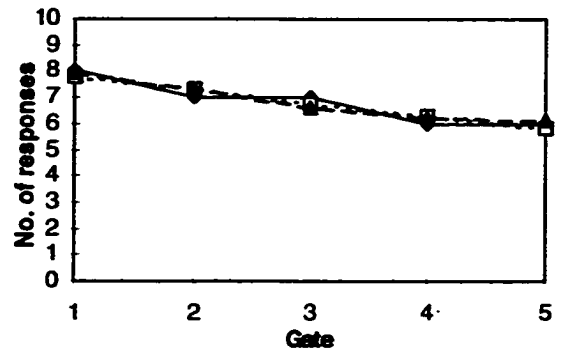
64 pastime /st/ L: 2.811 O: 1.582 m: 3-4



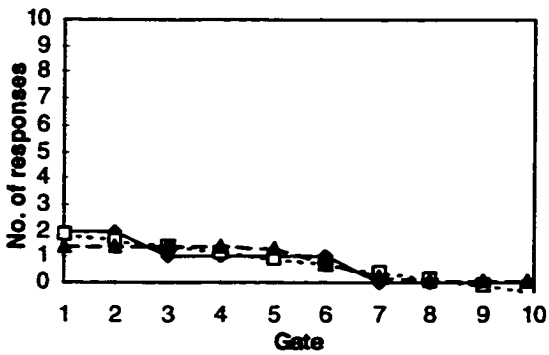
65 skate /sk/ L: 1.951 O: 1.544 m: 8-9



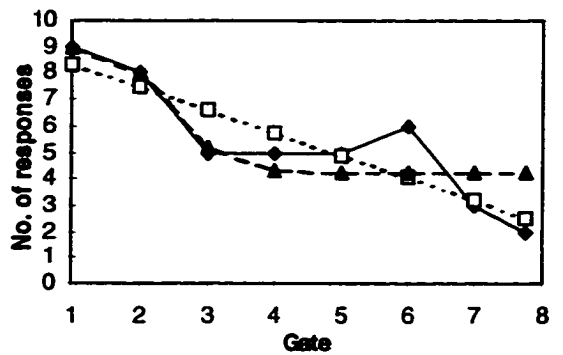
69 string /tr/ L: 0.548 O: 0.553 m: 2-3



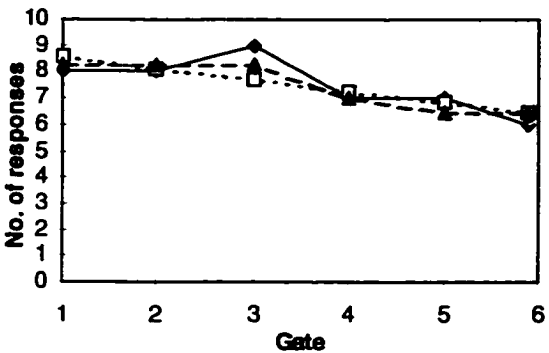
66 mask /sk/ L: 0.859 O: 1.122 m: 6-7



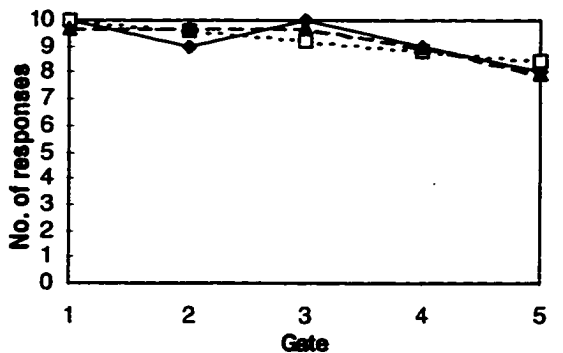
70 Detroit /tr/ L: 2.842 O: 3.293 m: 2-3



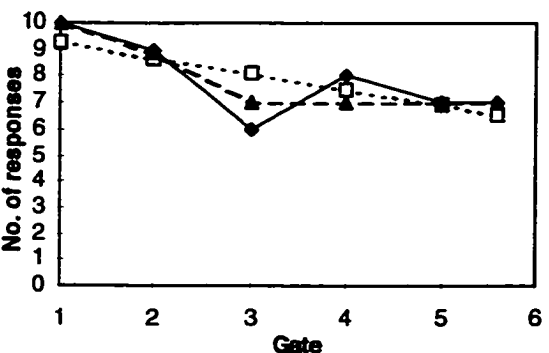
67 discount /sk/ L: 1.524 O: 1.102 m: 3-4



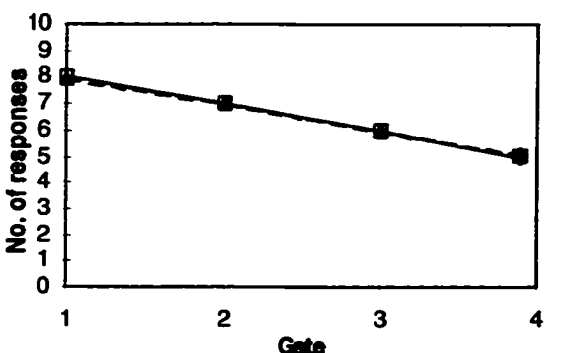
71 crops /kr/ L: 1.095 O: 0.824 m: 4-5



68 train /tr/ L: 2.336 O: 1.419 m: 2-3

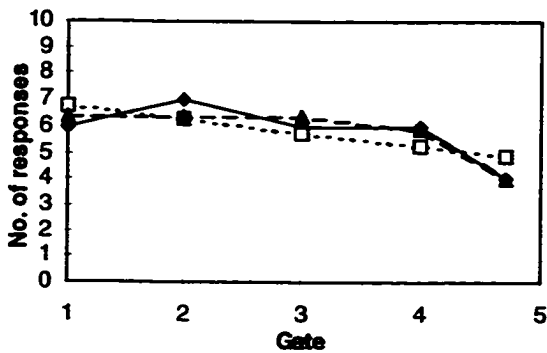


72 scrap /kr/ L: 0.056 O: 0.125 m: 2-3

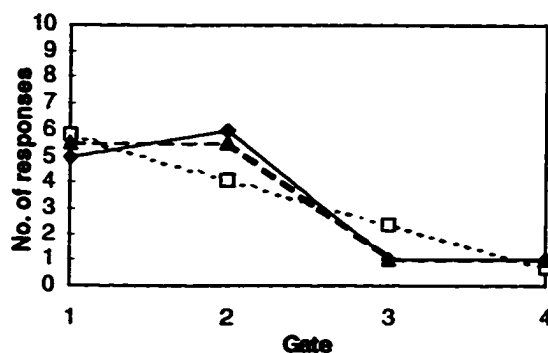




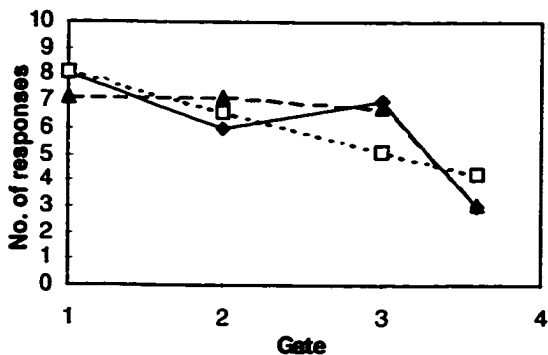
73 acrobat /kr/ L: 1.599 O: 0.821 m: 4-5



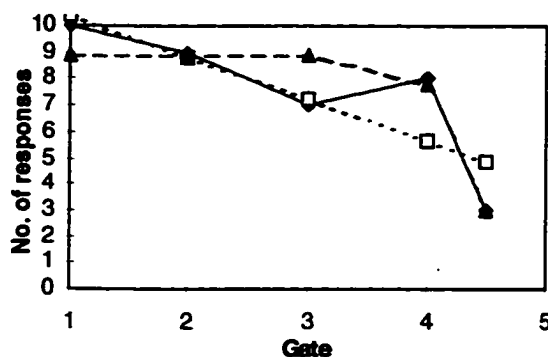
77 split /pl/ L: 2.510 O: 0.707 m: 2-3



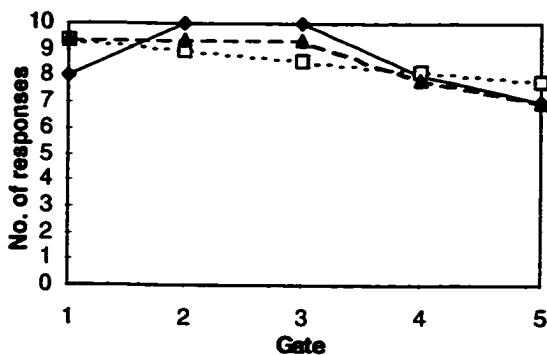
74 drop /dr/ L: 2.328 O: 1.457 m: 3-4



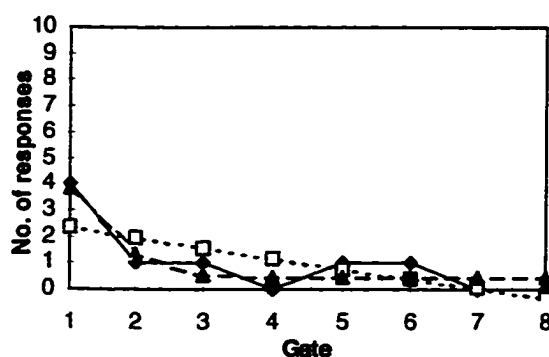
78 twelve /tw/ L: 3.036 O: 2.205 m: 4-5



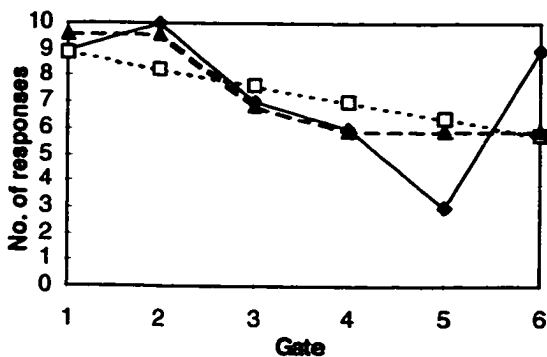
75 groan /gr/ L: 2.366 O: 1.639 m: 3-4



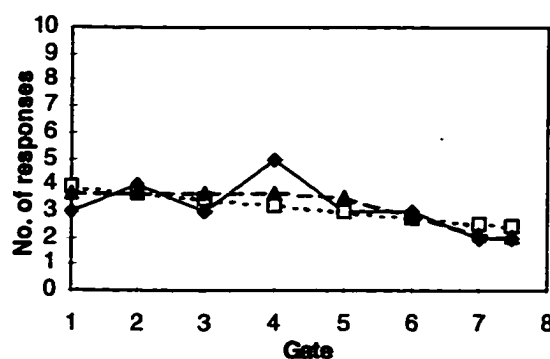
79 court /rt/ L: 2.430 O: 1.241 m: 1-2



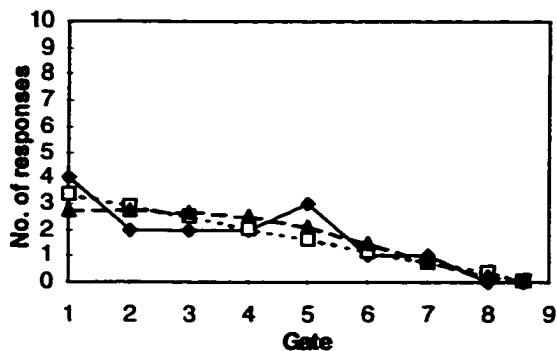
76 plain /pl/ L: 5.140 O: 4.312 m: 2-3



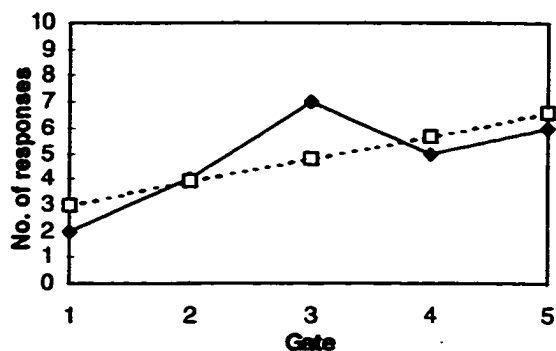
80 cork /rk/ L: 2.194 O: 1.764 m: 6-7



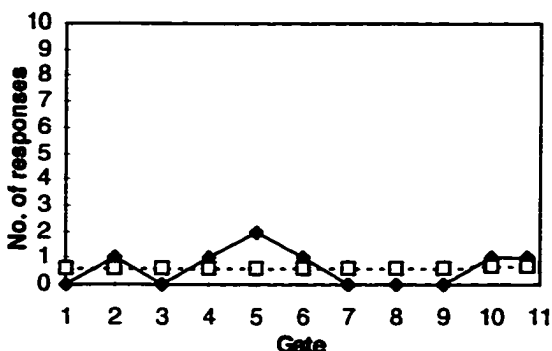
81 help /ɪp/ L: 1.898 O: 1.991 m: 6-7



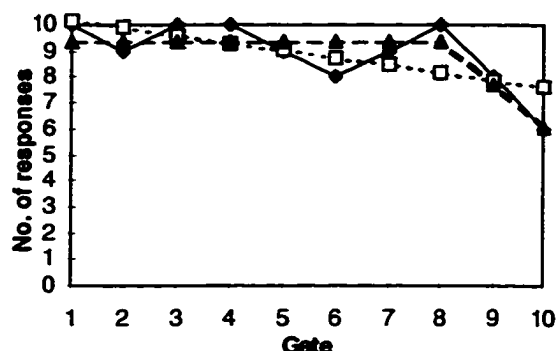
85 unconcealed /ns/ L: 2.588



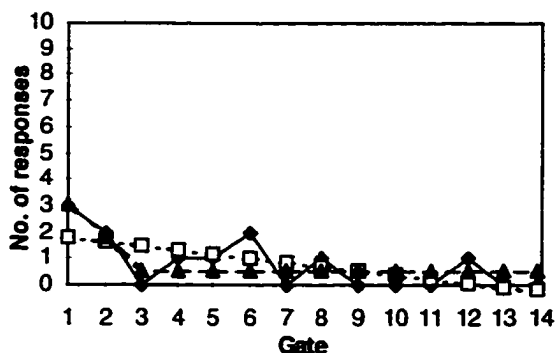
82 fans /nz/ L: 2.130



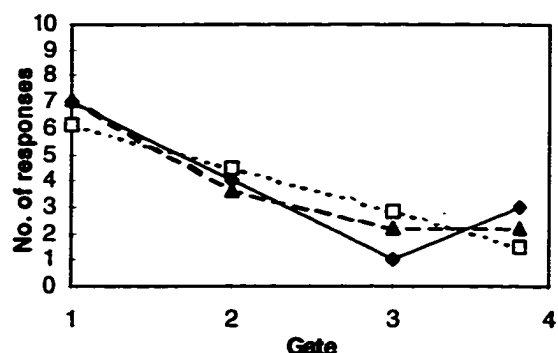
86 snow /sn/ L: 2.865 O: 1.991 m: 9-10



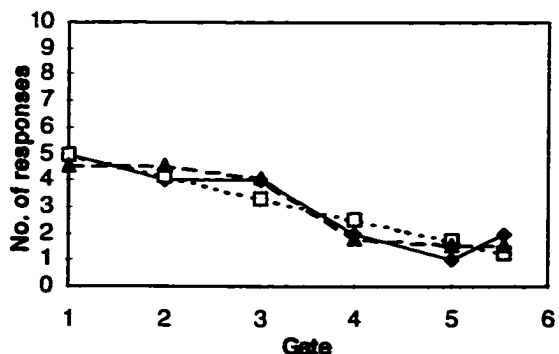
83 dance /ns/ L: 2.668 O: 2.250 m: 2-3



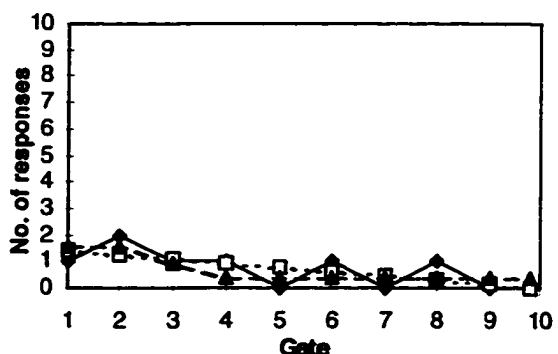
87 Disney /zn/ L: 2.560 O: 1.493 m: 1-2



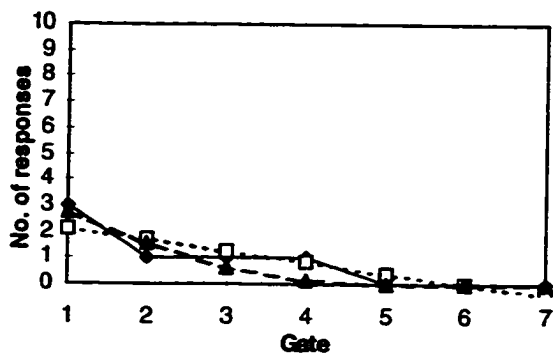
84 fancy /ns/ L: 1.336 O: 1.037 m: 3-4



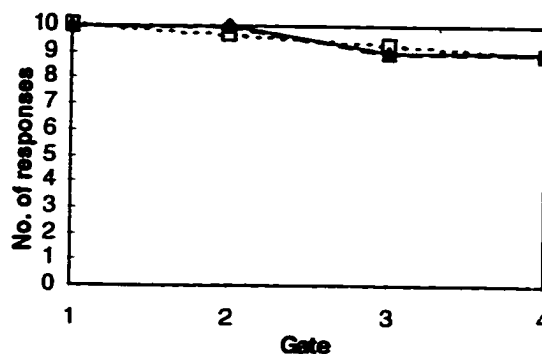
88 farm /rm/ L: 1.486 O: 1.494 m: 3-4



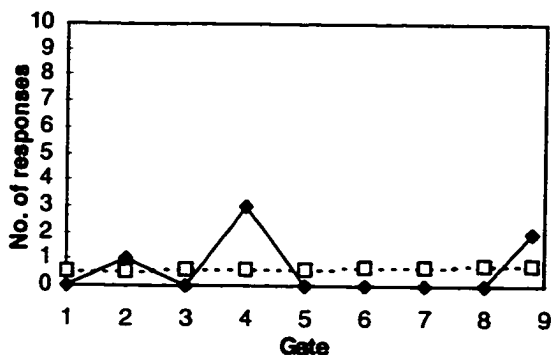
89 corn /rn/ L: 1.309 O: 1.097 m: 1-2



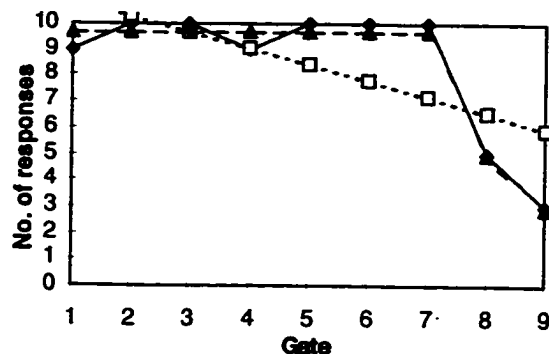
93 fragile /fr/ L: 0.447 O: 0.043 m: 2-3



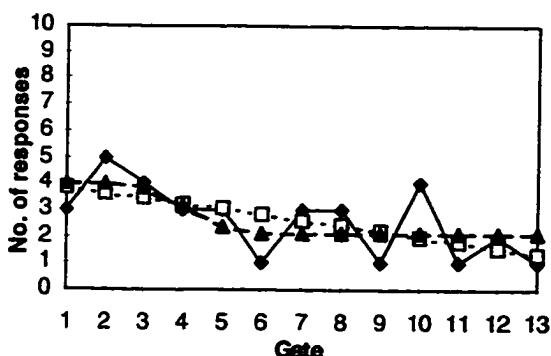
90 film /lm/ L: 3.154



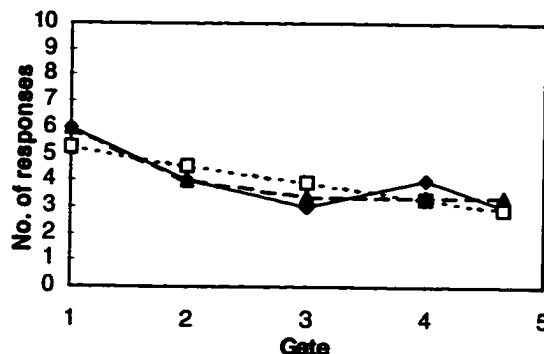
94 sleep /sl/ L: 5.491 O: 1.200 m: 7-8



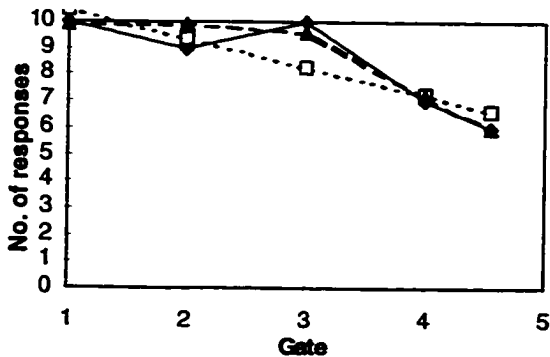
91 ranch /ntʃ/ L: 3.625 O: 3.550 m: 4-5



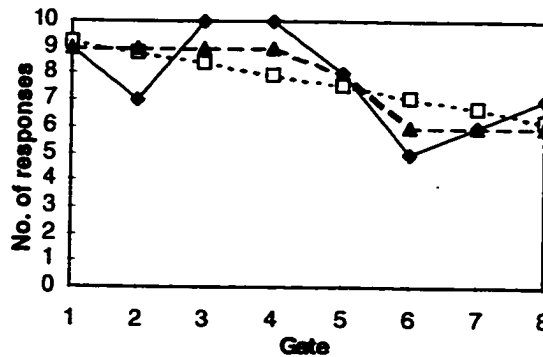
95 Iceland /sl/ L: 1.526 O: 0.820 m: 1-2



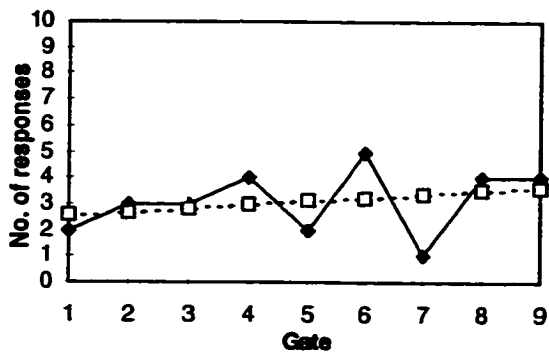
92 flash /fl/ L: 1.920 O: 0.989 m: 3-4



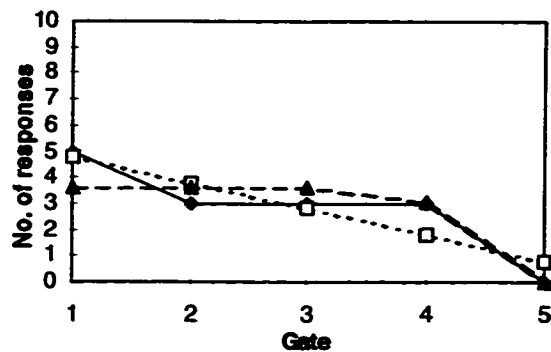
96 swan /sw/ L: 3.973 O: 2.829 m: 5-6



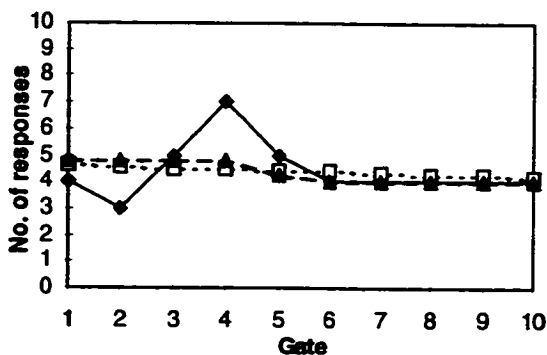
97 golf /l/ L: 3.438



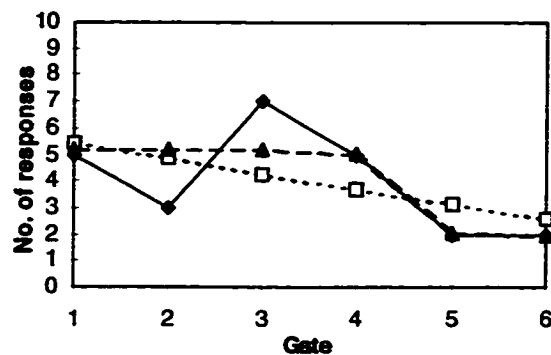
101 cultural /tʃ/ L: 1.673 O: 1.639 m: 4-5



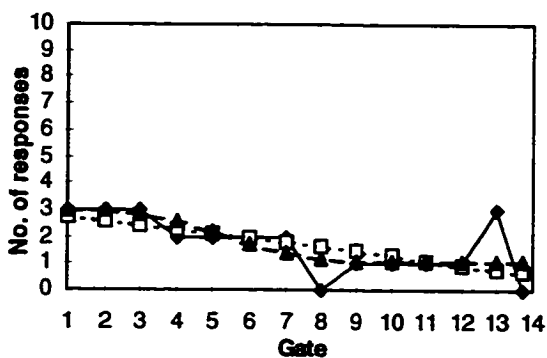
98 wharf /r/ L: 3.195 O: 3.062 m: 4-5



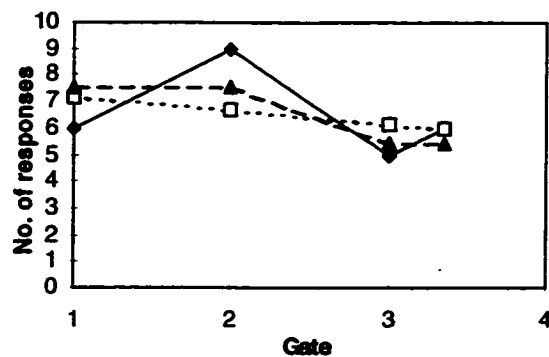
102 marginal /rdʒ/ L: 3.780 O: 2.844 m: 4-5



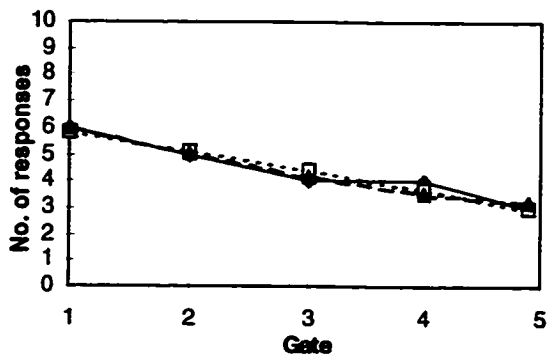
99 false /s/ L: 2.984 O: 2.673 m: 5-6



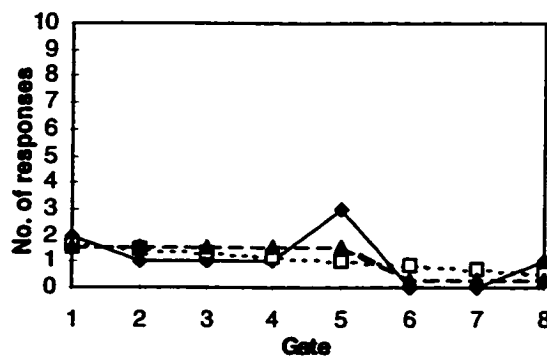
103 optical /pt/ L: 2.858 O: 2.239 m: 2-3



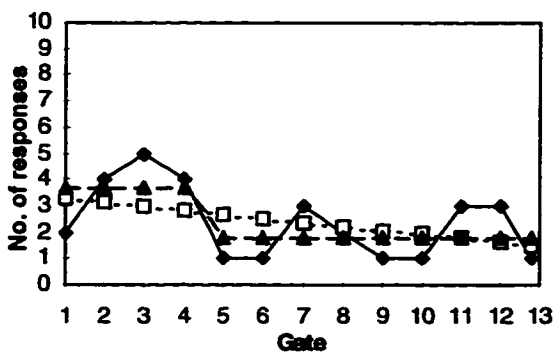
100 calcium /s/ L: 0.550 O: 0.554 m: 2-3



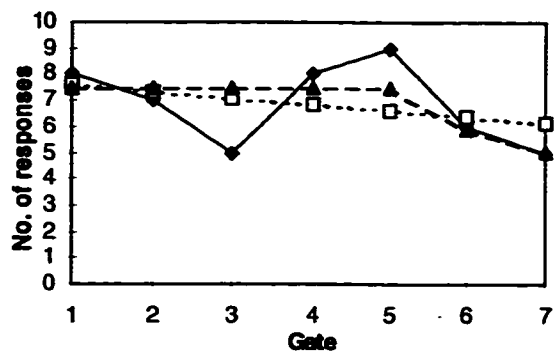
104 pact /kt/ L: 2.423 O: 1.967 m: 5-6



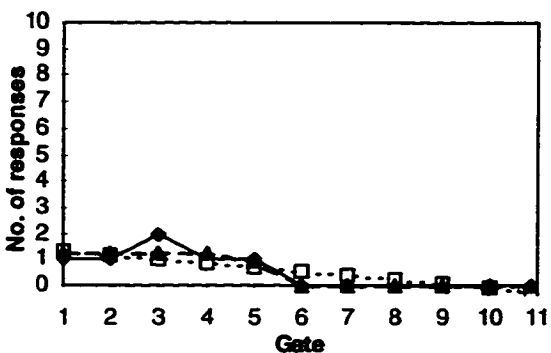
105 coughs /fs/ L: 4.371 O: 3.511 m: 4-5



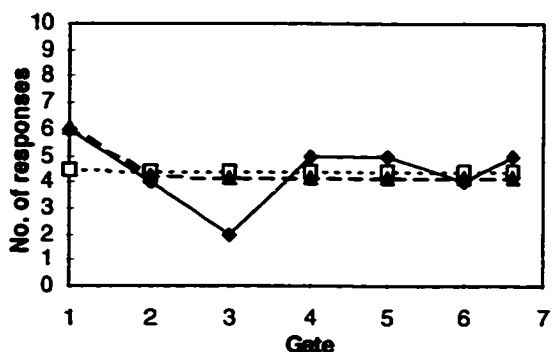
109 biopsy /a<sup>h</sup>a/ L: 3.620 O: 3.044 m: 5-6



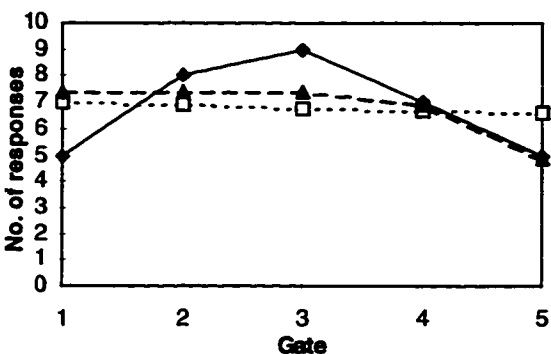
106 nerves /vz/ L: 1.330 O: 0.869 m: 5-6



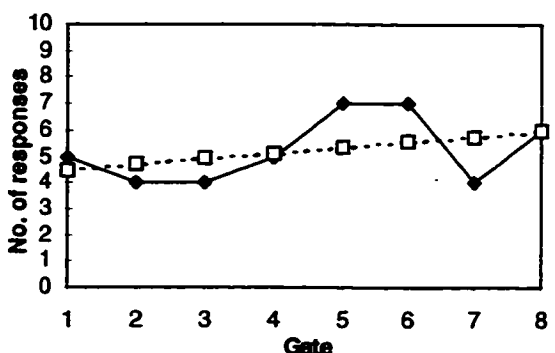
110 biography /a<sup>h</sup>a/ L: 3.116 O: 2.619 m: 1-2



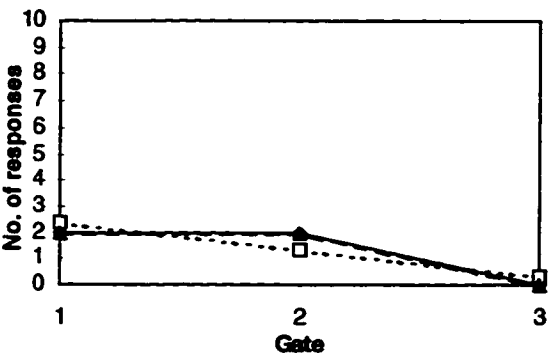
107 amnesty /mn/ L: 3.564 O: 2.965 m: 4-5



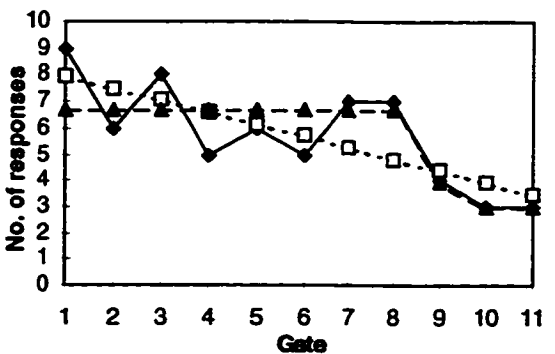
111 biotech /a<sup>h</sup>o<sup>w</sup>/ L: 3.094



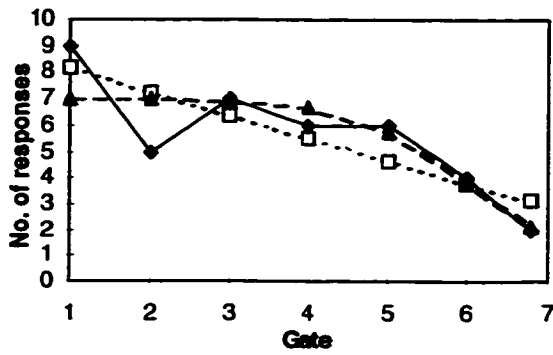
108 garlic /rl/ L: 0.816 O: 0.004 m: 2-3



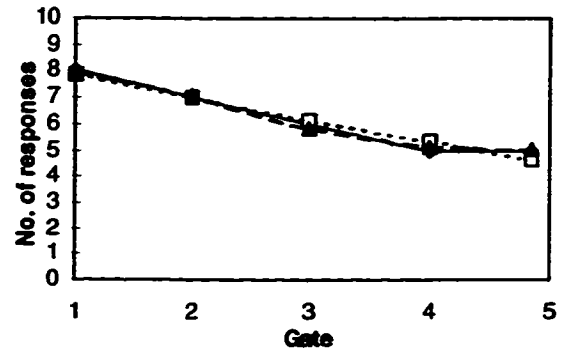
112 eon /ia/ L: 4.044 O: 3.734 m: 8-9



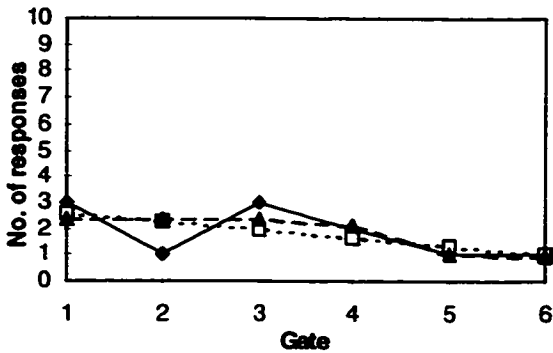
113 diagonal /a<sup>j</sup>æ/ L: 3.083 O: 2.932 m: 6-7



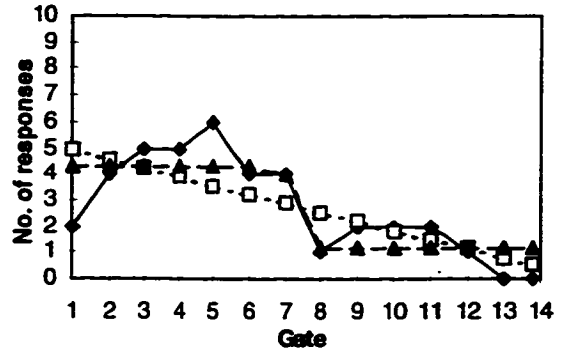
117 data /de<sup>j</sup>/ L: 0.554 O: 0.262 m: 2-3



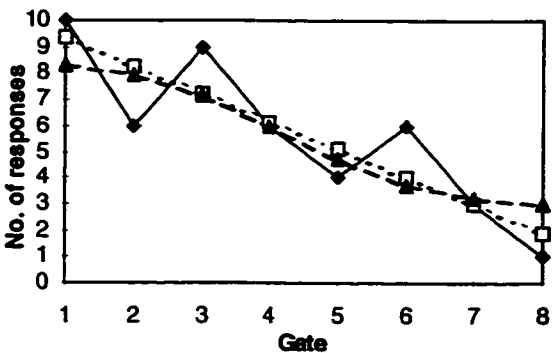
114 react /iæ/ L: 1.762 O: 1.640 m: 4-5



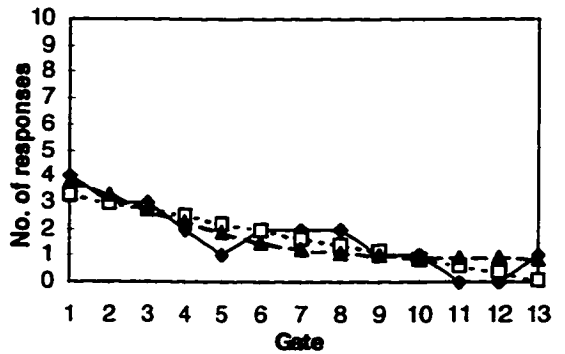
118 fade /e<sup>j</sup>d/ L: 4.712 O: 3.771 m: 7-8



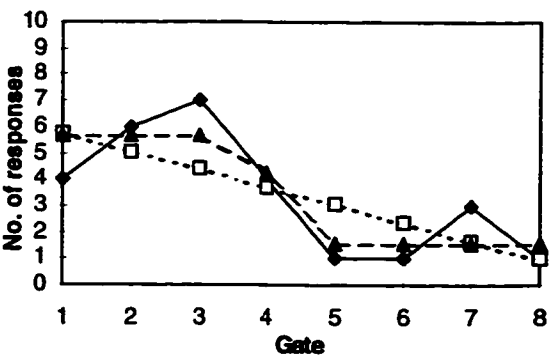
115 tiger /ta<sup>j</sup>/ L: 3.837 O: 4.445 m: 4-5



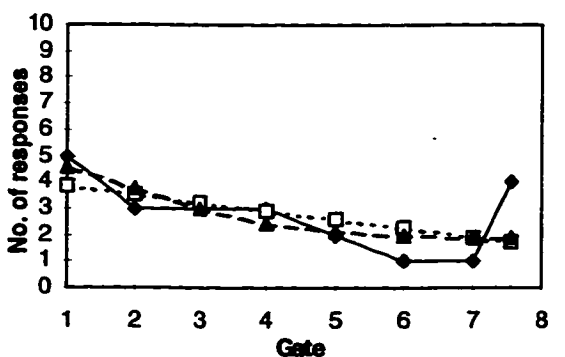
119 doubt /a<sup>w</sup>ʊ/ L: 2.027 O: 2.061 m: 2-3



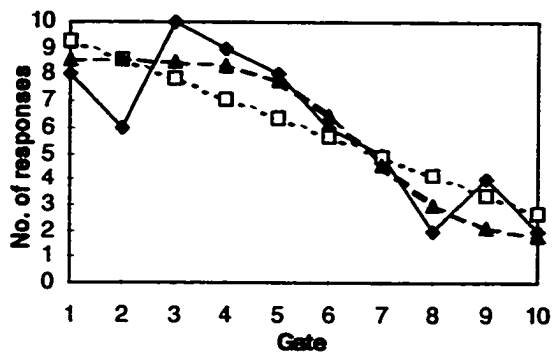
116 bite /a<sup>j</sup>t/ L: 4.305 O: 2.790 m: 4-5



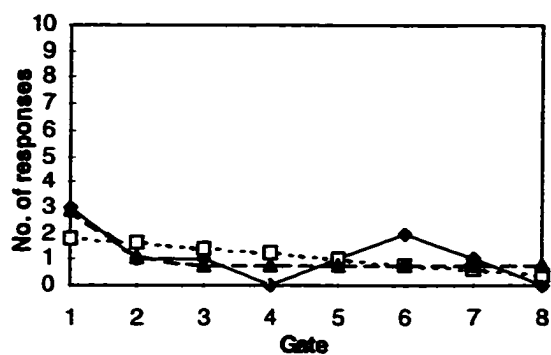
120 soybean /o<sup>j</sup>b/ L: 3.075 O: 2.691 m: 1-2



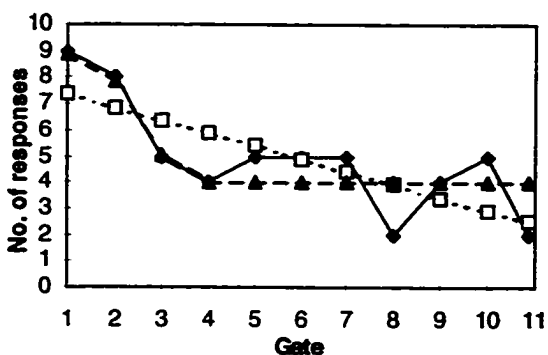
121 toad /toʷ/ L: 4.990 O: 3.794 m: 6-7



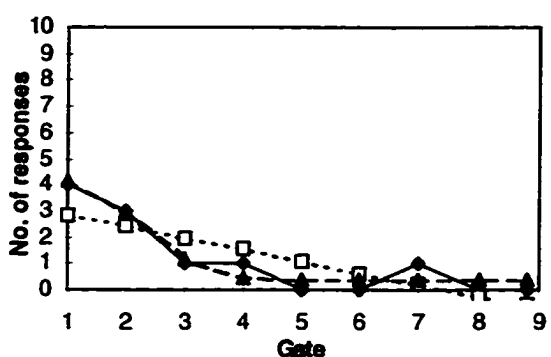
125 button /tʌ/ L: 2.270 O: 1.691 m: 1-2



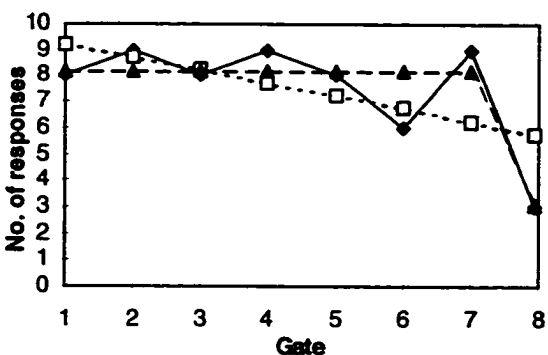
122 oats /oʷt/ L: 4.298 O: 3.469 m: 2-3



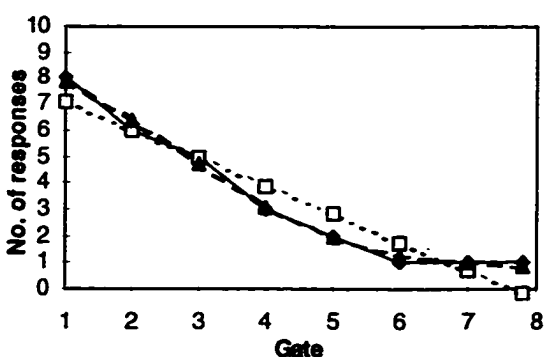
126 beetle /t/ L: 2.349 O: 1.153 m: 2-3



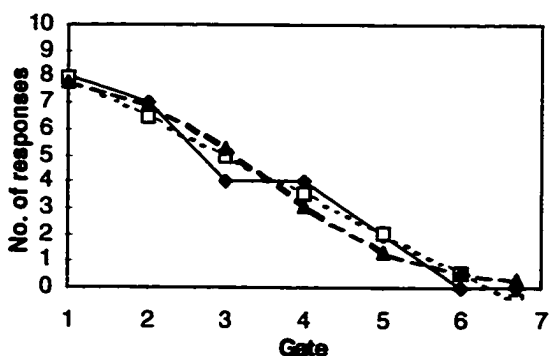
123 courage /kə/ L: 4.431 O: 2.623 m: 7-8



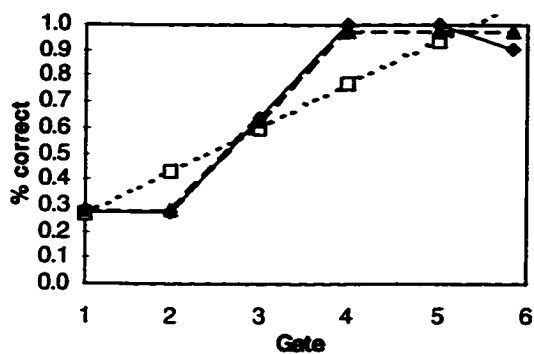
127 apple /p/ L: 2.078 O: 0.621 m: 2-3



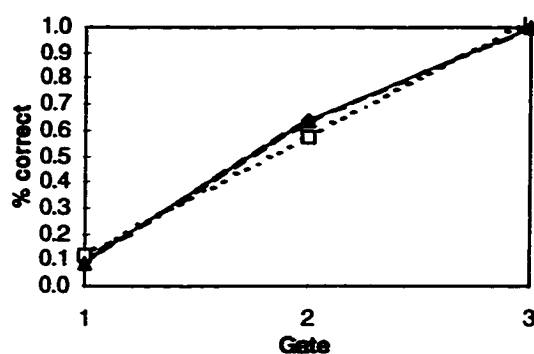
124 circle /ə/ L: 1.429 O: 1.791 m: 3-4



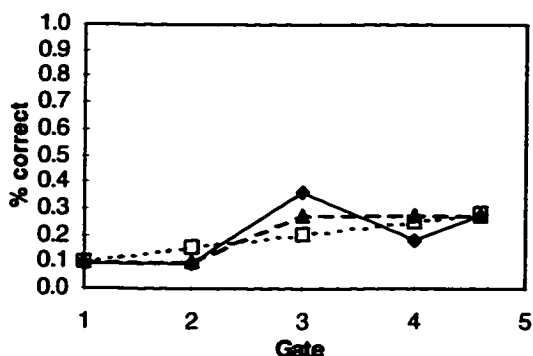
1 tip /tɪ/ L: 0.337 O: 0.076 m: 3-4



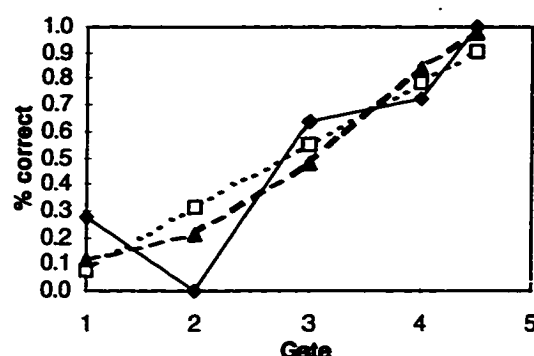
5 attic /tɪ/ L: 0.074 O: 0.007 m: 1-2



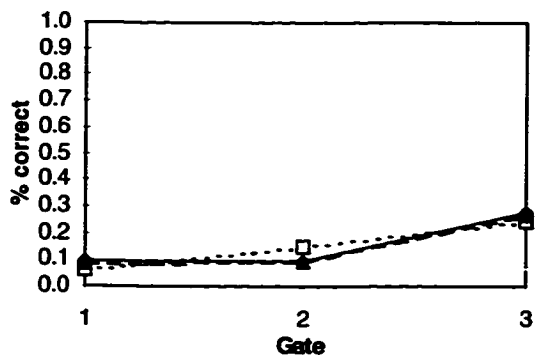
2 stiff /tɪ/ L: 0.187 O: 0.129 m: 2-3



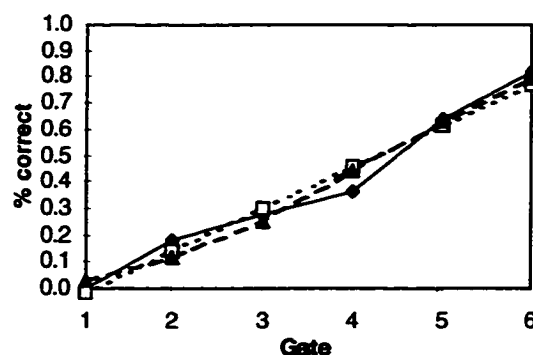
6 custom /kʌ/ L: 0.396 O: 0.327 m: 3-4



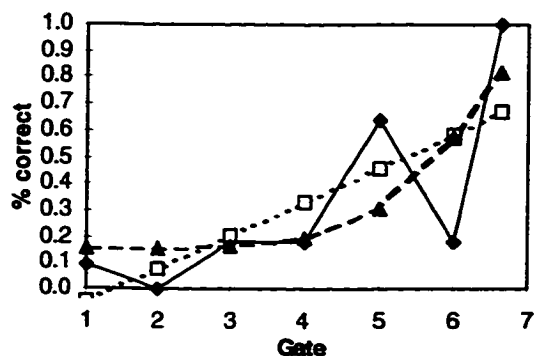
3 Tibet /tɪ/ L: 0.074 O: 0.004 m: 2-3



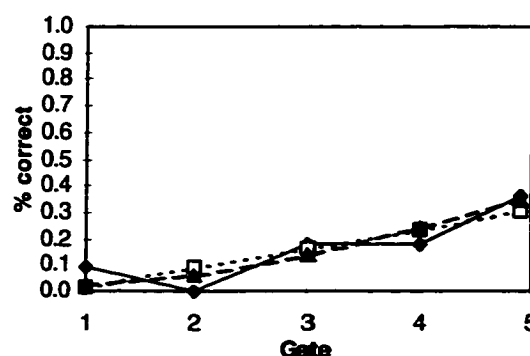
7 skull /kʌ/ L: 0.118 O: 0.111 m: 4-5



4 petition /tɪ/ L: 0.594 O: 0.570 m: 6-7

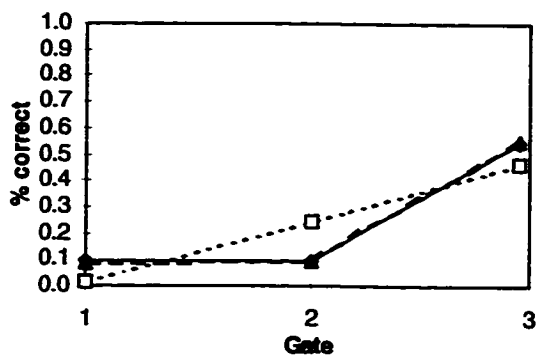


8 accompany /kʌ/ L: 0.144 O: 0.120 m: 4-5

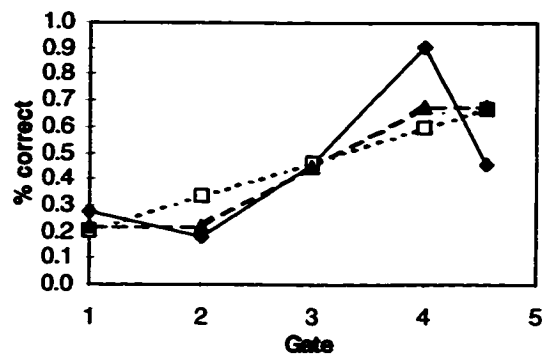




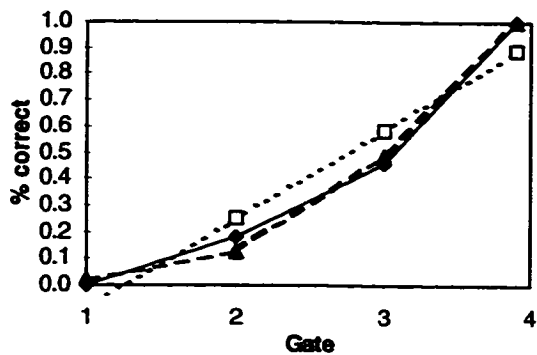
9 caboose /kə/ L: 0.190 O: 0.005 m: 2-3



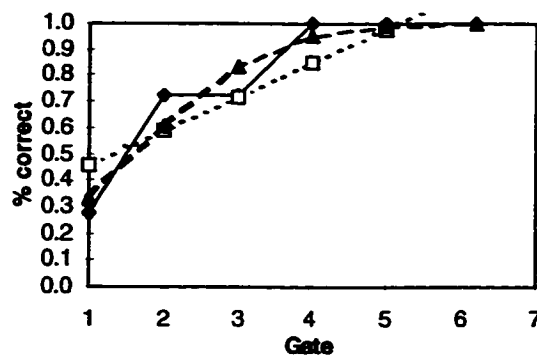
13 fitness /t/ L: 0.414 O: 0.328 m: 3-4



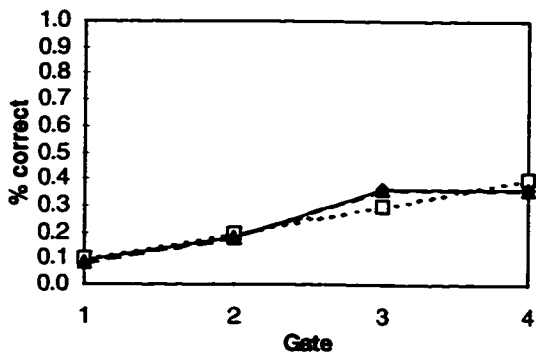
10 academic /kə/ L: 0.204 O: 0.066 m: 3-4



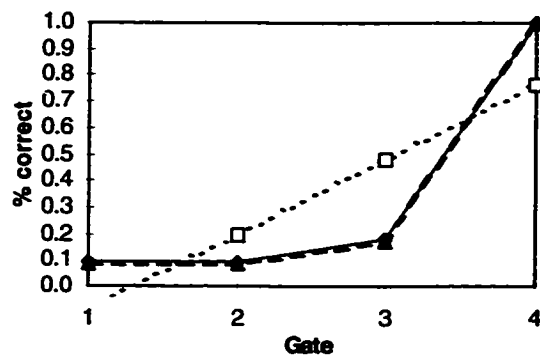
14 Italian /t/ L: 0.309 O: 0.178 m: 1-2



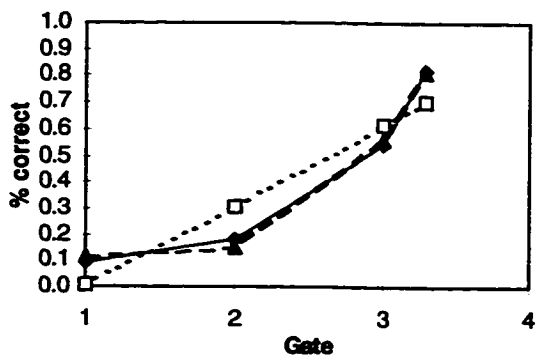
11 duck /dʌ/ L: 0.076 O: 0.003 m: 2-3



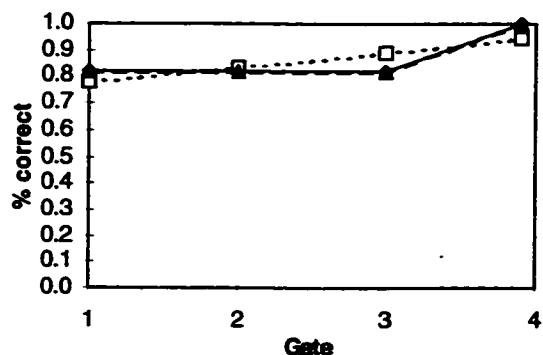
15 committee /t/ L: 0.433 O: 0.007 m: 3-4



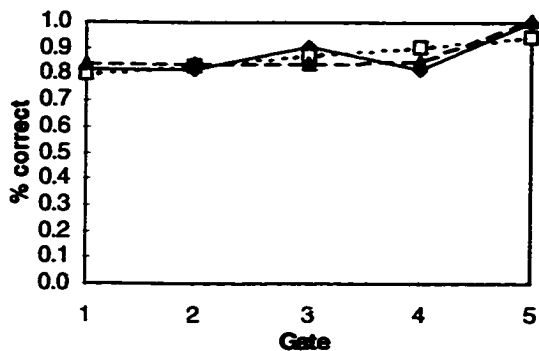
12 citizen /t/ L: 0.203 O: 0.048 m: 3-4



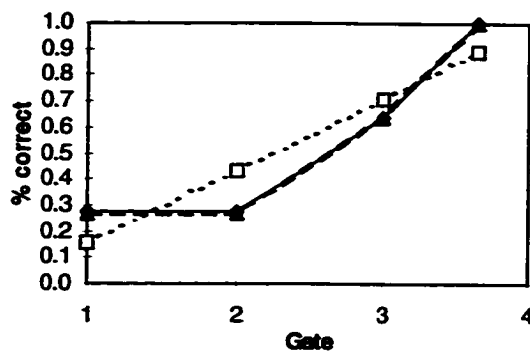
16 unity /t/ L: 0.103 O: 0.003 m: 3-4



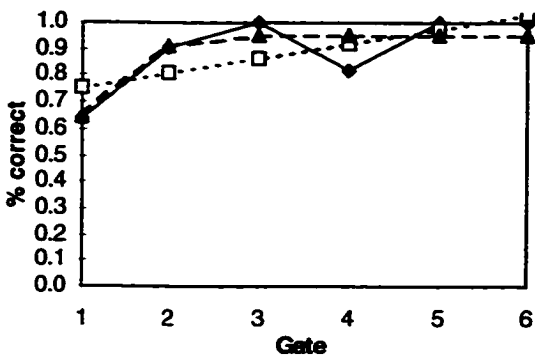
17 bucket /ʌk/ L: 0.115 O: 0.082 m: 4-5



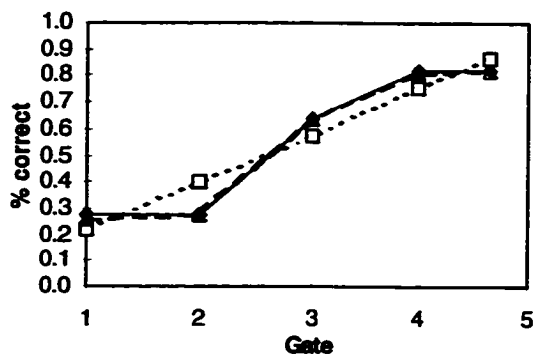
21 muddy /ʌd/ L: 0.238 O: 0.004 m: 3-4



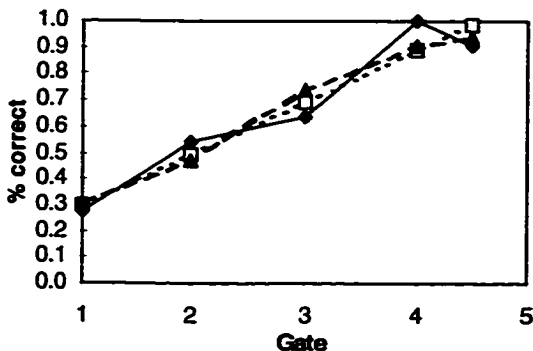
18 mechanical /æk/ L: 0.232 O: 0.158 m: 1-2



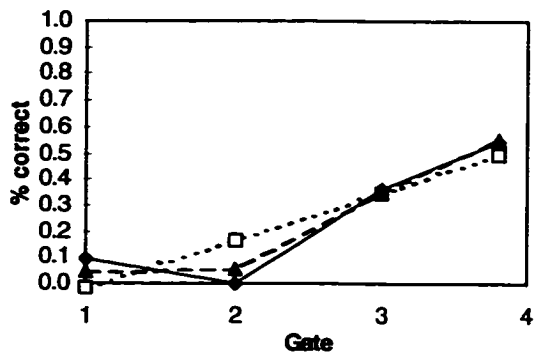
22 cadenza /əd/ L: 0.170 O: 0.015 m: 2-3



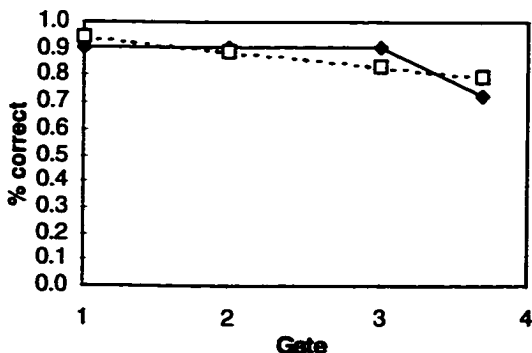
19 indicate /ək/ L: 0.157 O: 0.162 m: 2-3



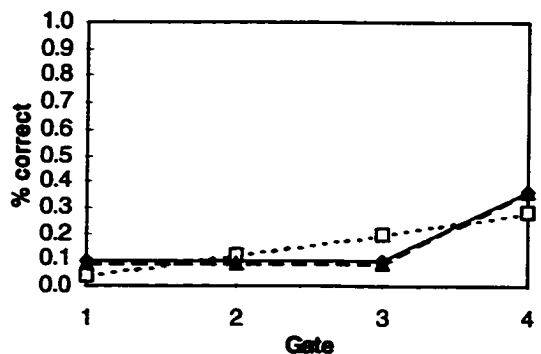
23 medicine /mɛ/ L: 0.205 O: 0.069 m: 2-3

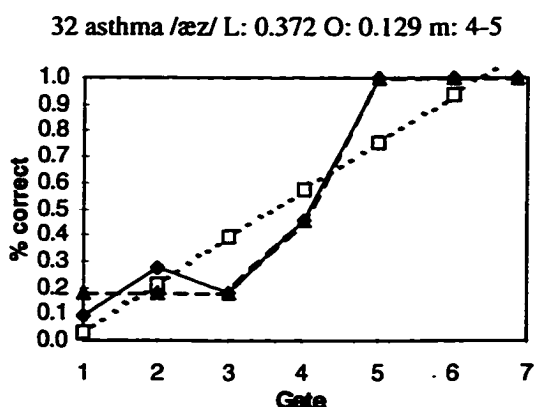
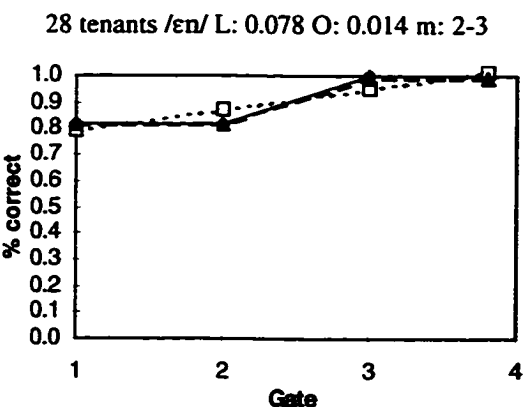
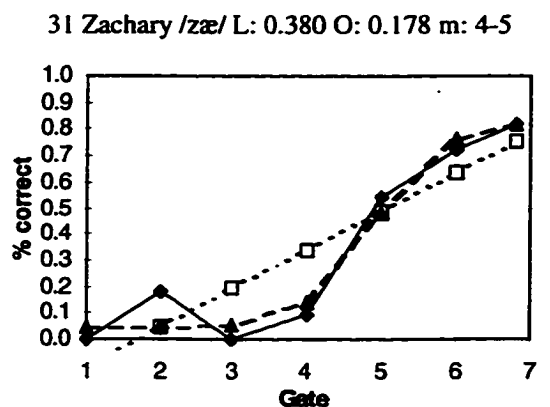
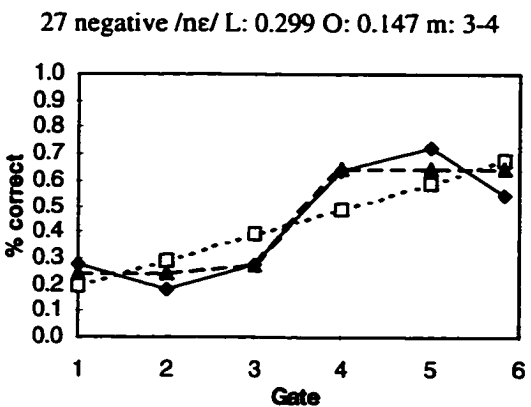
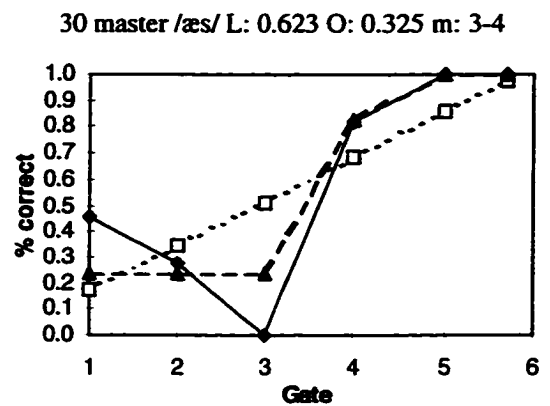
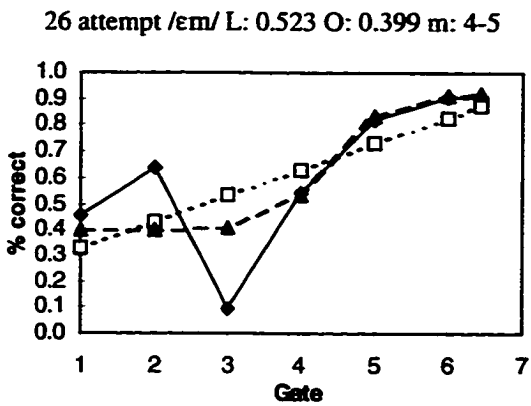
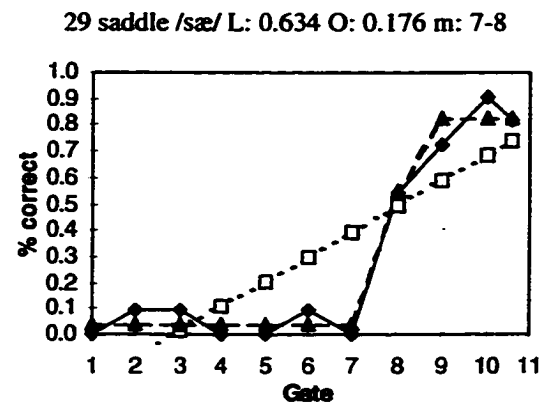
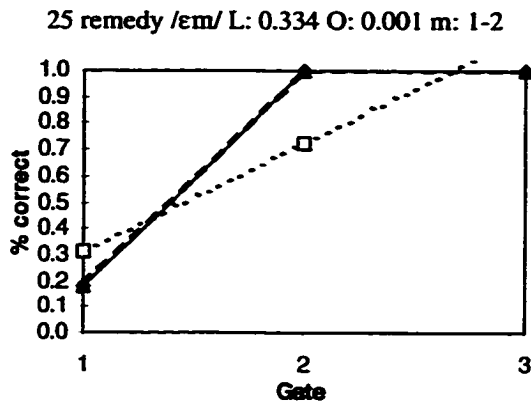


20 induction /ʌk/ L: 0.109

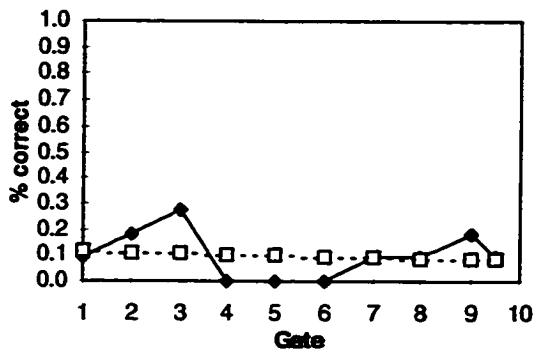


24 immense /mɛ/ L: 0.149 O: 0.002 m: 3-4

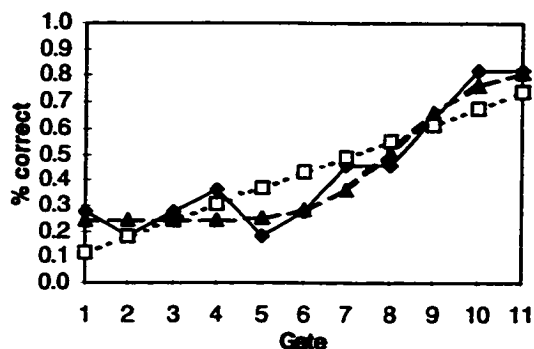




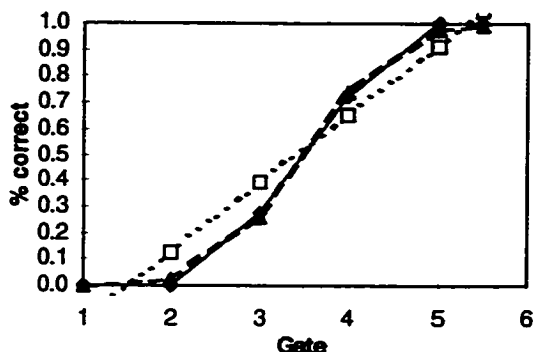
33 shell /ʃe/ L: 0.269



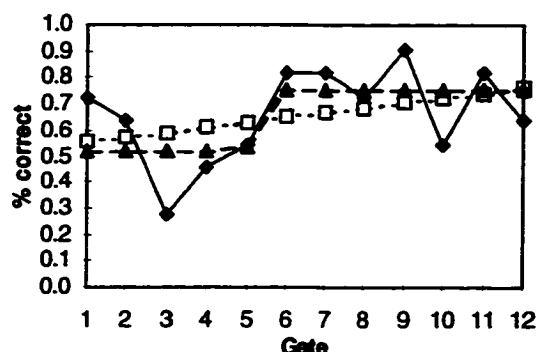
37 leaf /if/ L: 0.353 O: 0.197 m: 8-9



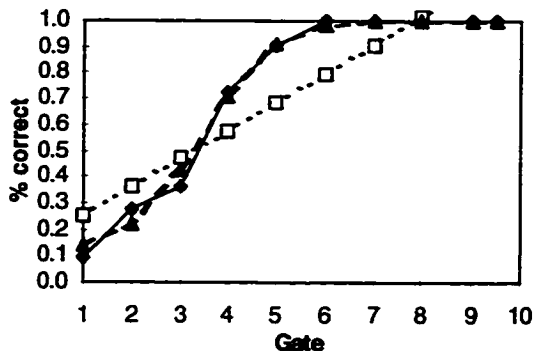
34 session /ɛʃ/ L: 0.251 O: 0.042 m: 3-4



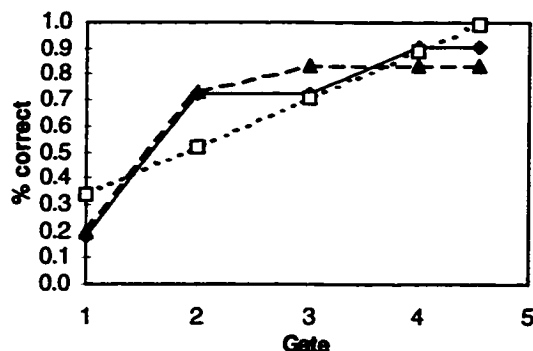
38 relief /if/ L: 0.562 O: 0.465 m: 5-6



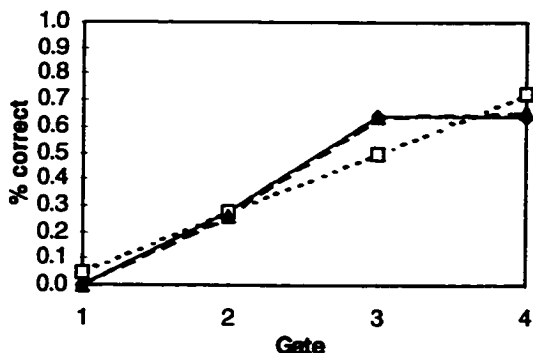
35 fees /fi/ L: 0.462 O: 0.099 m: 3-4



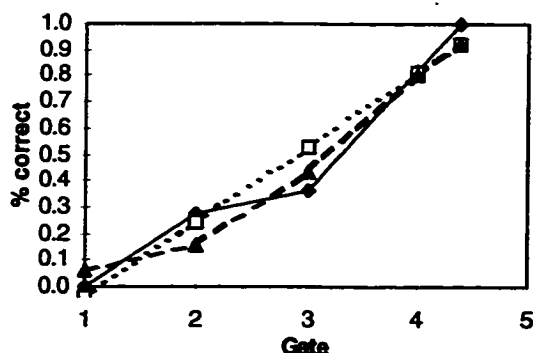
39 vacuum /væ/ L: 0.272 O: 0.150 m: 1-2

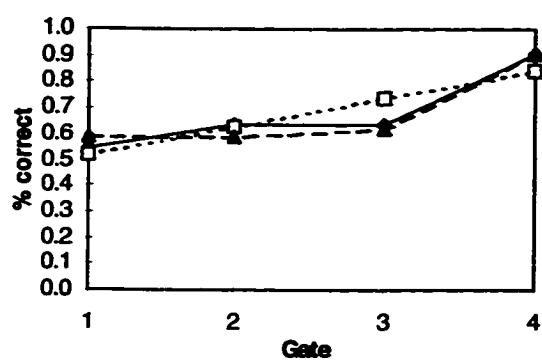


36 unfeeling /fi/ L: 0.170 O: 0.017 m: 2-3

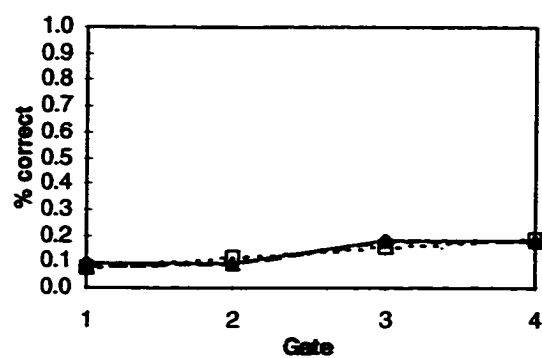
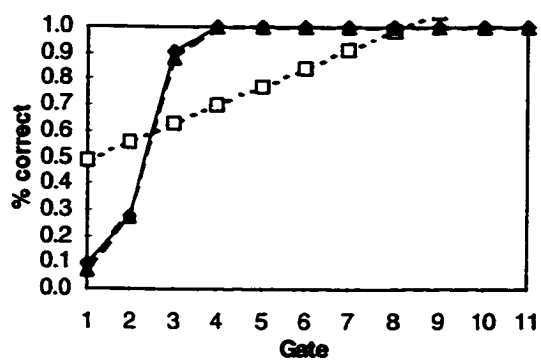


40 ravish /æv/ L: 0.187 O: 0.168 m: 3-4

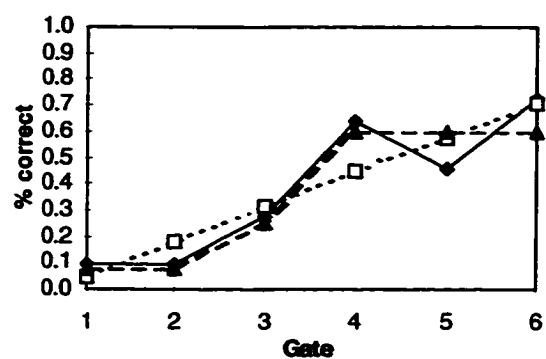


41 trail /re<sup>i</sup>/ L: 0.122 O: 0.066 m: 3-4

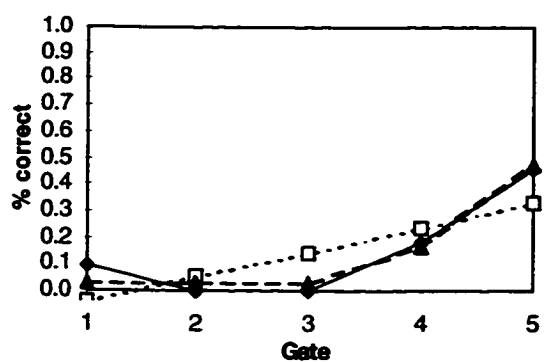
45 yellow /jε/ L: 0.041 O: 0.003 m: 2-3

42 fair /e<sup>i</sup>r/ L: 0.741 O: 0.033 m: 2-3

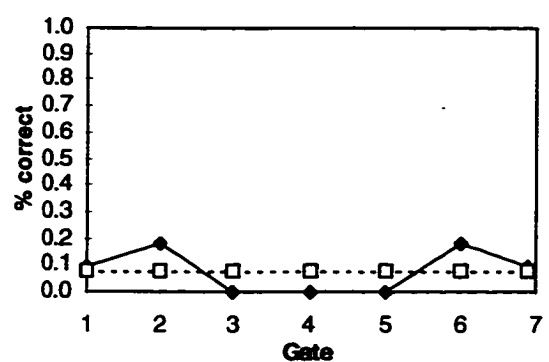
46 watch /wə/ L: 0.252 O: 0.198 m: 3-4



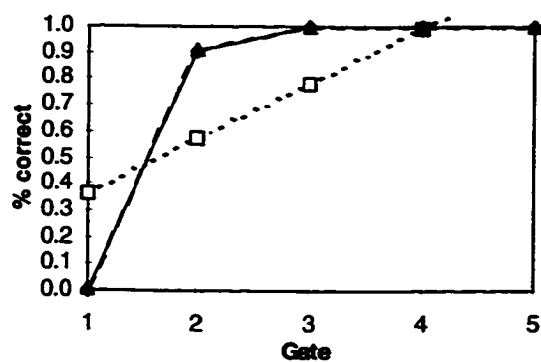
43 lever /ɛ/ L: 0.244 O: 0.077 m: 4-5



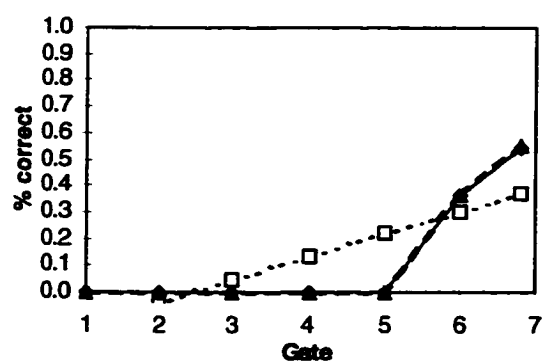
47 chapel /tʃæ/ L: 0.200



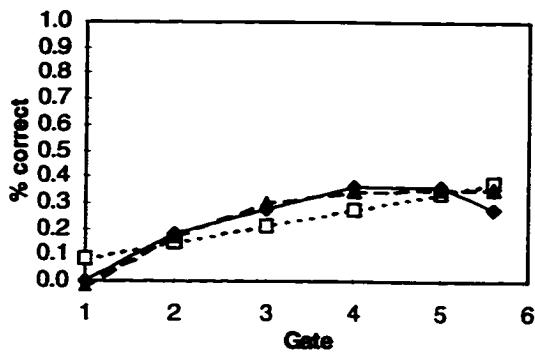
44 elevator /ɛl/ L: 0.577 O: 0.006 m: 1-2



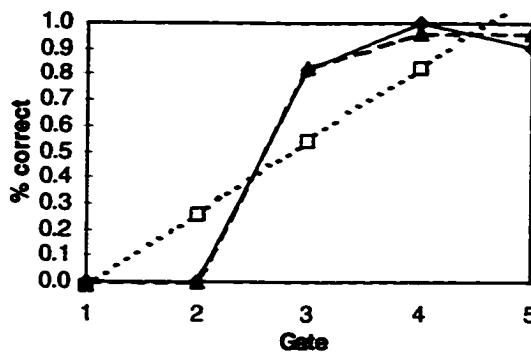
48 latches /ætʃ/ L: 0.343 O: 0.006 m: 5-6



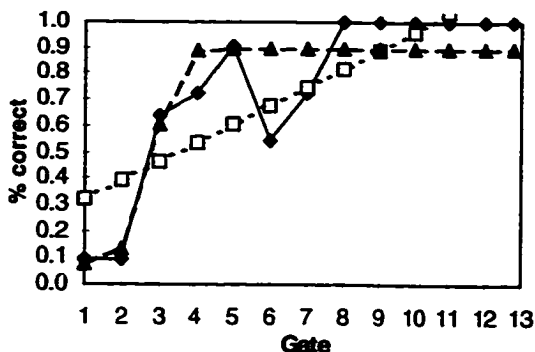
49 jump /dʒʌ/ L: 0.177 O: 0.092 m: 1-2



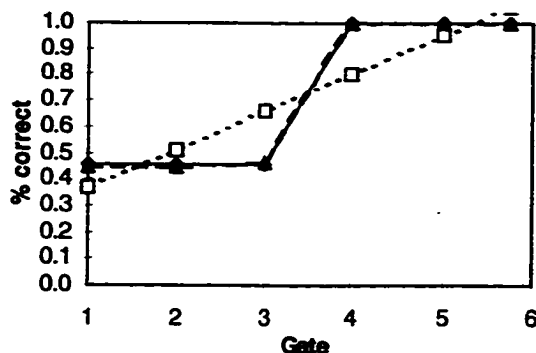
53 reinterpret /nt/ L: 0.463 O: 0.065 m: 2-3



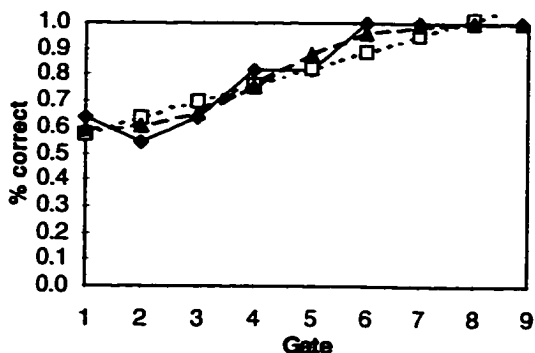
50 judge /ʌdʒ/ L: 0.636 O: 0.494 m: 2-3



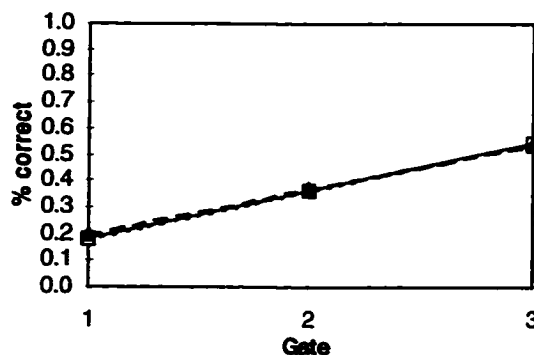
54 band /nd/ L: 0.311 O: 0.010 m: 3-4



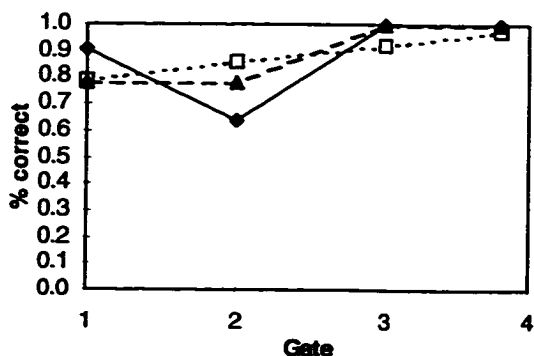
51 bent /nt/ L: 0.198 O: 0.122 m: 4-5



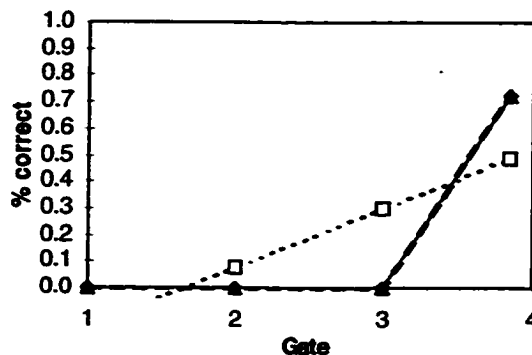
55 wander /nd/ L: 0.000 O: 0.014 m: 1-2



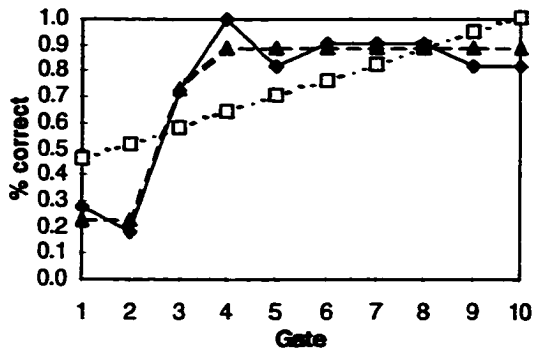
52 sentiment /nt/ L: 0.263 O: 0.193 m: 2-3



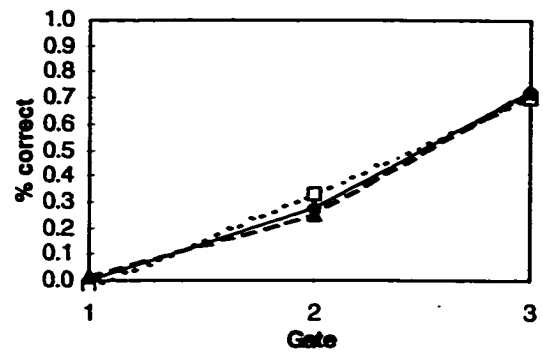
56 reconditioned /nd/ L: 0.417 O: 0.003 m: 3-4



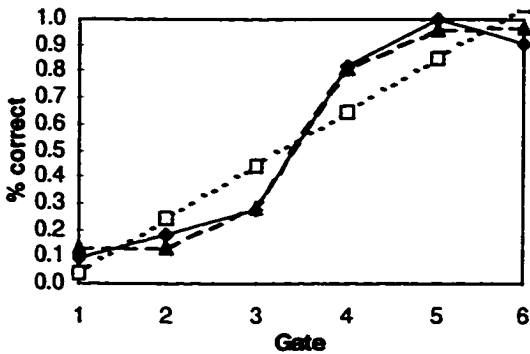
57 axe /ks/ L: 0.627 O: 0.181 m: 2-3



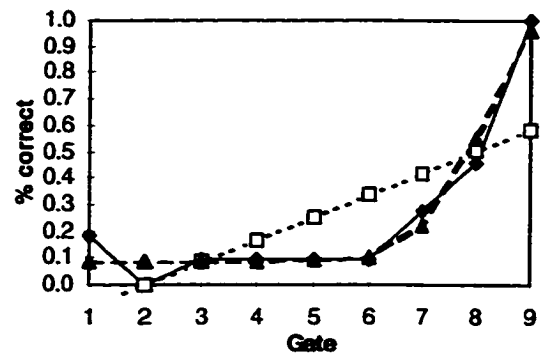
61 Betsy /ts/ L: 0.074 O: 0.025 m: 2-3



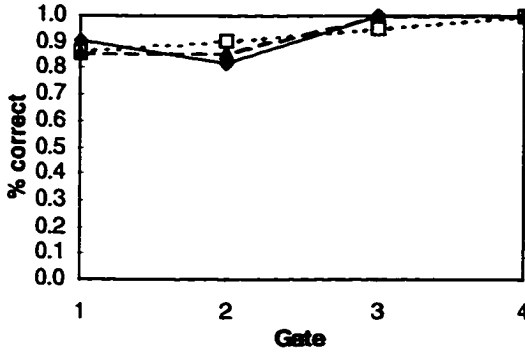
58 hacksaw /ks/ L: 0.329 O: 0.093 m: 3-4



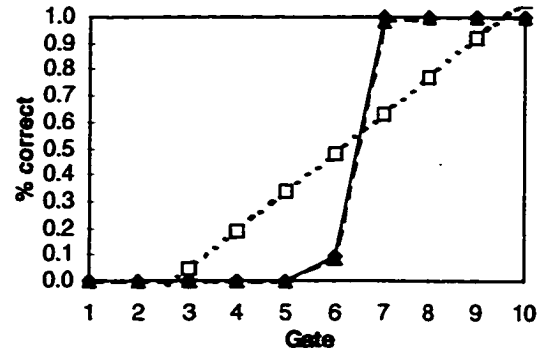
62 stop /st/ L: 0.597 O: 0.177 m: 8-9



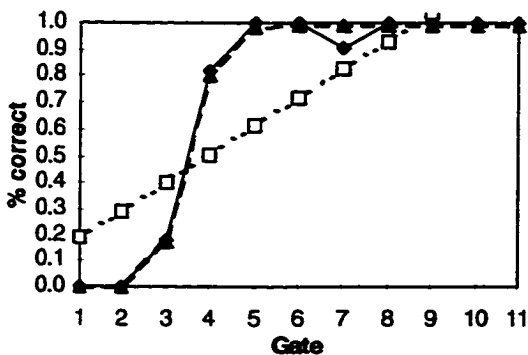
59 unacceptable /ks/ L: 0.111 O: 0.064 m: 2-3



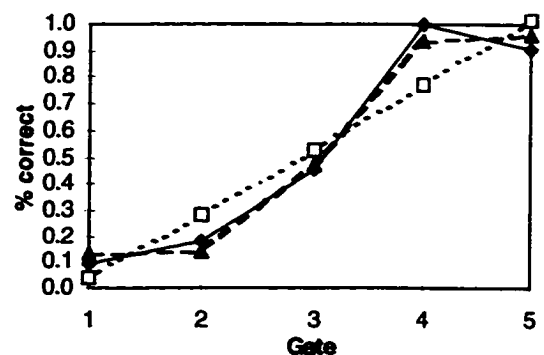
63 based /st/ L: 0.759 O: 0.014 m: 6-7



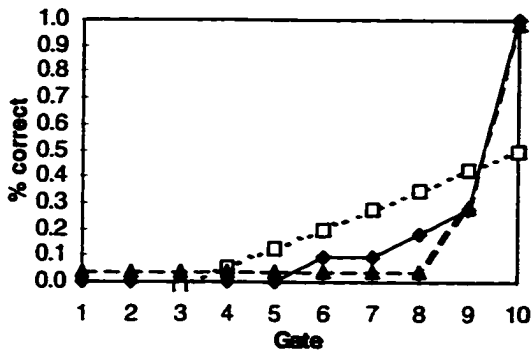
60 cats /ts/ L: 0.769 O: 0.088 m: 3-4



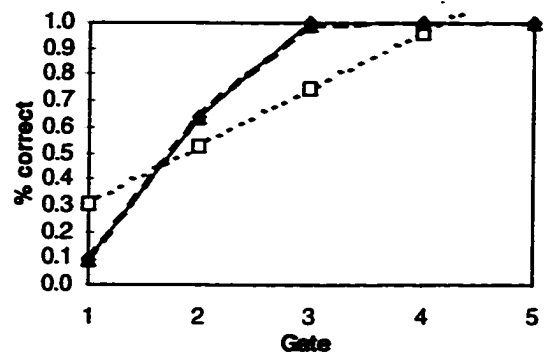
64 pastime /st/ L: 0.286 O: 0.103 m: 3-4



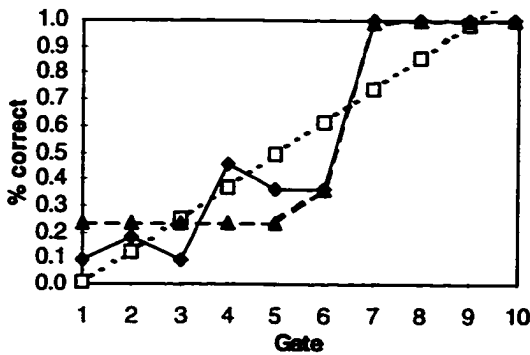
65 skate /sk/ L: 0.638 O: 0.181 m: 9-10



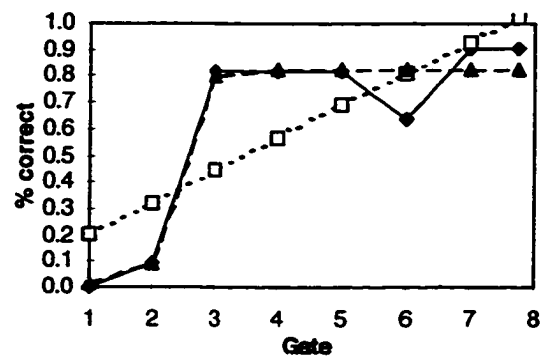
69 string /tr/ L: 0.398 O: 0.008 m: 1-2



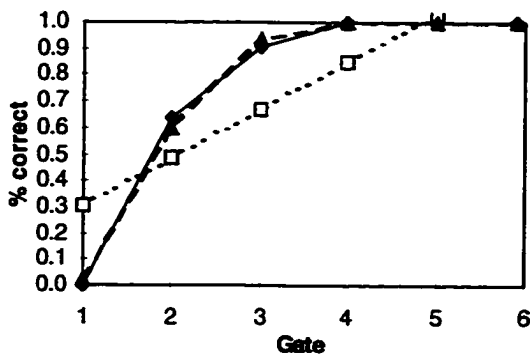
66 mask /sk/ L: 0.469 O: 0.331 m: 6-7



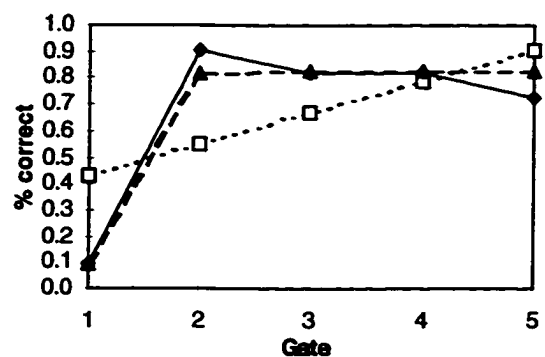
70 Detroit /tr/ L: 0.598 O: 0.224 m: 2-3



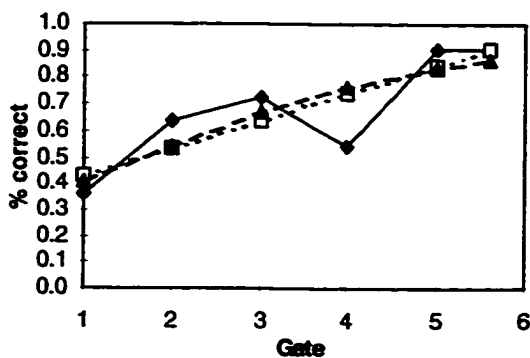
67 discount /sk/ L: 0.485 O: 0.056 m: 1-2



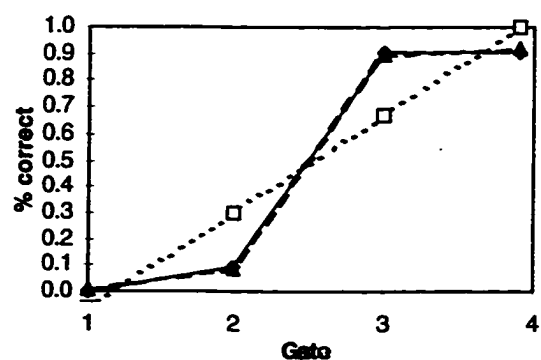
71 crops /kr/ L: 0.548 O: 0.132 m: 1-2



68 train /tr/ L: 0.256 O: 0.264 m: 1-2

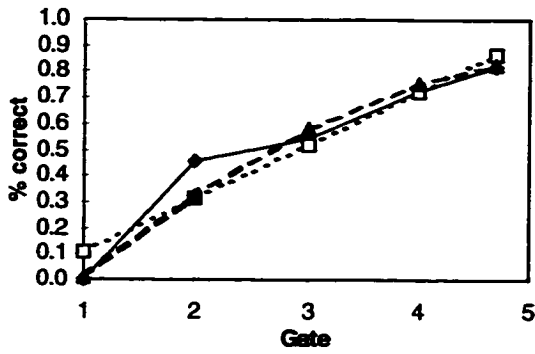


72 scrap /kr/ L: 0.338 O: 0.022 m: 2-3

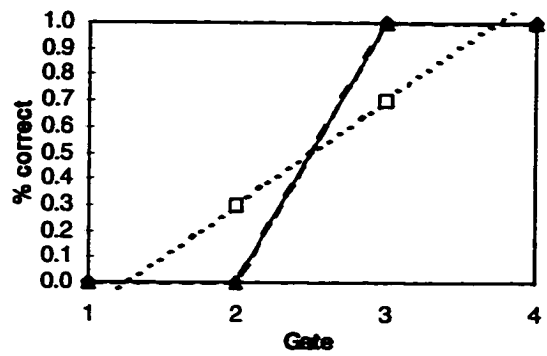




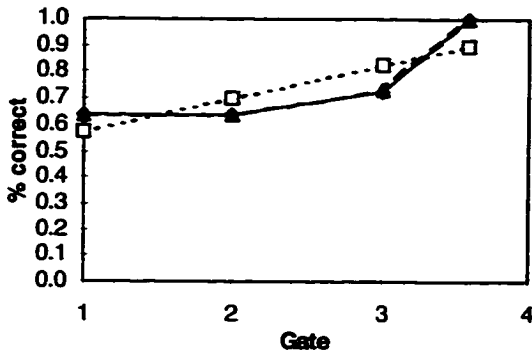
73 acrobat /kr/ L: 0.186 O: 0.138 m: 1-2



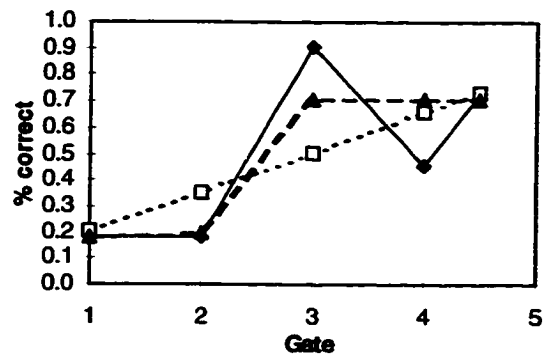
77 split /pl/ L: 0.447 O: 0.000 m: 2-3



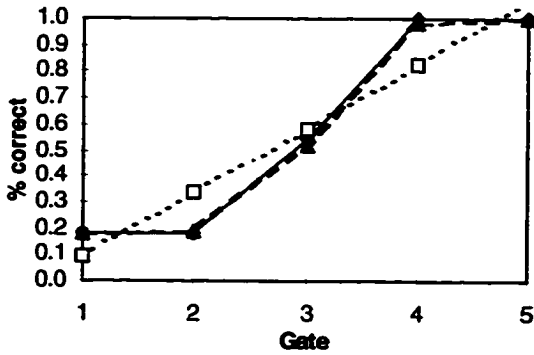
74 drop /dr/ L: 0.165 O: 0.007 m: 3-4



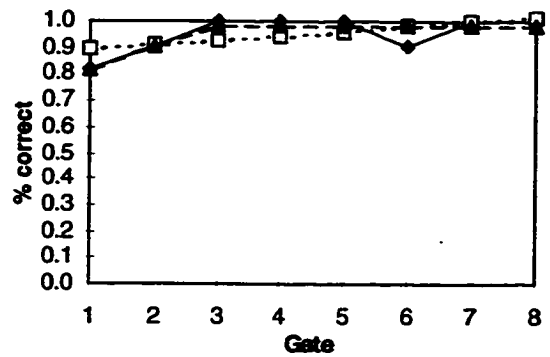
78 twelve /tw/ L: 0.484 O: 0.325 m: 2-3



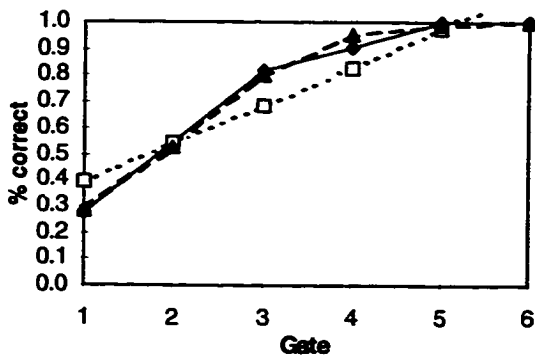
75 groan /gr/ L: 0.262 O: 0.034 m: 3-4



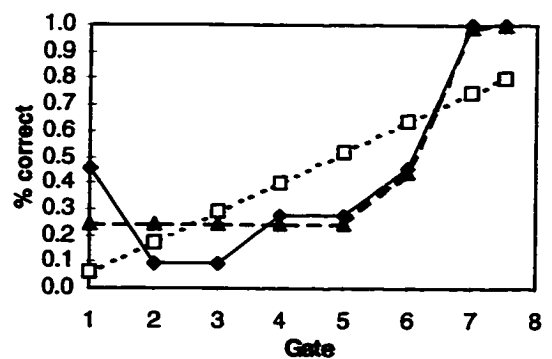
79 court /rt/ L: 0.143 O: 0.083 m: 2-3



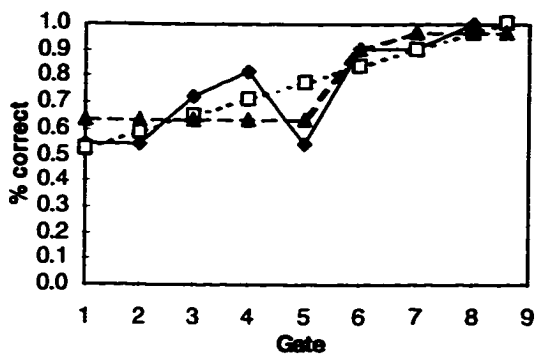
76 plain /pl/ L: 0.232 O: 0.055 m: 2-3



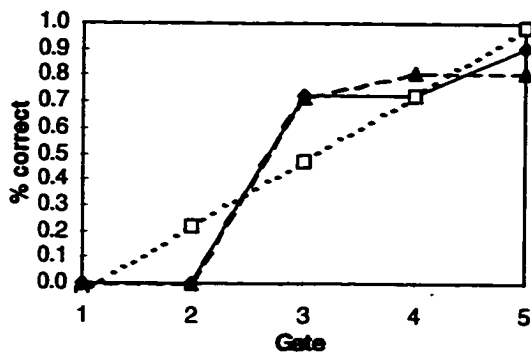
80 cork /rk/ L: 0.643 O: 0.305 m: 6-7



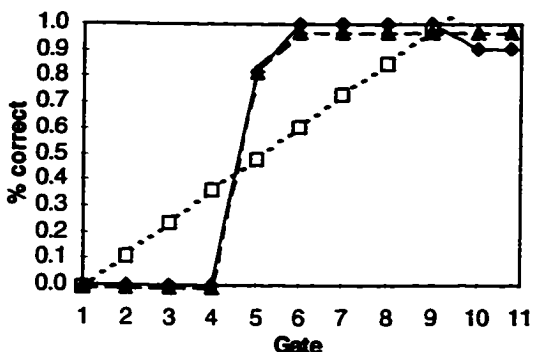
81 help /lp/ L: 0.280 O: 0.268 m: 5-6



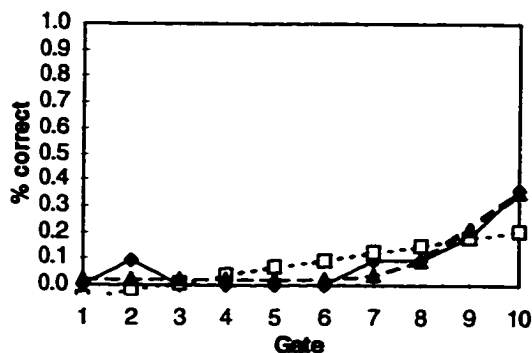
85 unconcealed /ns/ L: 0.345 O: 0.129 m: 2-3



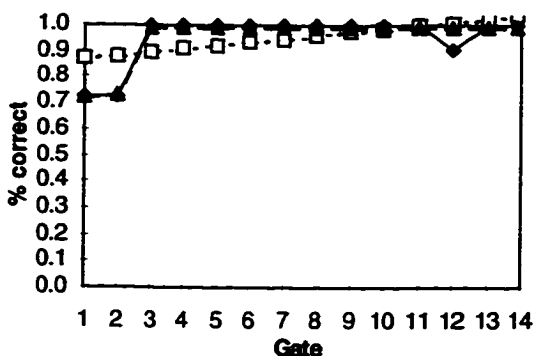
82 fans /nz/ L: 0.823 O: 0.106 m: 4-5



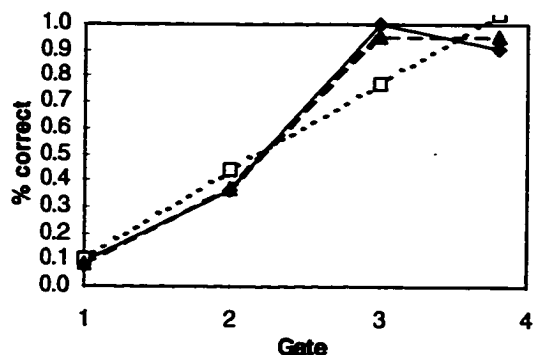
86 snow /sn/ L: 0.241 O: 0.106 m: 9-10



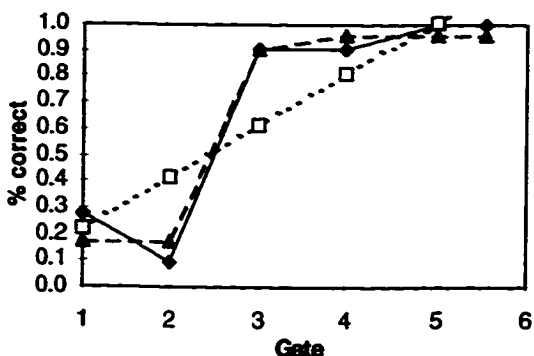
83 dance /ns/ L: 0.303 O: 0.087 m: 2-3



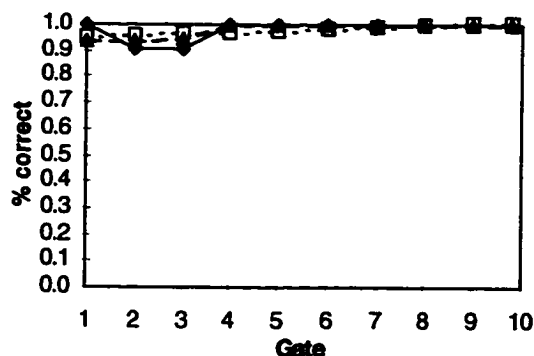
87 Disney /zn/ L: 0.273 O: 0.065 m: 2-3



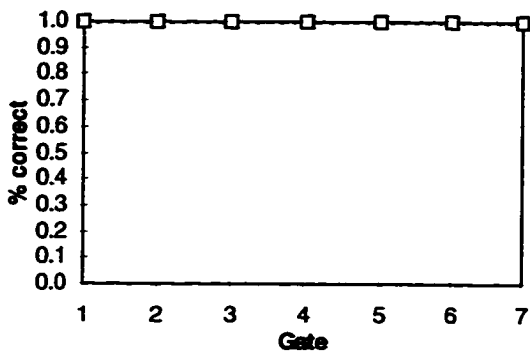
84 fancy /ns/ L: 0.468 O: 0.150 m: 2-3



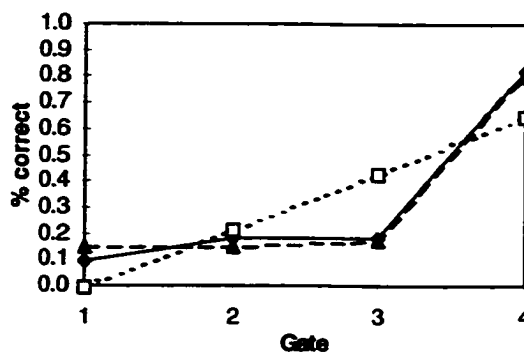
88 farm /rm/ L: 0.098 O: 0.075 m: 3-4



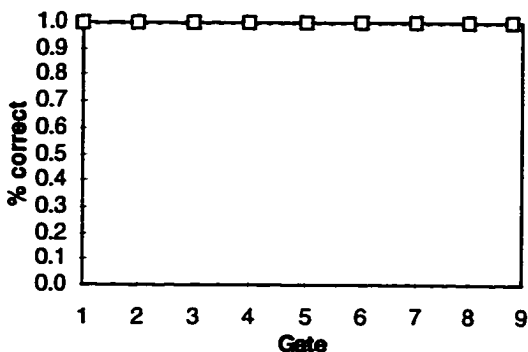
89 corn /rn/ L: 0.000



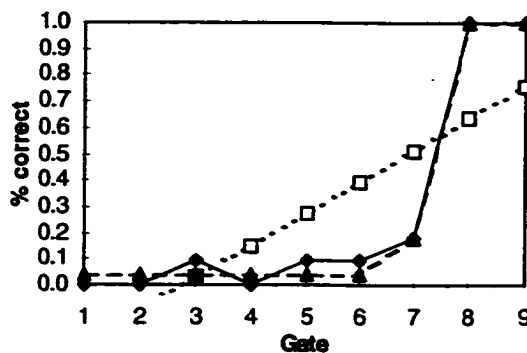
93 fragile /fr/ L: 0.318 O: 0.069 m: 3-4



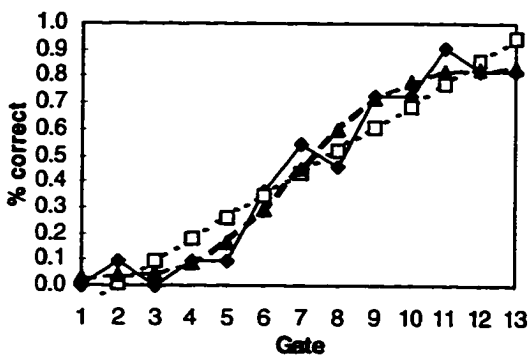
90 film /fm/ L: 0.000



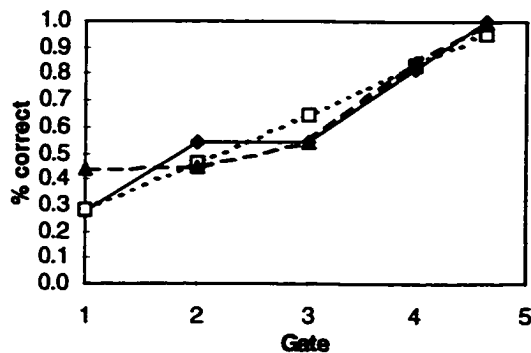
94 sleep /sl/ L: 0.712 O: 0.112 m: 7-8



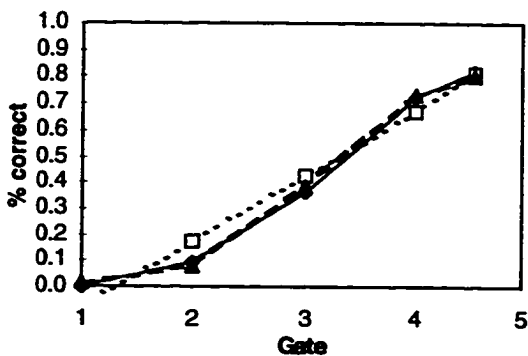
91 ranch /ntʃ/ L: 0.358 O: 0.245 m: 6-7



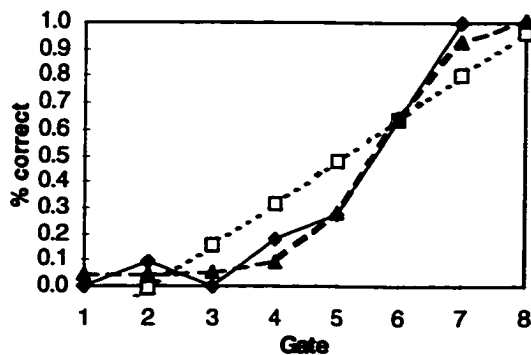
95 Iceland /sl/ L: 0.140 O: 0.197 m: 3-4



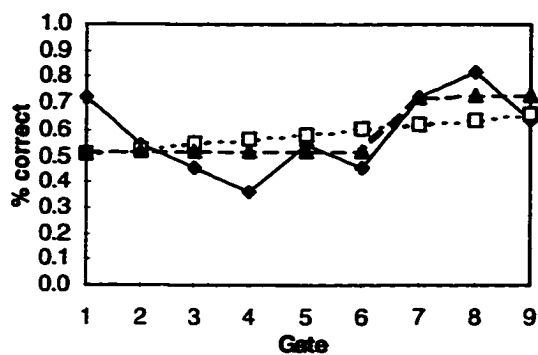
92 flash /fl/ L: 0.139 O: 0.033 m: 3-4



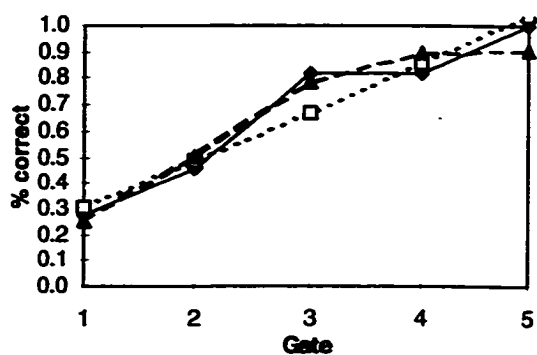
96 swan /sw/ L: 0.404 O: 0.144 m: 5-6



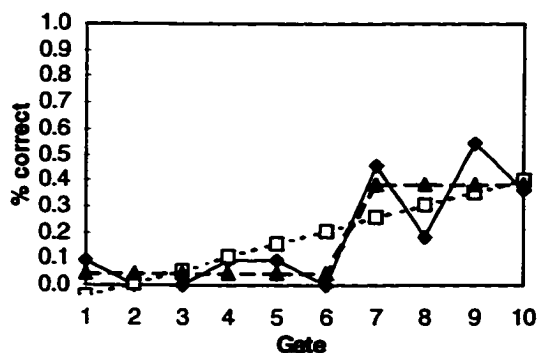
97 golf /lf/ L: 0.405 O: 0.306 m: 6-7



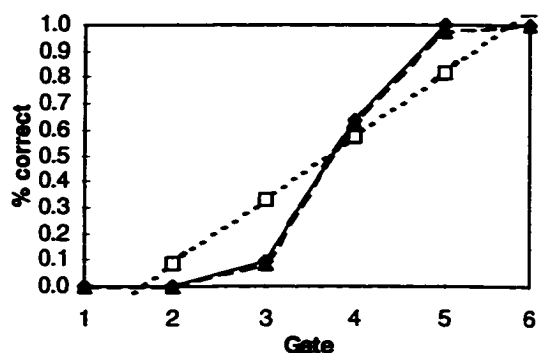
101 cultural /ltʃ/ L: 0.163 O: 0.135 m: 2-3



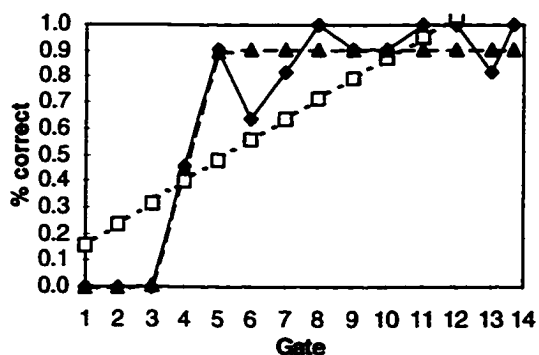
98 wharf /rf/ L: 0.401 O: 0.291 m: 6-7



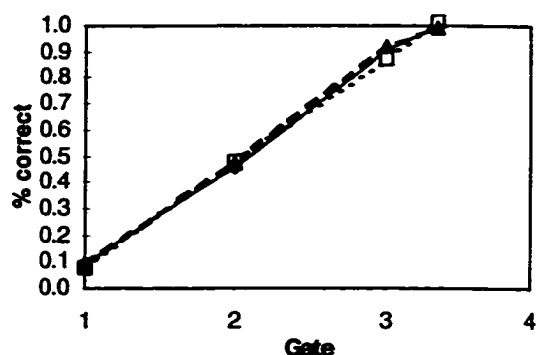
102 marginal /rdʒ/ L: 0.361 O: 0.021 m: 3-4



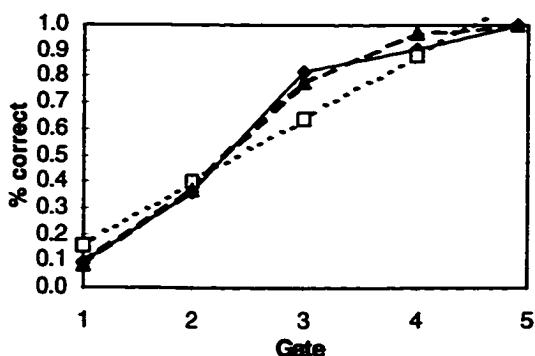
99 false /s/ L: 0.792 O: 0.351 m: 4-5



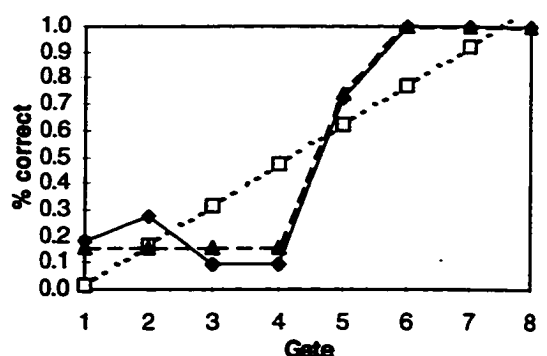
103 optical /pt/ L: 0.045 O: 0.024 m: 2-3



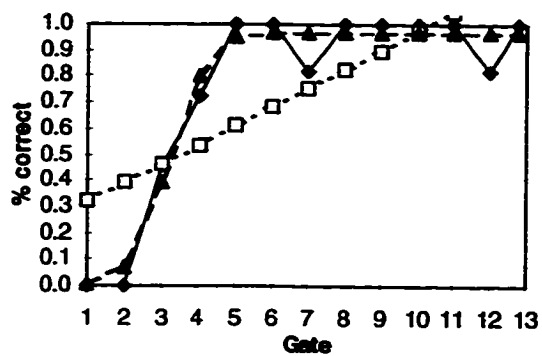
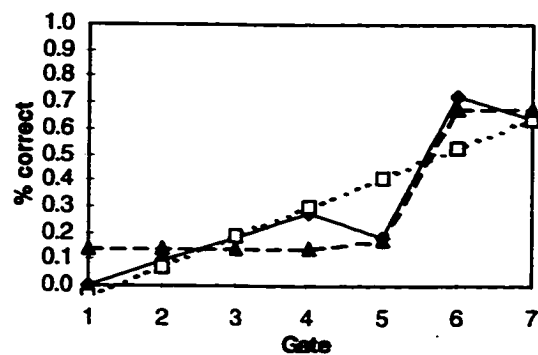
100 calcium /s/ L: 0.219 O: 0.073 m: 2-3



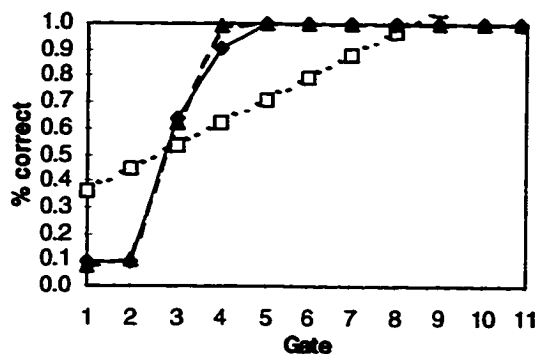
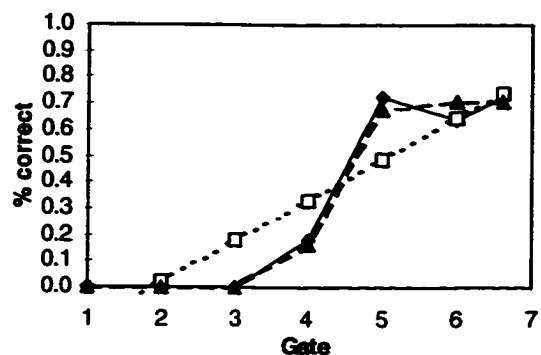
104 pact /kt/ L: 0.555 O: 0.151 m: 4-5



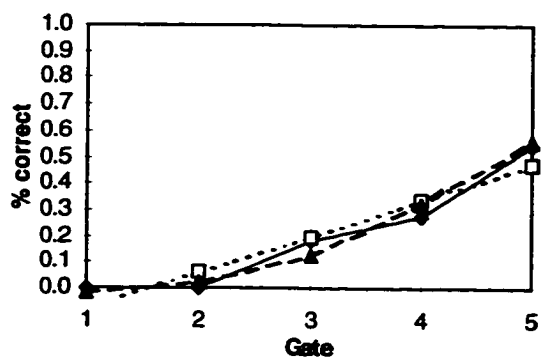
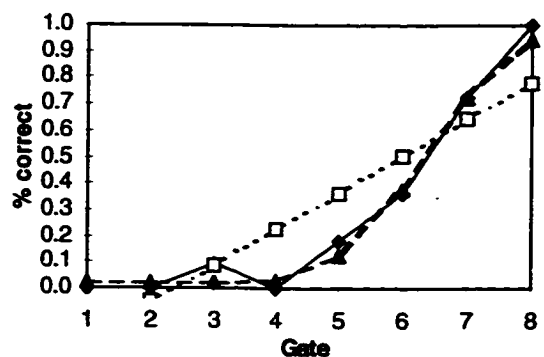
105 coughs /fs/ L: 0.844 O: 0.262 m: 3-4

109 biopsy /a<sup>i</sup>a/ L: 0.311 O: 0.213 m: 5-6

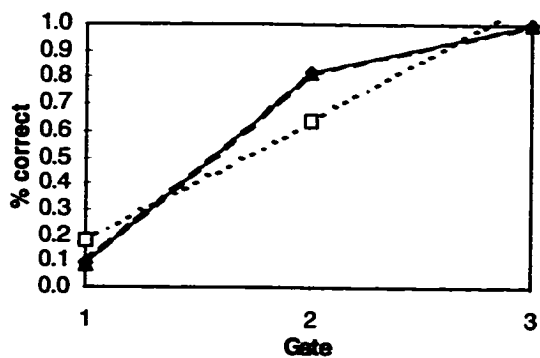
106 nerves /vz/ L: 0.711 O: 0.087 m: 2-3

110 biography /a<sup>i</sup>a/ L: 0.360 O: 0.092 m: 4-5

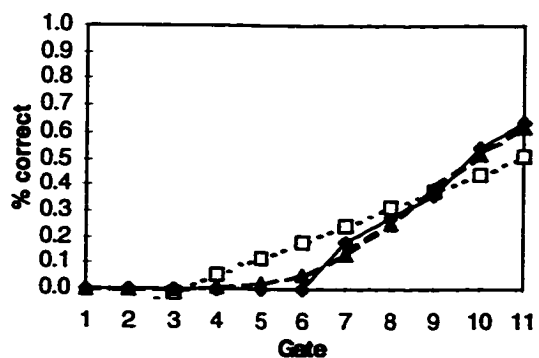
107 amnesty /mn/ L: 0.138 O: 0.080 m: 4-5

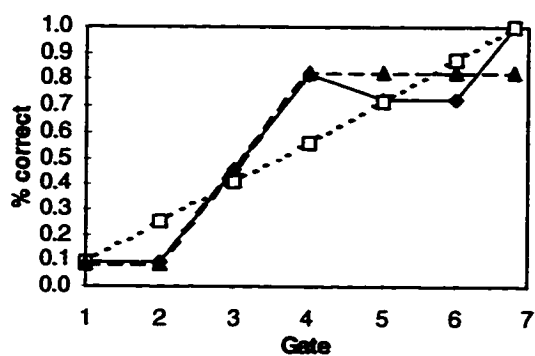
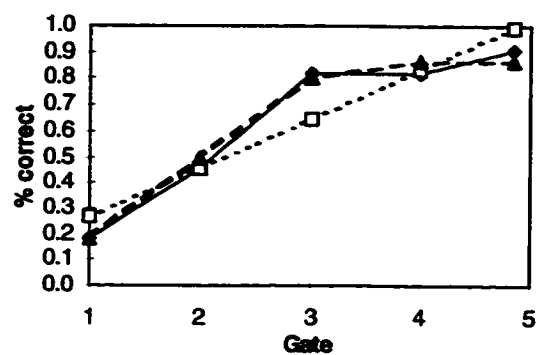
111 biotech /a<sup>i</sup>o<sup>w</sup>/ L: 0.445 O: 0.116 m: 6-7

108 garlic /rI/ L: 0.223 O: 0.001 m: 1-2

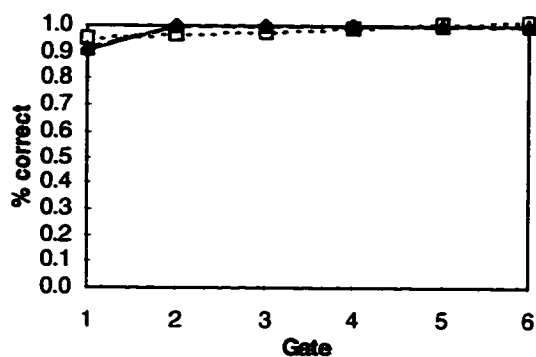
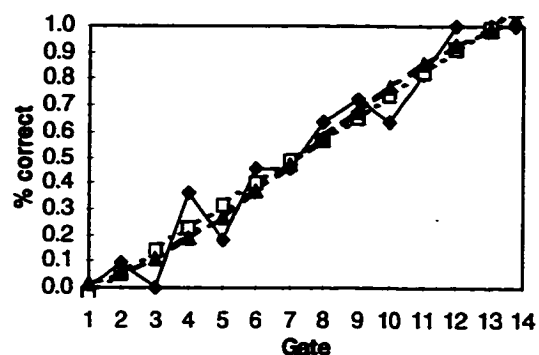
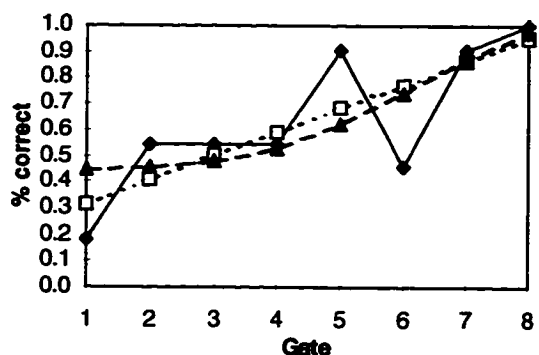
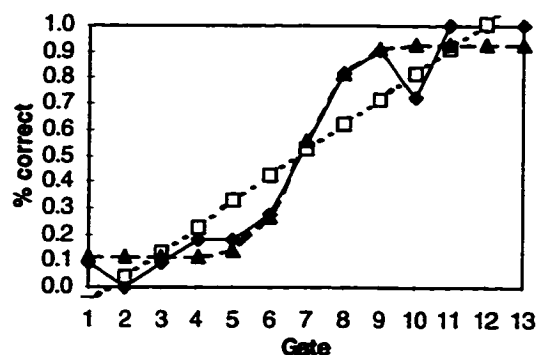
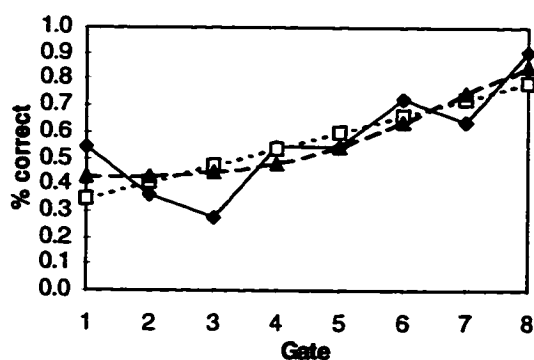
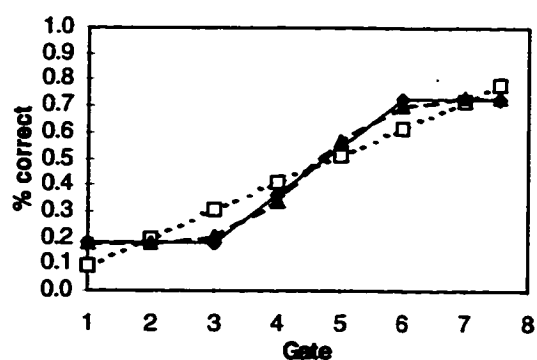


112 eon /ia/ L: 0.331 O: 0.090 m: 8-9

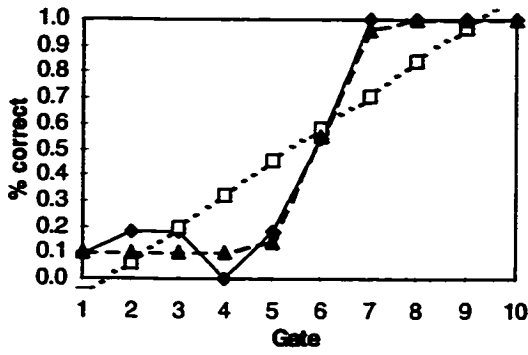


113 diagonal /a<sup>i</sup>æ/ L: 0.339 O: 0.223 m: 3-4117 data /de<sup>i</sup>/ L: 0.212 O: 0.075 m: 2-3

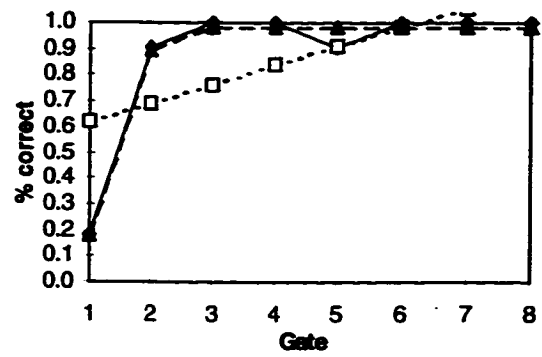
114 react /iæ/ L: 0.063 O: 0.001 m: 1-2

118 fade /e<sup>i</sup>d/ L: 0.307 O: 0.307 m: 7-8115 tiger /ta<sup>i</sup>/ L: 0.445 O: 0.501 m: 6-7119 doubt /a<sup>w</sup>ʊ/ L: 0.420 O: 0.277 m: 6-8116 bite /a<sup>i</sup>t/ L: 0.334 O: 0.278 m: 6-7120 soybean /o<sup>i</sup>b/ L: 0.202 O: 0.049 m: 4-5

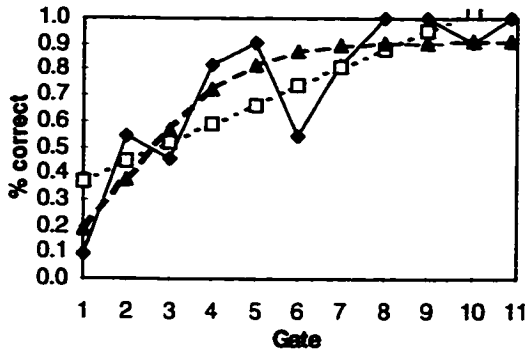
121 toad /toʷ/ L: 0.580 O: 0.164 m: 6-7



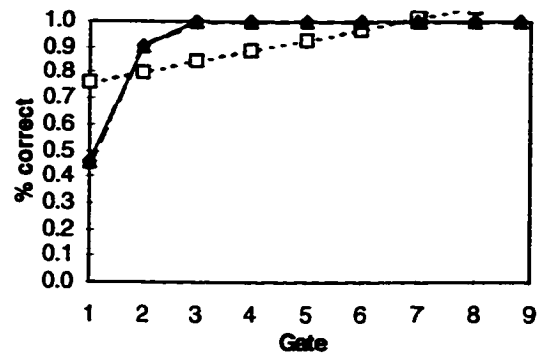
125 button /tʌ/ L: 0.583 O: 0.084 m: 1-2



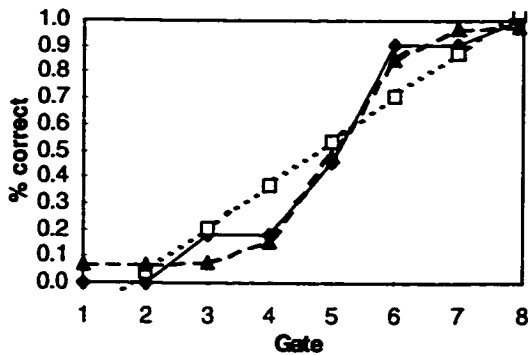
122 oats /oʷt/ L: 0.529 O: 0.454 m: 1-2



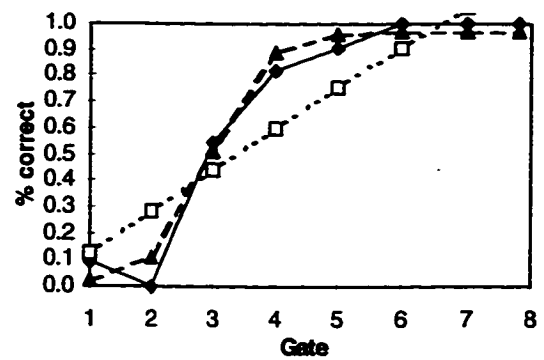
126 beetle /t/ L: 0.399 O: 0.003 m: 1-2



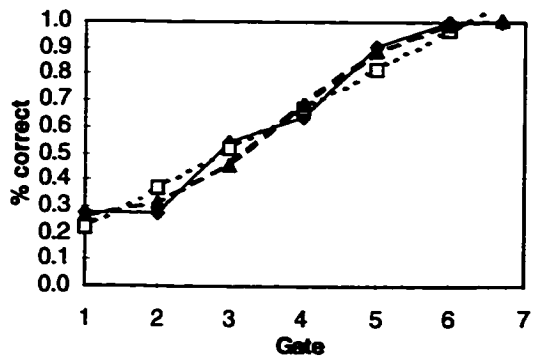
123 courage /kə/ L: 0.325 O: 0.173 m: 5-6



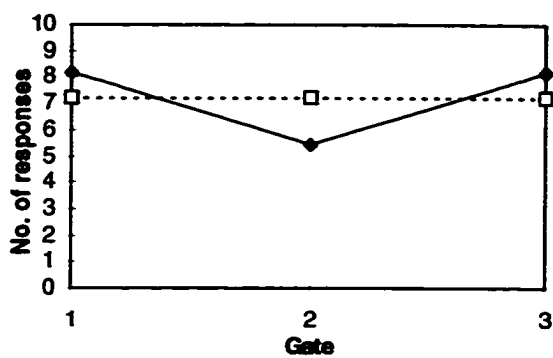
127 apple /p/ L: 0.464 O: 0.171 m: 2-3



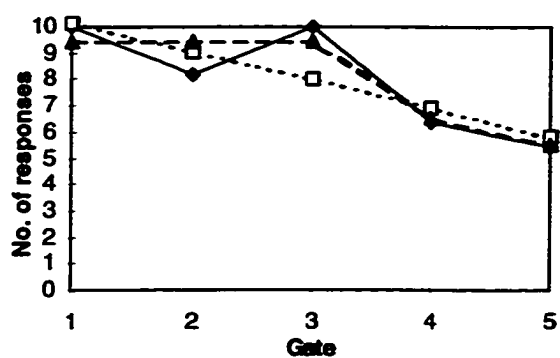
124 circle /ək/ L: 0.169 O: 0.113 m: 3-4



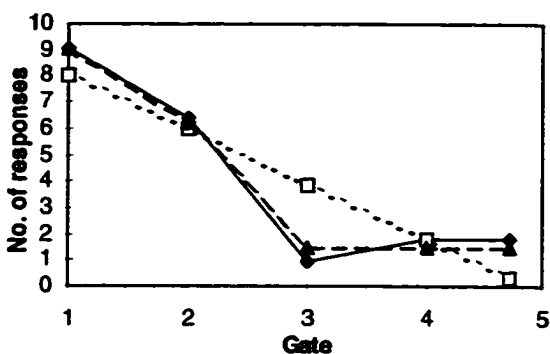
1 /todana/ [to] L: 2.227



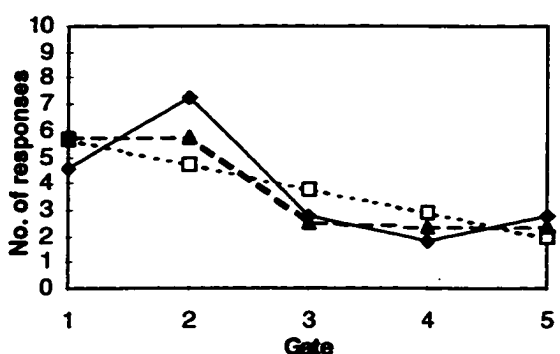
5 /hakama/ [ka] L: 2.300 O: 1.506 m: 3-4



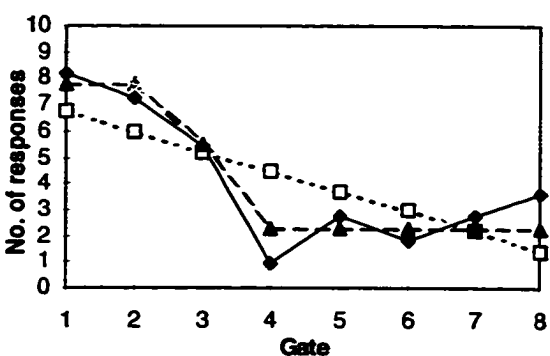
2 /tatoe'ru/ [to] L: 3.502 O: 0.762 m: 2-3



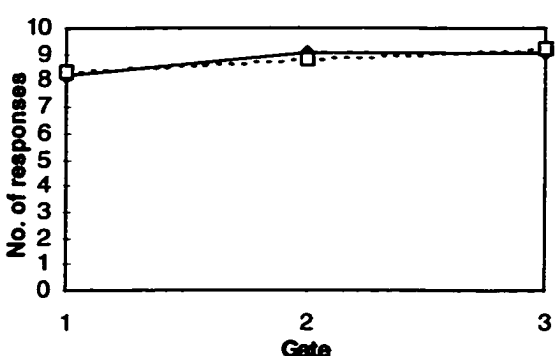
6 /sya'kai/ [ka] L: 3.252 O: 2.068 m: 2-3



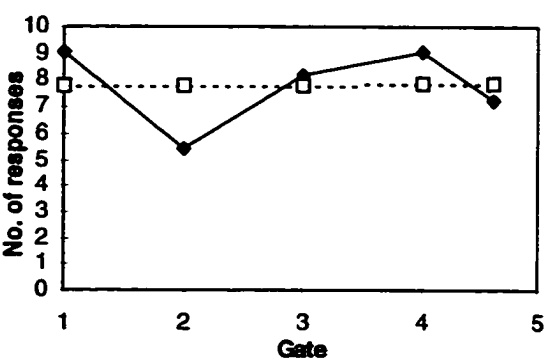
3 /ka'to/ [to] L: 4.882 O: 2.182 m: 3-4



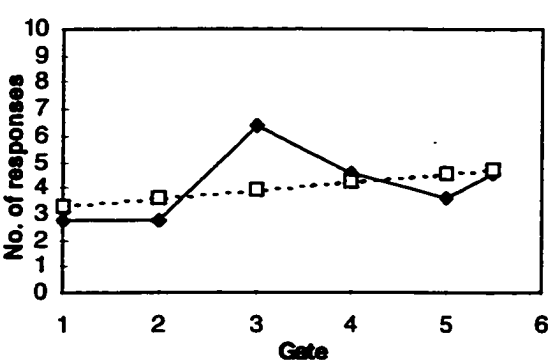
7 /dama'ru/ [da] L: 0.371



4 /kakari'iN/ [ka] L: 3.041

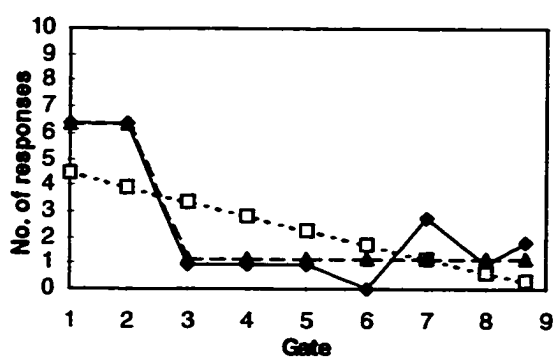


8 /midare'ru/ [da] L: 2.830

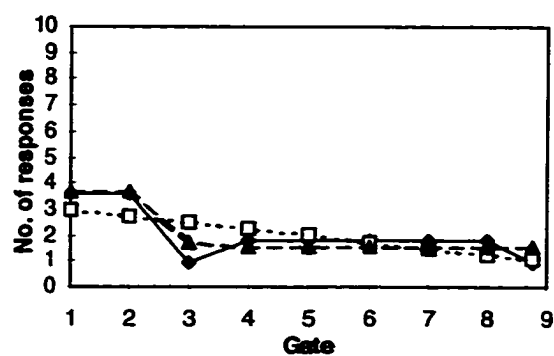




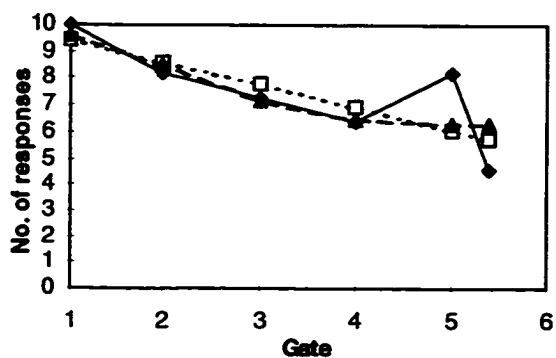
9 /ku'da/ [da] L: 5.385 O: 2.120 m: 2-3



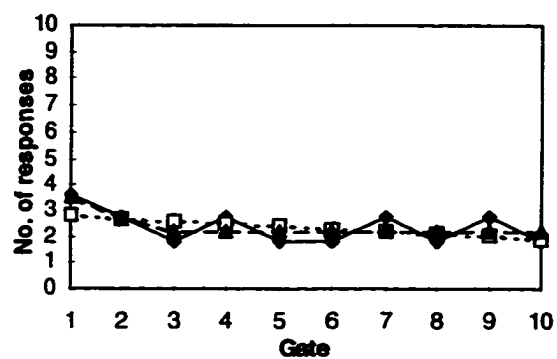
13 /hatake/ [ak] L: 2.095 O: 1.186 m: 2-3



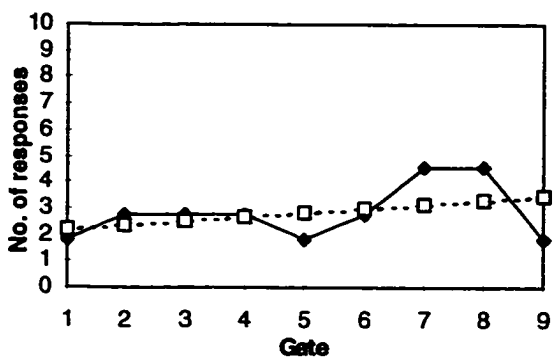
10 /hotoke/ [ot] L: 2.625 O: 2.623 m: 2-3



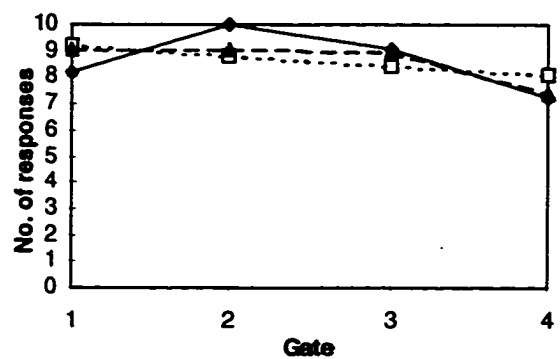
14 /ha'yaku/ [ak] L: 1.681 O: 1.267 m: 1-2



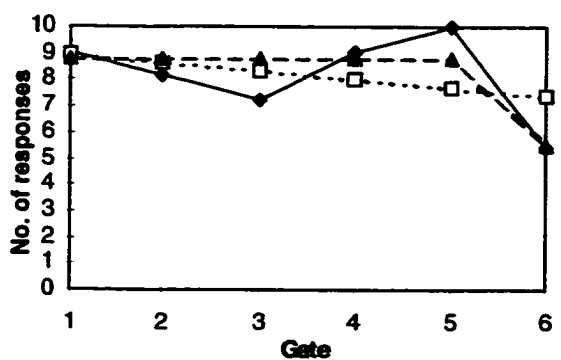
11 /himoto/ [ot] L: 2.761



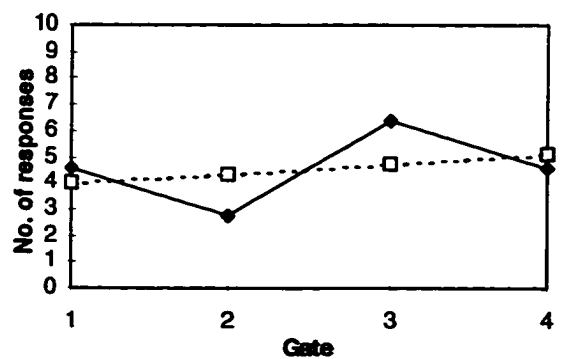
15 /kadai/ [ad] L: 1.863 O: 1.298 m: 3-4



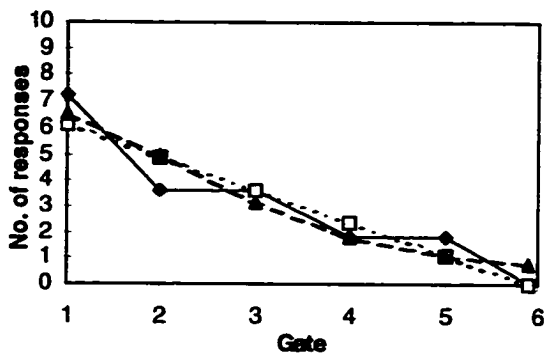
12 /hakobu/ [ak] L: 3.395 O: 2.082 m: 5-6



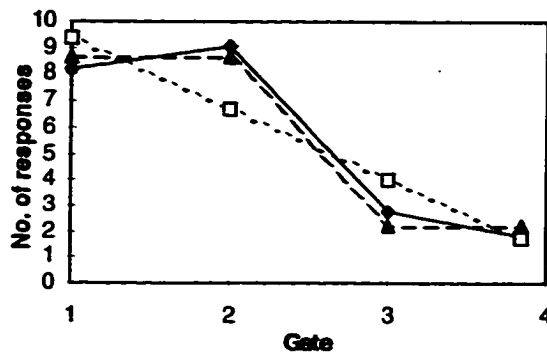
16 /hanada'yoru/ [ad] L: 2.439



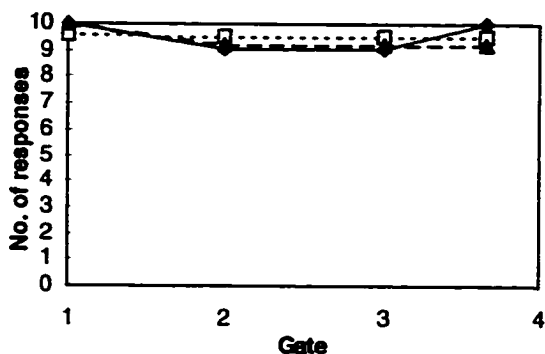
17 /ka'nada/ [ad] L: 1.907 O: 1.907 m: 2-3



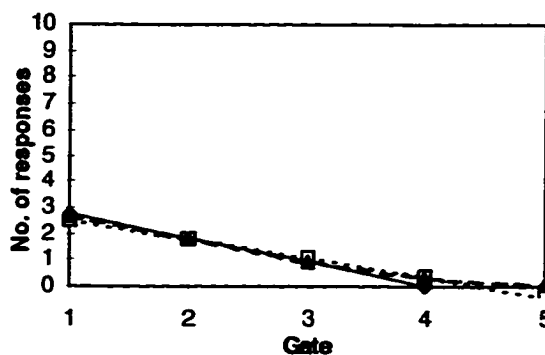
21 /kemuri/ [em] L: 2.971 O: 0.928 m: 2-3



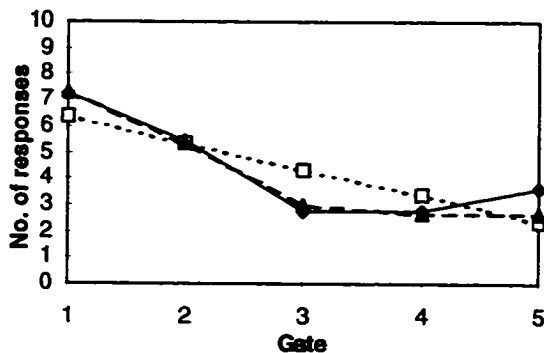
18 /megumi/ [me] L: 0.906 O: 0.815 m: 1-2



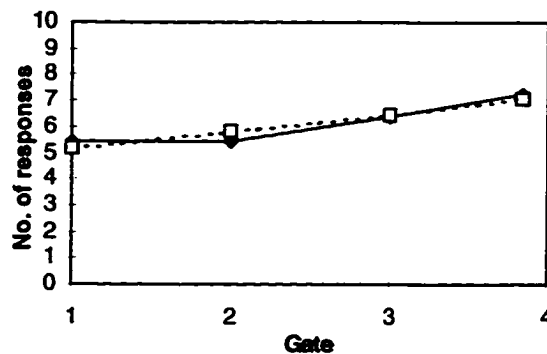
22 /tabemo'no/ [em] L: 0.575 O: 0.321 m: 2-3



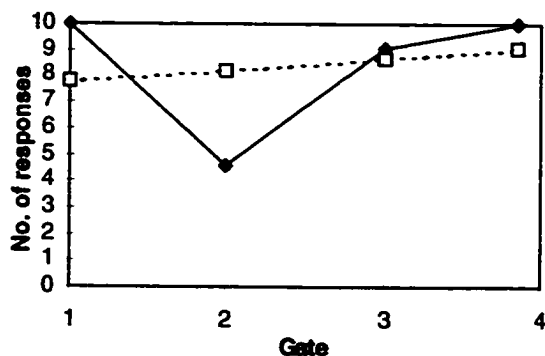
19 /tomeru/ [me] L: 2.353 O: 0.978 m: 2-3



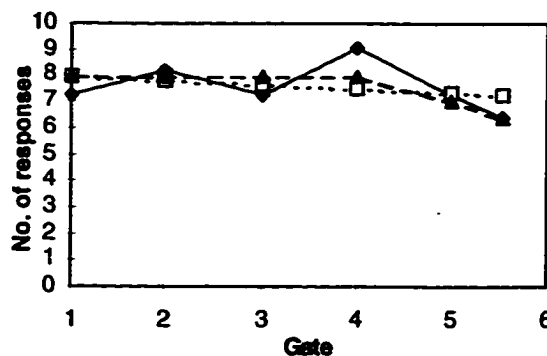
23 /teni'motu/ [en] L: 0.536



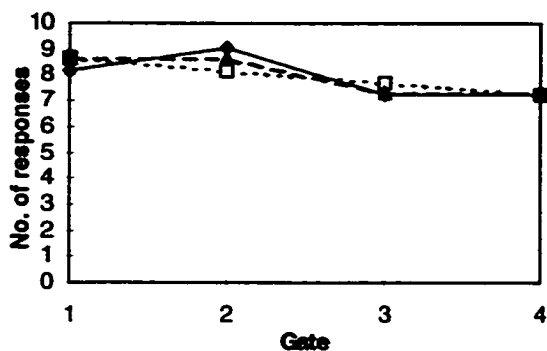
20 /nemui/ [ne] L: 4.421



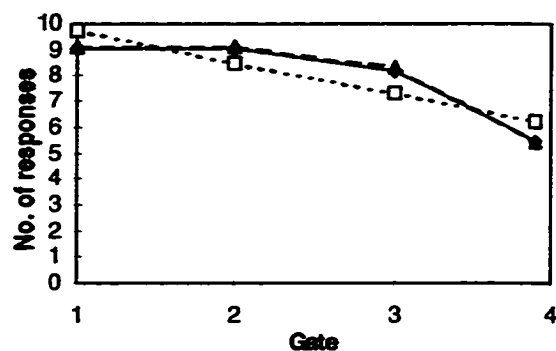
24 /soda'tu/ [so] L: 2.025 O: 1.533 m: 5-6



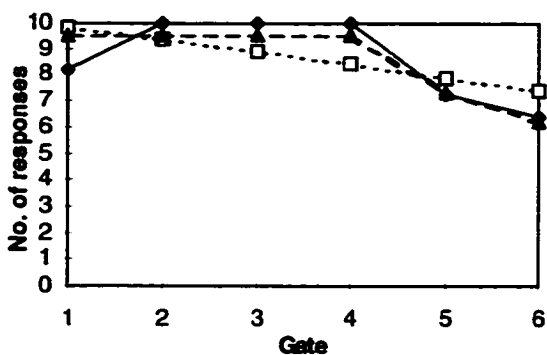
25 /zabu'toN/ [za] L: 1.113 O: 0.650 m: 2-3



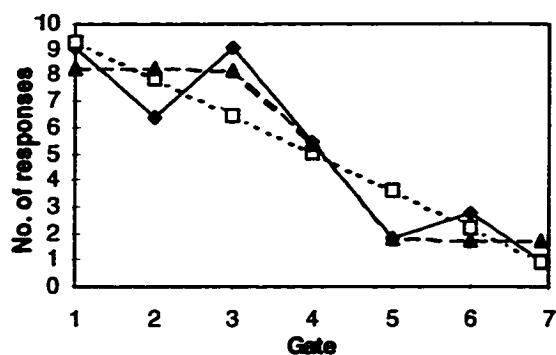
29 /kazari/ [az] L: 1.443 O: 0.190 m: 3-4



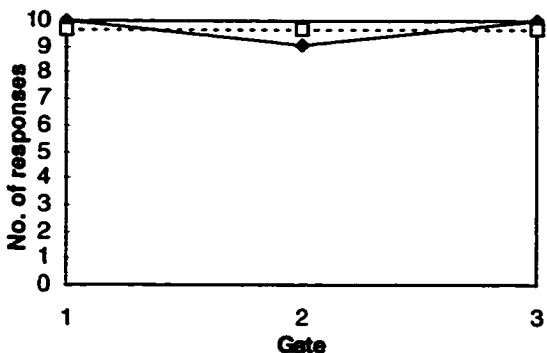
26 /syabe'ru/ [ʃa] L: 2.924 O: 1.587 m: 4-5



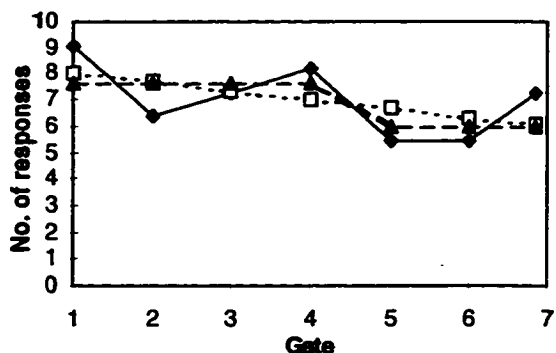
30 /basyo/ [aʃ] L: 3.600 O: 2.605 m: 4-5



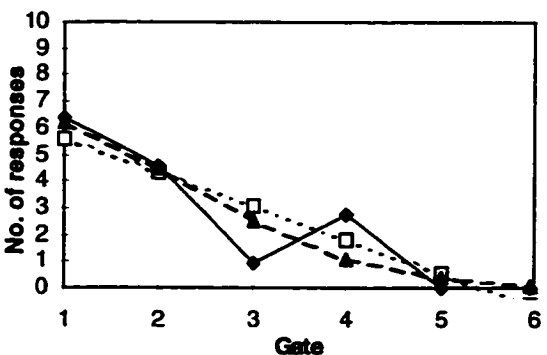
27 /hokeN/ [ho] L: 0.742



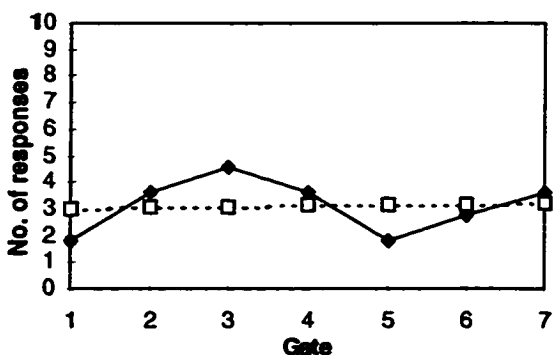
31 /gohoo/ [oh] L: 2.833 O: 2.530 m: 4-5



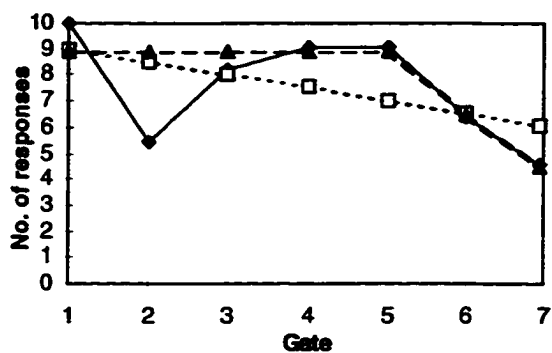
28 /zyosei/ [os] L: 2.622 O: 2.331 m: 2-3



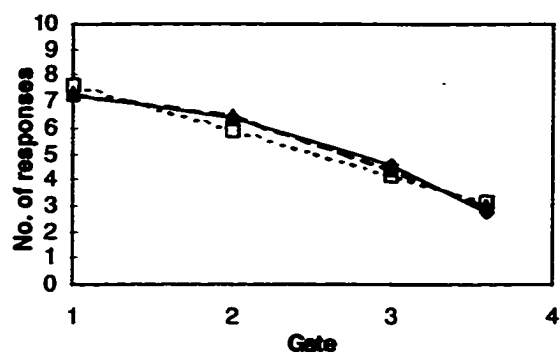
32 /wahuku/ [aʃ] L: 2.519



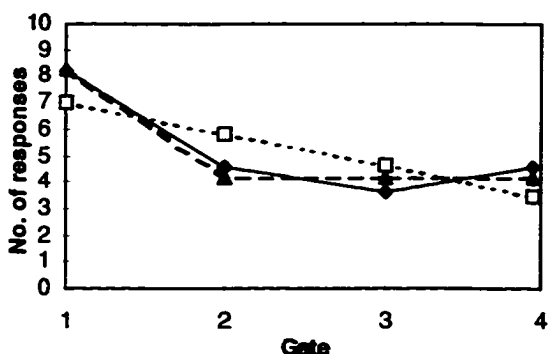
33 /dohyoo/ [oç] L: 4.405 O: 3.684 m: 5-6



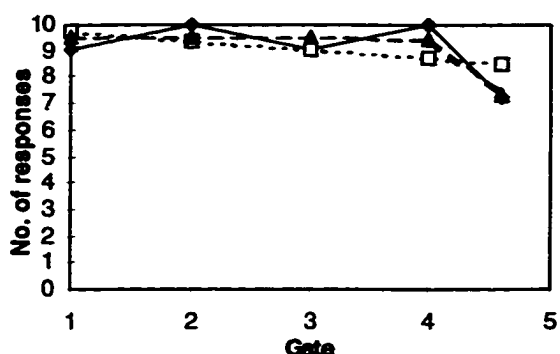
37 /huyoo/ [uj] L: 0.803 O: 0.295 m: 3-4



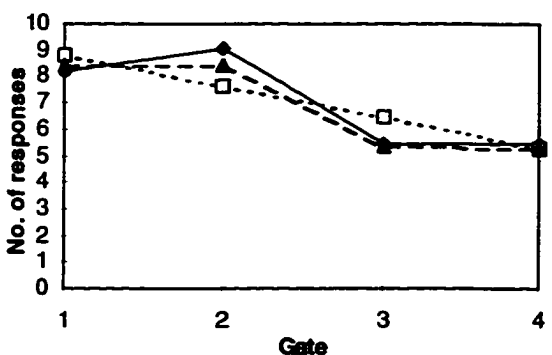
34 /harada'tu/ [ra] L: 2.253 O: 0.752 m: 1-2



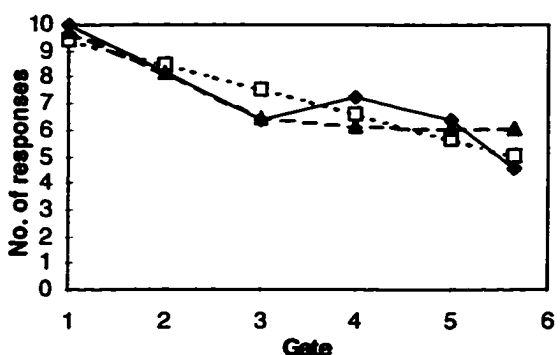
38 /mawari/ [aw] L: 1.992 O: 0.931 m: 4-5



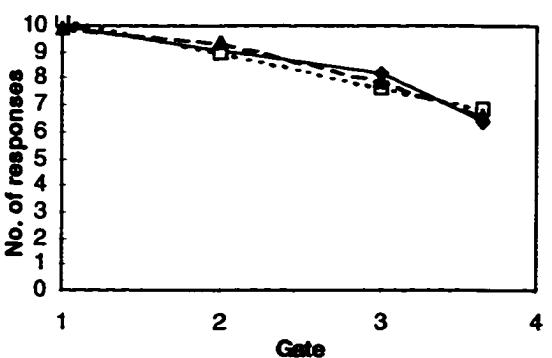
35 /yubi/ [ju] L: 1.885 O: 0.711 m: 2-3



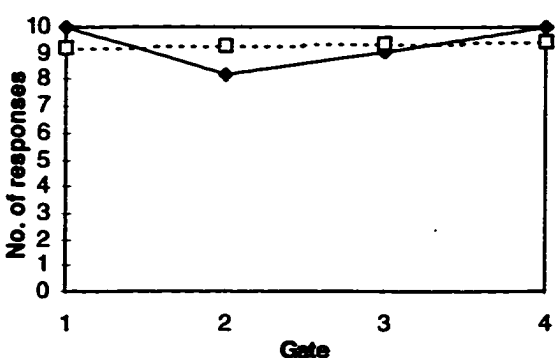
39 /tyazuke/ [tja] L: 1.735 O: 1.977 m: 2-3



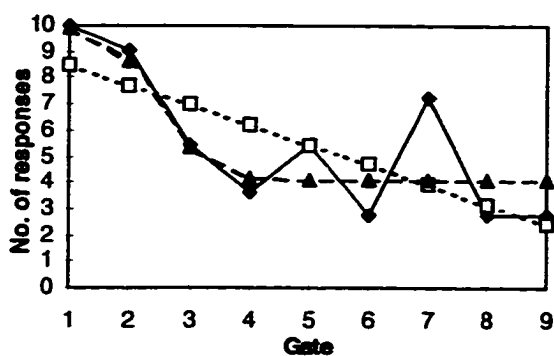
36 /kara'i/ [ar] L: 0.748 O: 0.522 m: 3-4



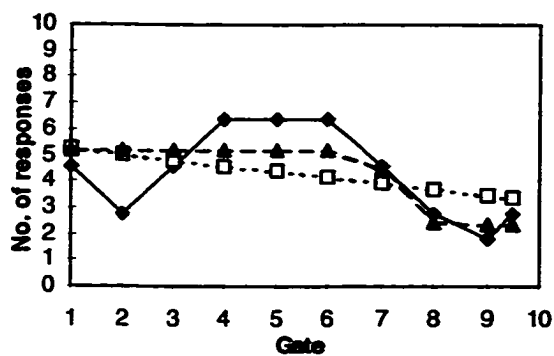
40 /zyokyo'ozyu/ [dzo] L: 1.494



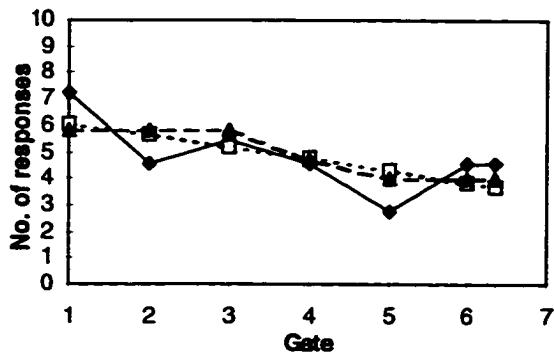
41 /matɪ/ [at̪] L: 5.327 O: 4.243 m: 2-3



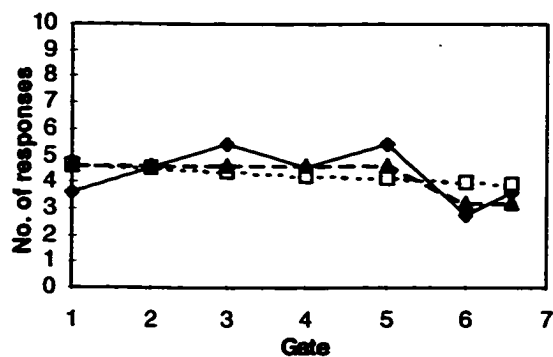
45 /teNkiN/ [ɲk] L: 4.738 O: 3.401 m: 7-8



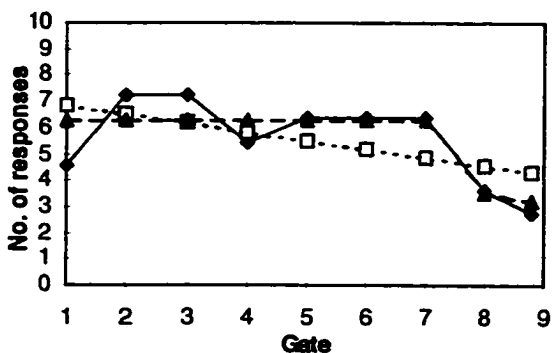
42 /tozi'ru/ [oɖʒ] L: 2.526 O: 2.469 m: 3-4



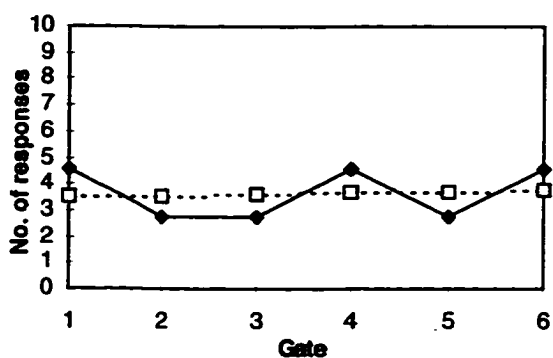
46 /kaNzeN/ [nz] L: 2.388 O: 1.656 m: 5-6



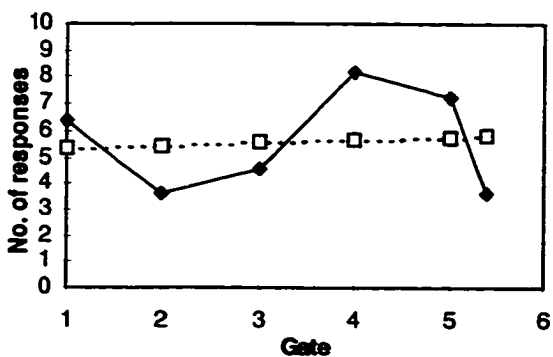
43 /haNtai/ [nt] L: 3.834 O: 2.435 m: 7-8



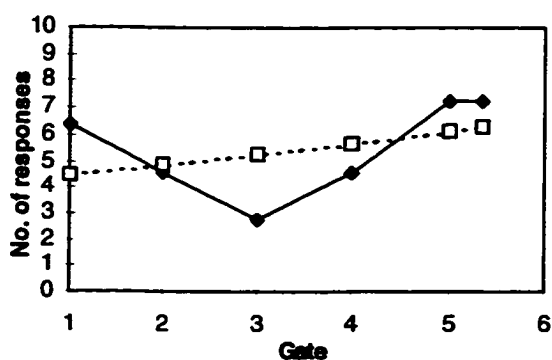
47 /seNsoo/ [ns] L: 2.216



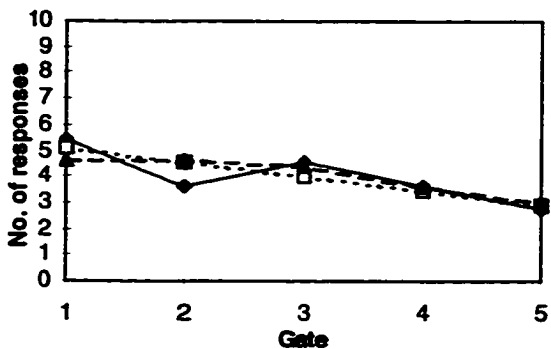
44 /kaNdoo/ [nd] L: 4.323



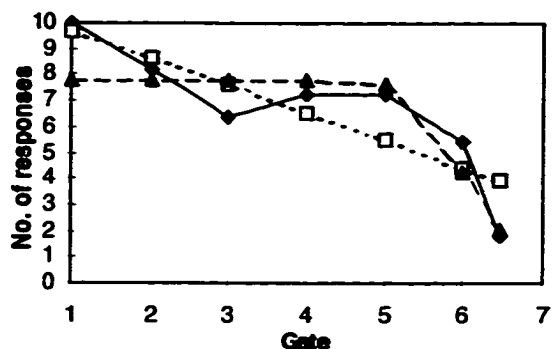
48 /keNritu/ [nr] L: 3.740



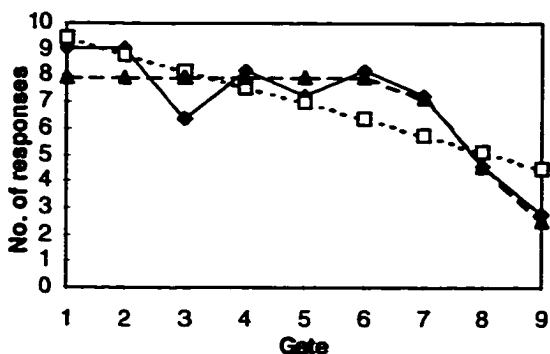
49 /koNyaku/ [ɲj] L: 1.150 O: 1.322 m: 3-4



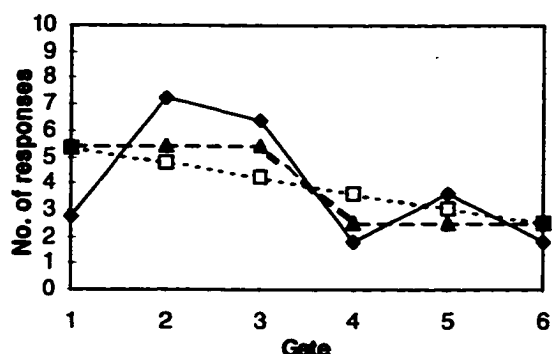
53 /kitamuki/ [kʰt] L: 3.342 O: 2.984 m: 6-7



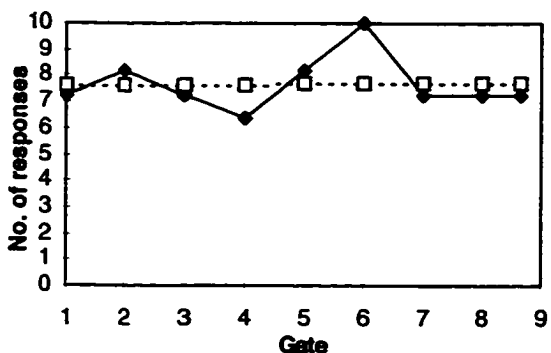
50 /kiNtyoo/ [nt̚] L: 3.634 O: 2.400 m: 7-8



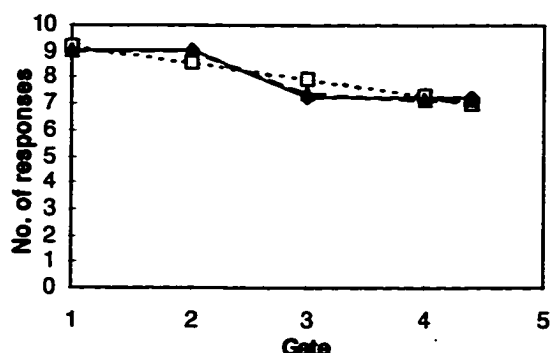
54 /kokutetu/ [kt] L: 4.673 O: 3.714 m: 3-4



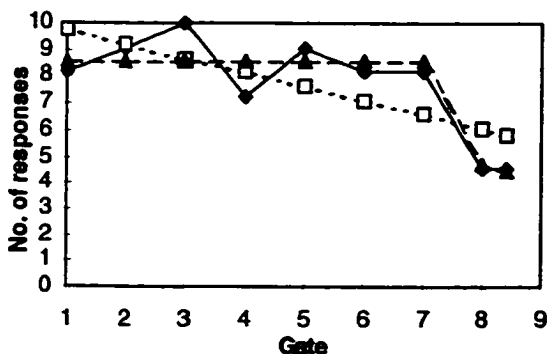
51 /sukunai/ [sk] L: 2.903



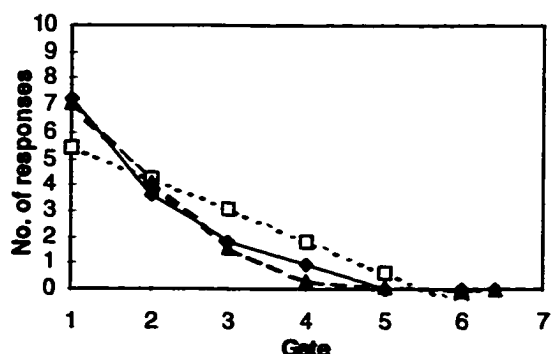
55 /kyaku/ [kj] L: 0.880 O: 0.156 m: 2-3



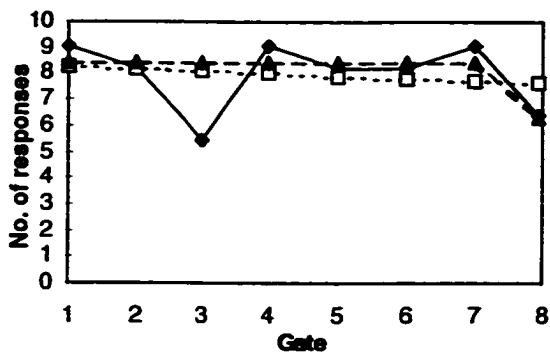
52 /sikaku/ [ʃʰk] L: 3.841 O: 2.183 m: 7-8



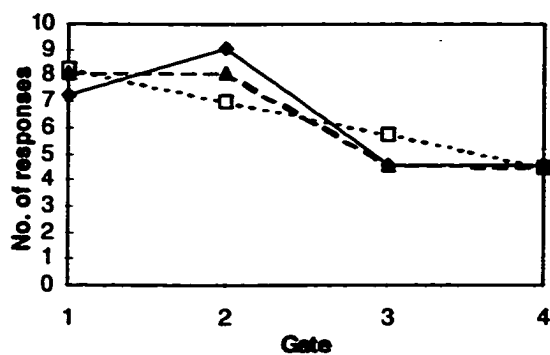
56 /dakyoo/ [kj] L: 2.808 O: 0.803 m: 1-2



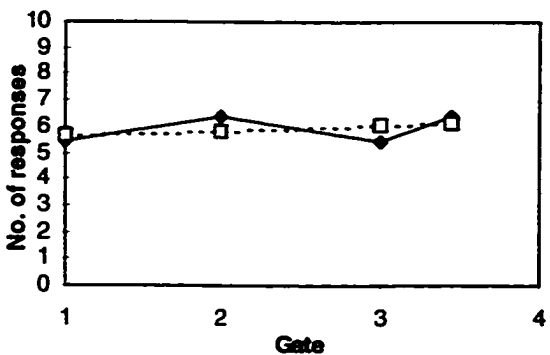
57 /hyoo/ [çj] L: 3.536 O: 3.237 m: 7-8



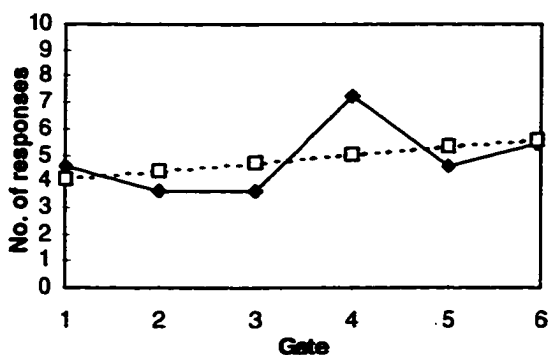
61 /sassoku/ [ss] L: 2.603 O: 1.294 m: 2-3



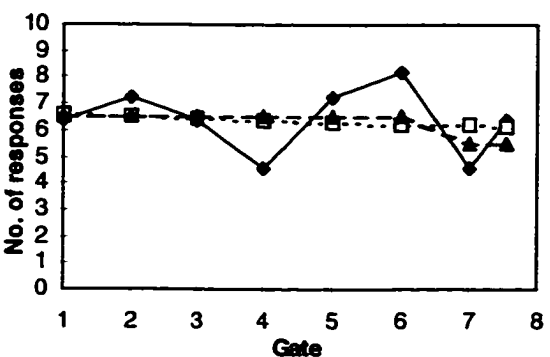
58 /ryokaN/ [rj] L: 0.840



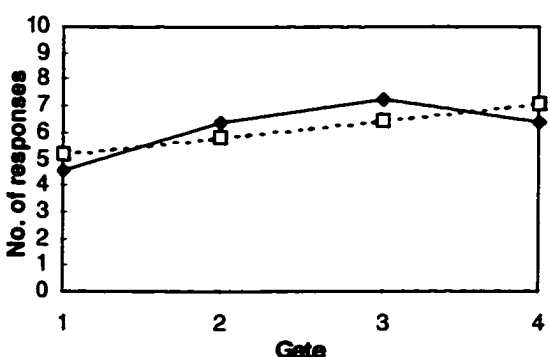
62 /hassya/ [ʃʃ] L: 2.768



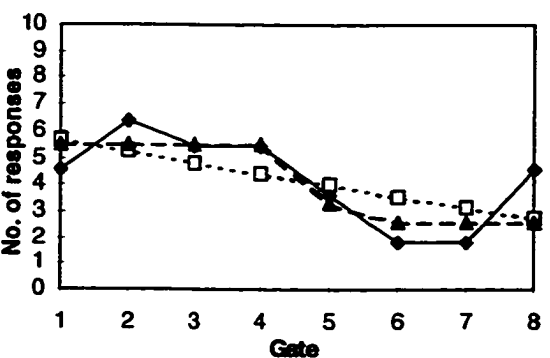
59 /mottaina'i/ [tt] L: 3.373 O: 3.070 m: 6-7



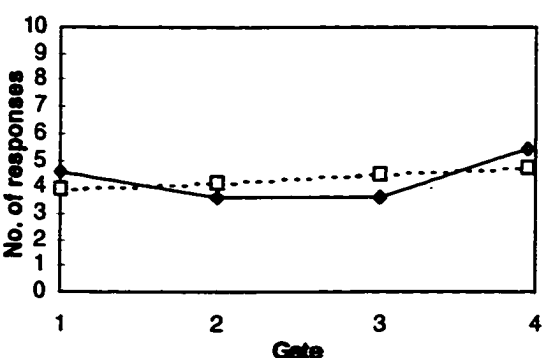
63 /teNmetu/ [mm] L: 1.379



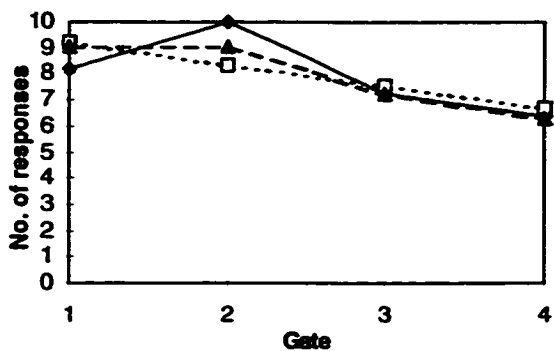
60 /sakka/ [kk] L: 3.500 O: 2.606 m: 4-5



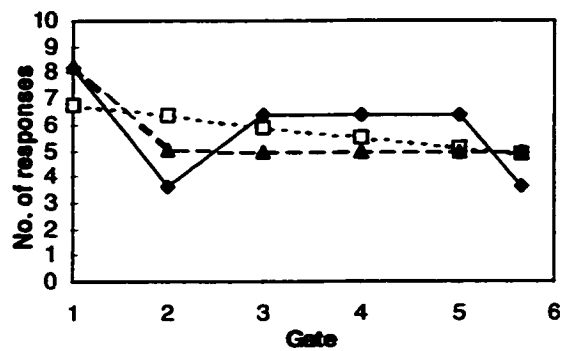
64 /aNnaizyo/ [nn] L: 1.386



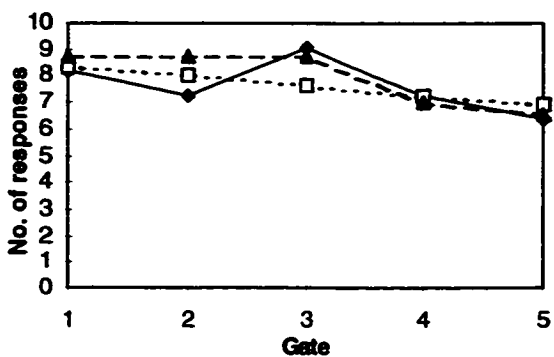
65 /tootyaku/ [oo] L: 1.971 O: 1.289 m: 2-3



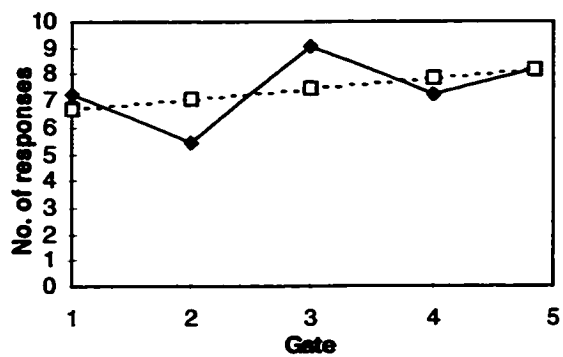
69 /siatu/ [ia] L: 3.645 O: 3.050 m: 1-2



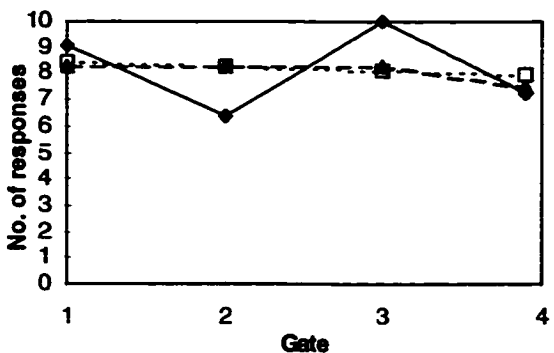
66 /keigo/ [ee] L: 1.725 O: 1.636 m: 3-4



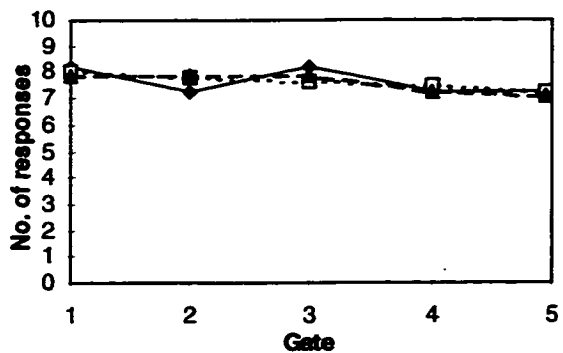
70 /kaeri'miti/ [ae] L: 2.440



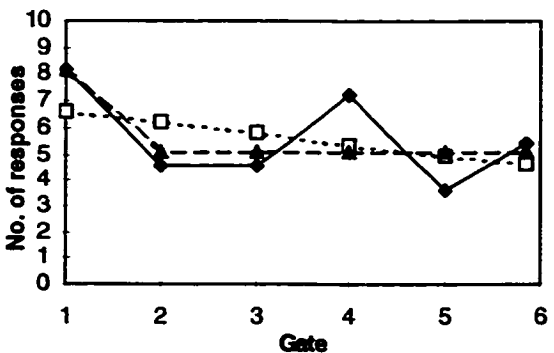
67 /syuukaN/ [uu] L: 2.850 O: 2.713 m: 3-4



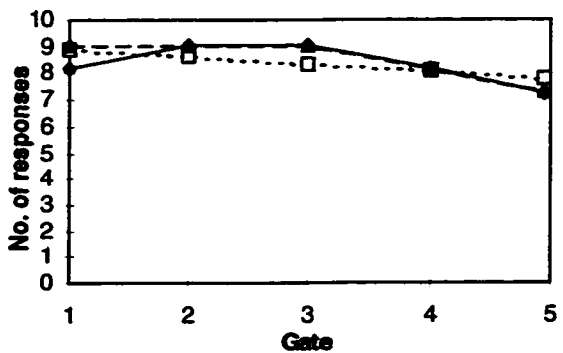
71 /taiko/ [ai] L: 0.813 O: 0.763 m: 3-4



68 /haori/ [ao] L: 3.579 O: 2.758 m: 1-2

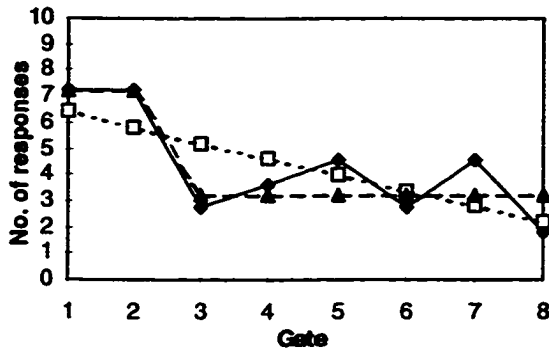


72 /koibito/ [oi] L: 1.259 O: 0.909 m: 4-5

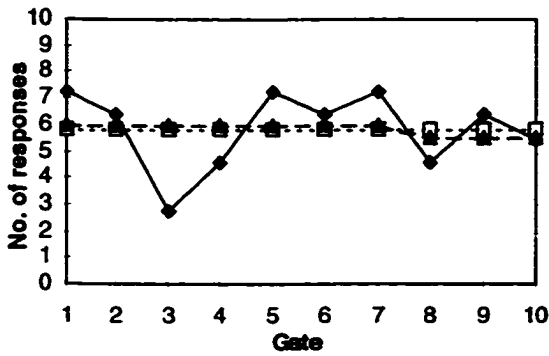




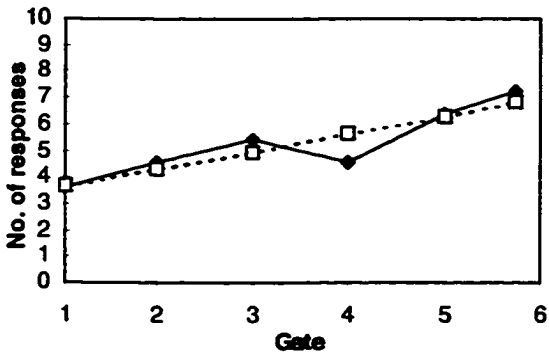
73 /teNiN/ [ēi] L: 3.730 O: 2.483 m: 2-3



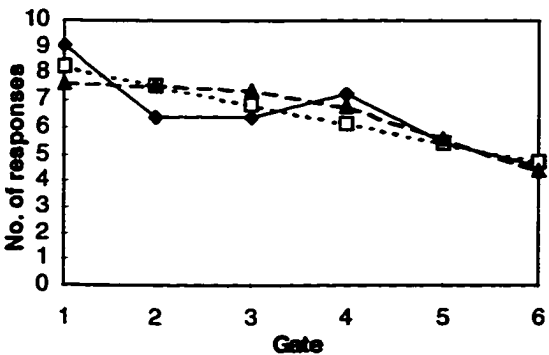
74 /hiN/ [iŋ] L: 4.491 O: 4.437 m: 7-8

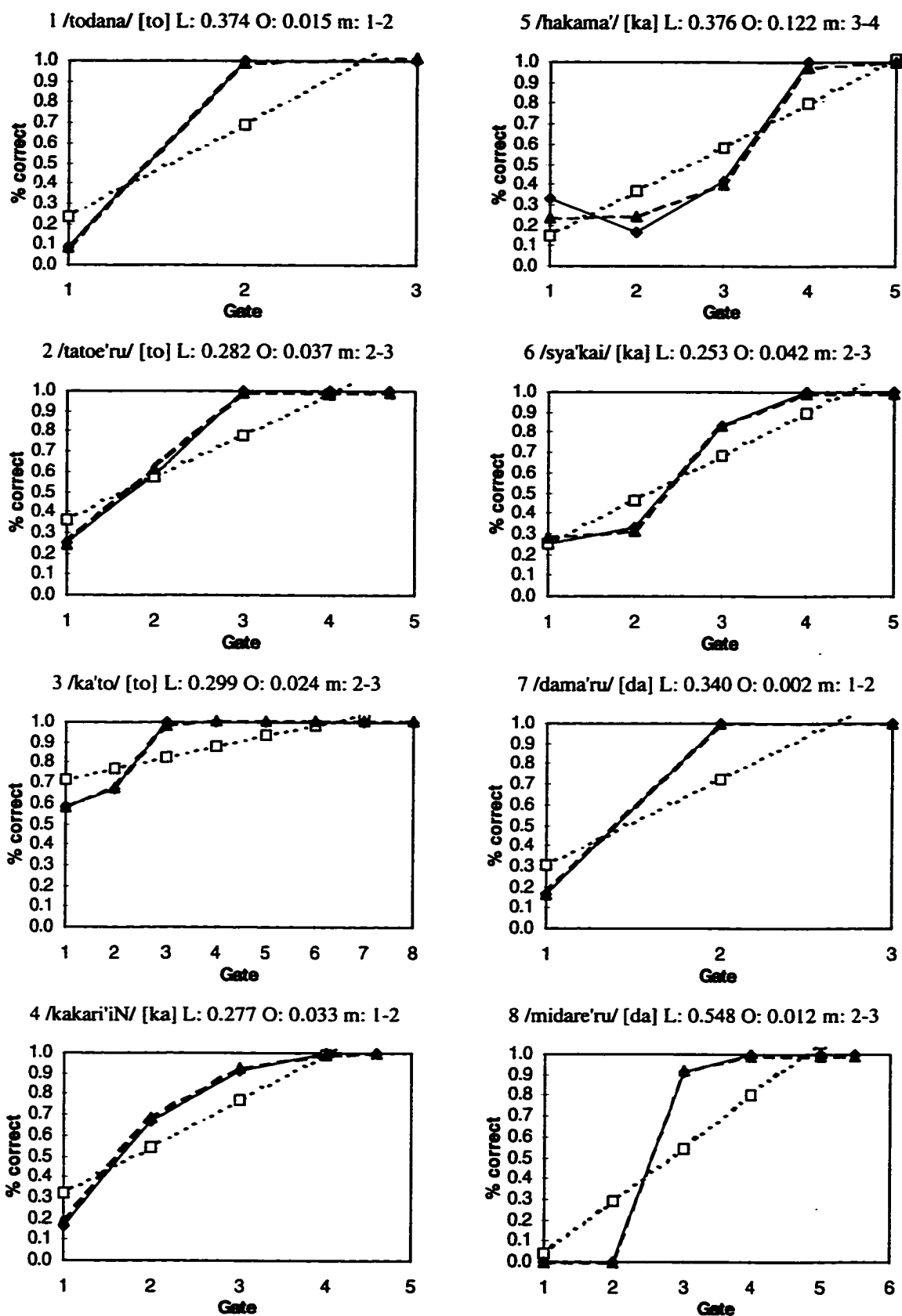


75 /maNne'Nhitu/ [an] L: 1.305

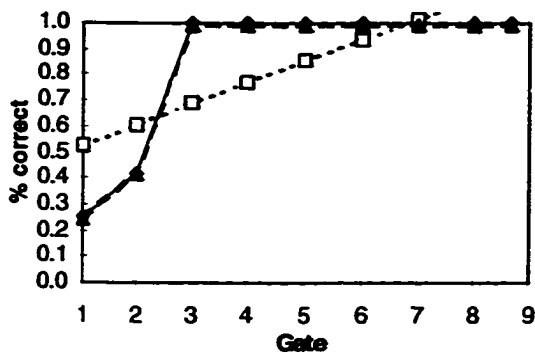


76 /seNmoN/ [em] L: 1.911 O: 2.247 m: 5-6

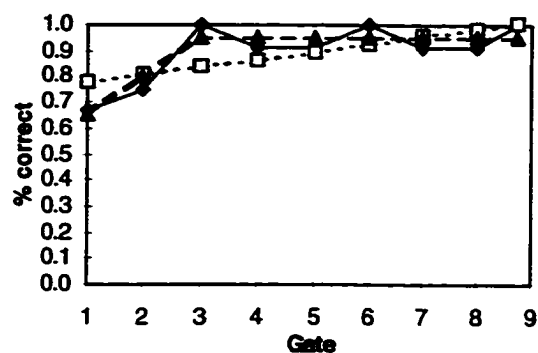




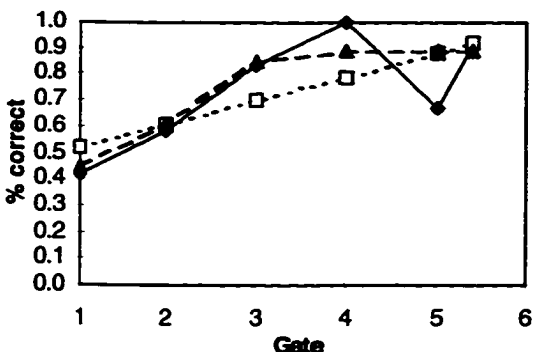
9 /ku'da/ [da] L: 0.566 O: 0.022 m: 2-3



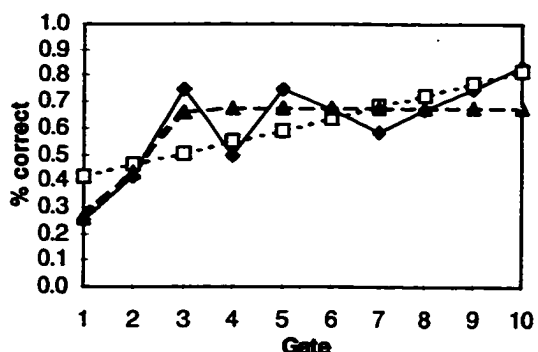
13 /hatake/ [ak] L: 0.238 O: 0.122 m: 2-3



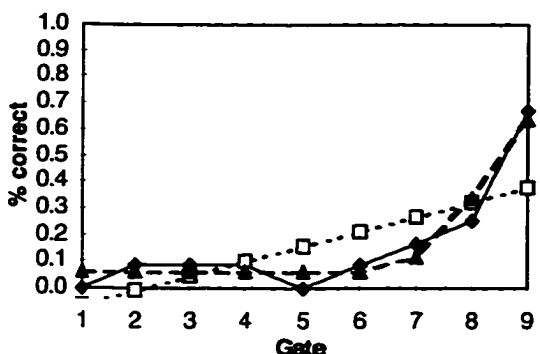
10 /hotoke/ [ot] L: 0.345 O: 0.254 m: 2-3



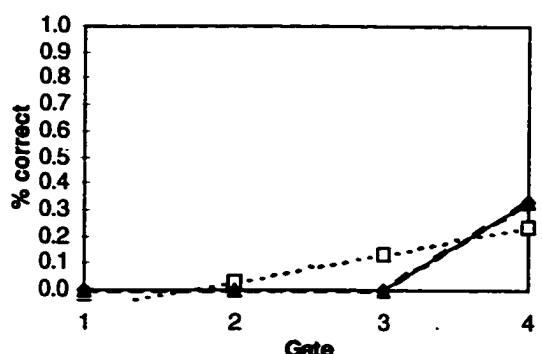
14 /ha'yaku/ [ak] L: 0.363 O: 0.288 m: 2-3



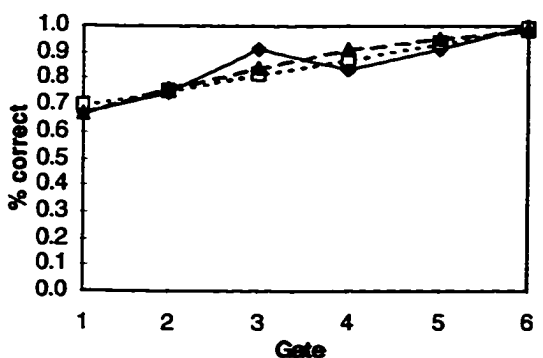
11 /himoto/ [ot] L: 0.393 O: 0.141 m: 8-9



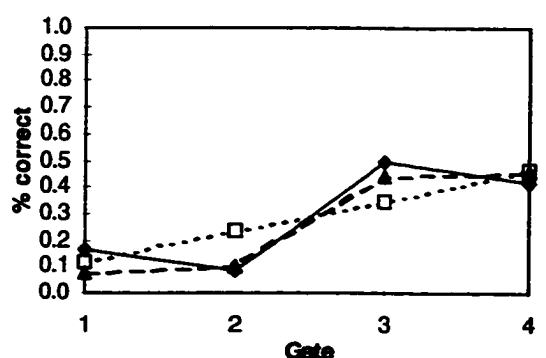
15 /kadai/ [ad] L: 0.183 O: 0.004 m: 3-4

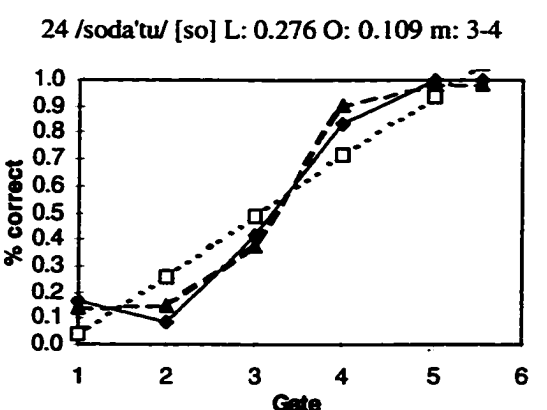
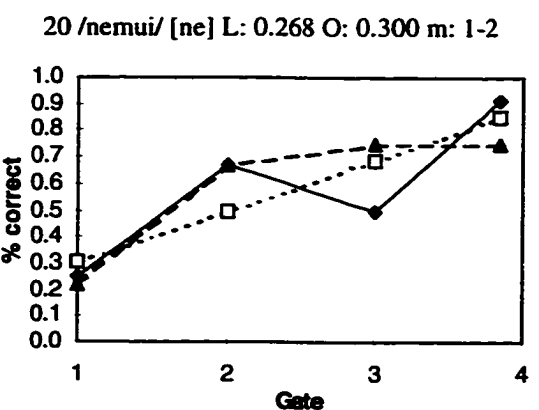
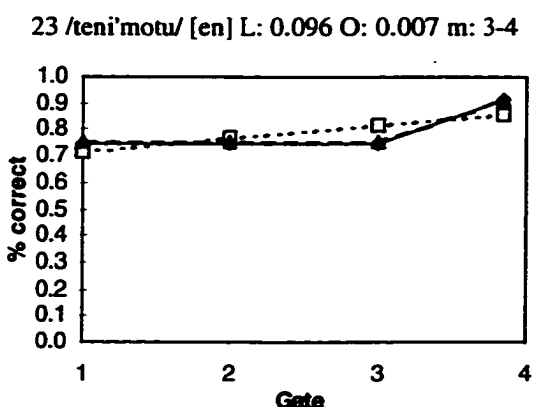
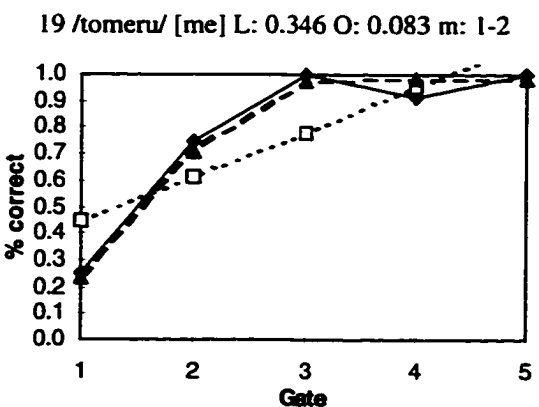
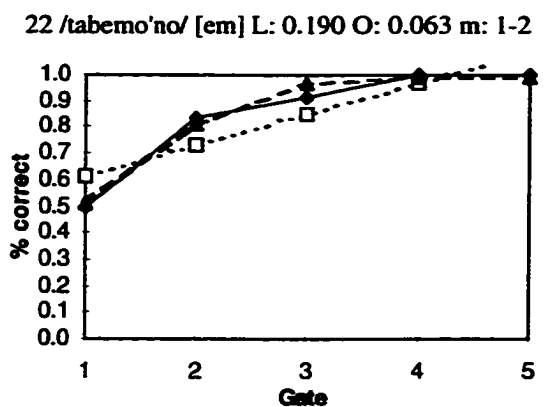
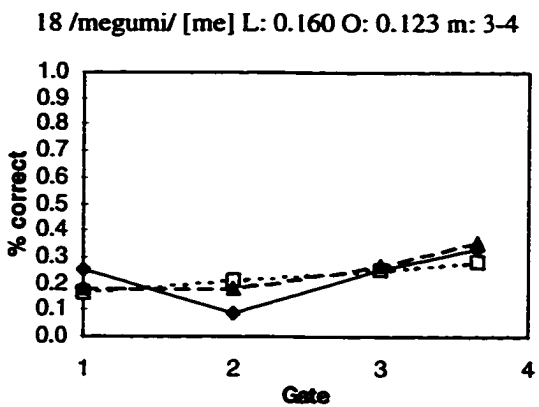
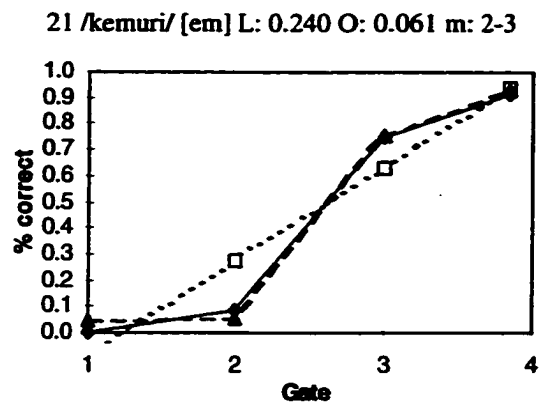
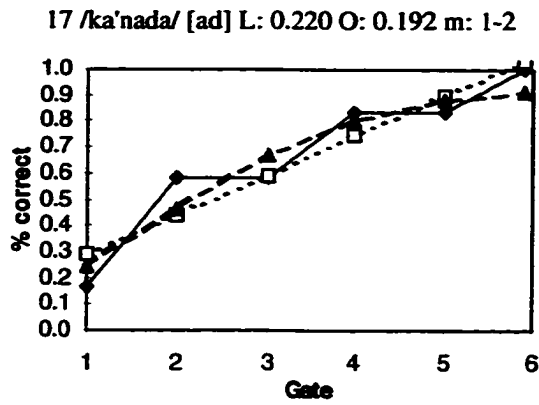


12 /hakobu/ [ak] L: 0.115 O: 0.115 m: 1-2

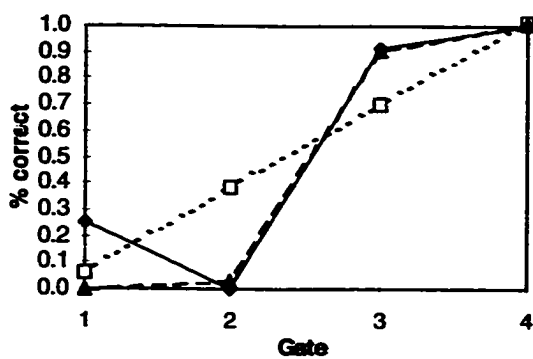


16 /hanada'yori/ [ad] L: 0.224 O: 0.114 m: 2-3

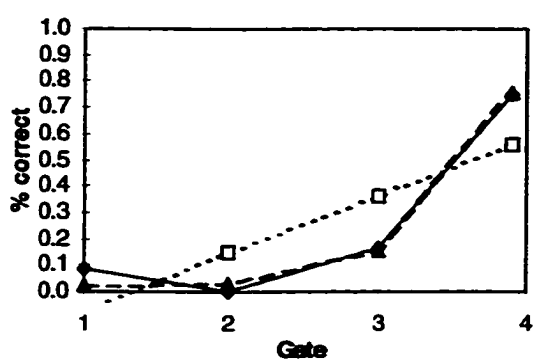




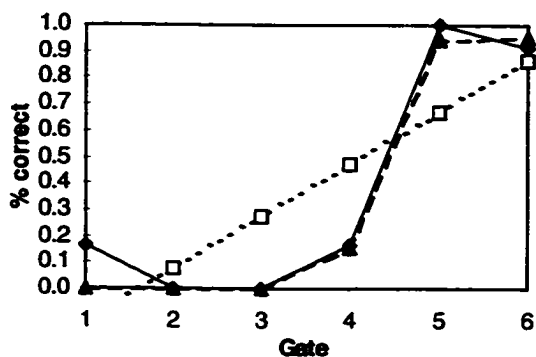
25 /zabu'toN/ [za] L: 0.477 O: 0.252 m: 2-3



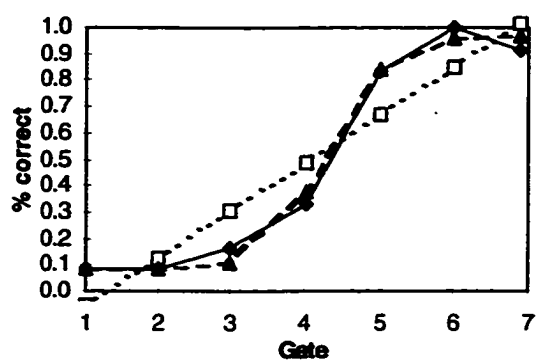
29 /kazari/ [az] L: 0.347 O: 0.066 m: 3-4



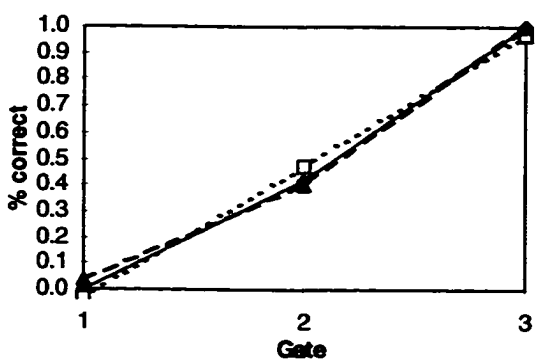
26 /syabe'ru/ [ja] L: 0.607 O: 0.180 m: 4-5



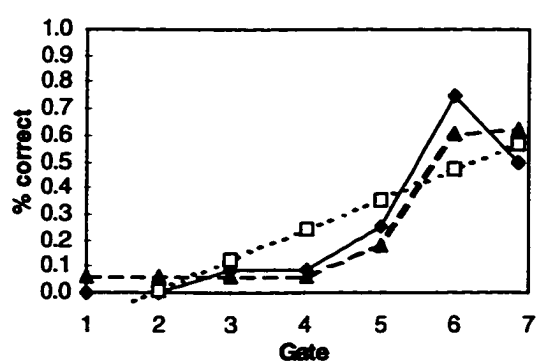
30 /basyo/ [a] L: 0.351 O: 0.094 m: 4-5



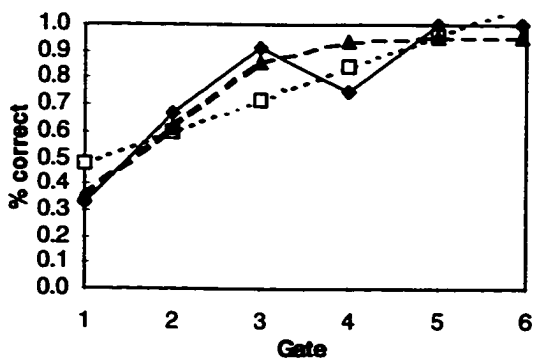
27 /hokeN/ [ho] L: 0.068 O: 0.043 m: 2-3



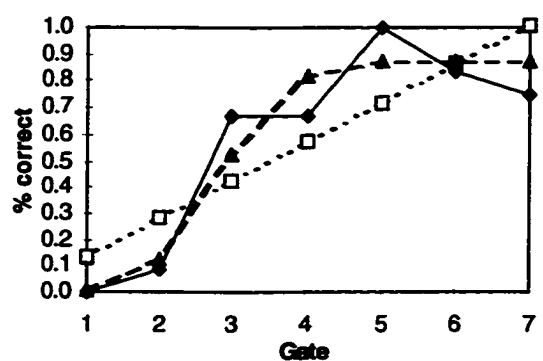
31 /gohoo/ [oh] L: 0.363 O: 0.222 m: 5-6



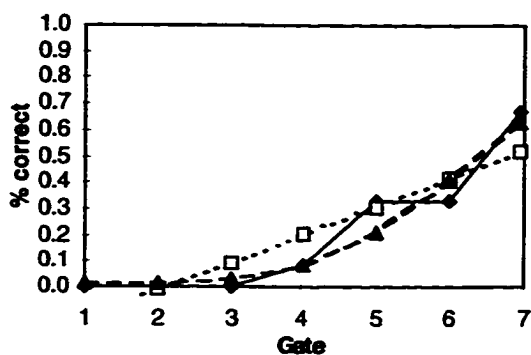
28 /zyosei/ [os] L: 0.282 O: 0.218 m: 1-2



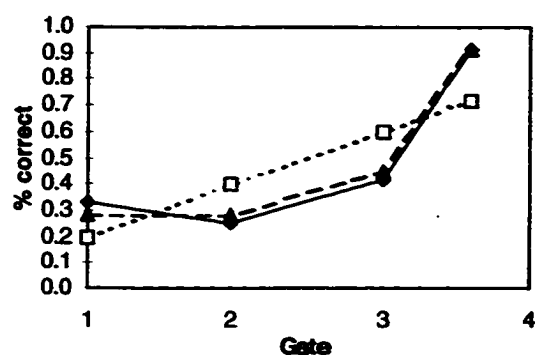
32 /wahuku/ [aφ] L: 0.521 O: 0.281 m: 2-3



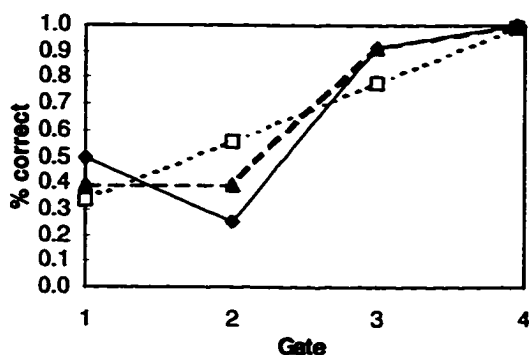
33 /dohyoo/ [oç] L: 0.259 O: 0.155 m: 6-7



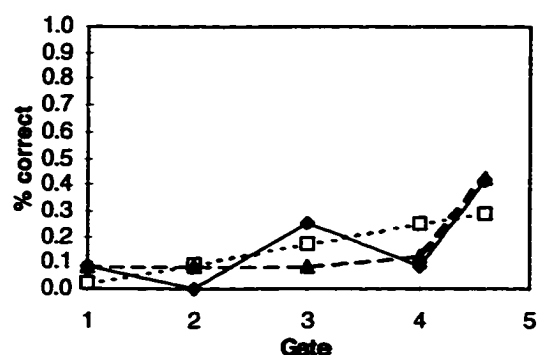
37 /huyoo/ [uj] L: 0.336 O: 0.069 m: 3-4



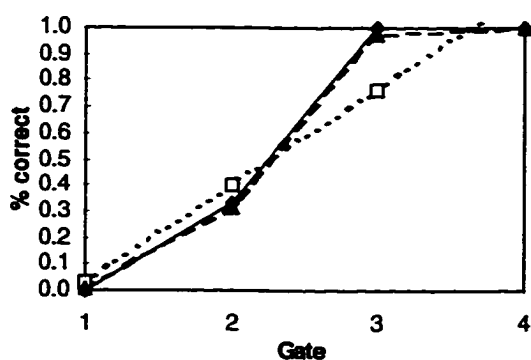
34 /harada'tu/ [ra] L: 0.375 O: 0.182 m: 2-3



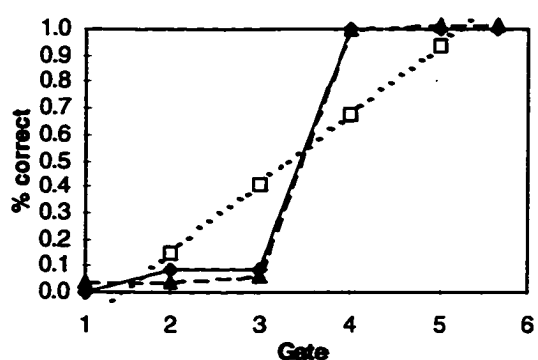
38 /mawari/ [aw] L: 0.248 O: 0.189 m: 4-5



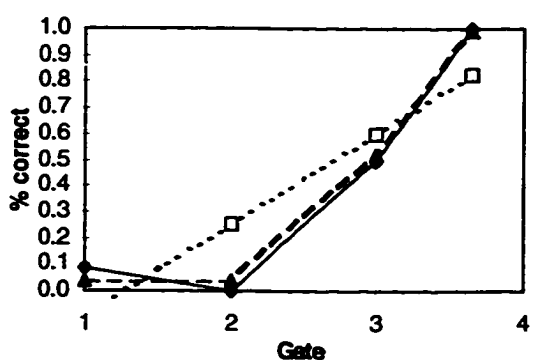
35 /yubi/ [ju] L: 0.279 O: 0.031 m: 2-3



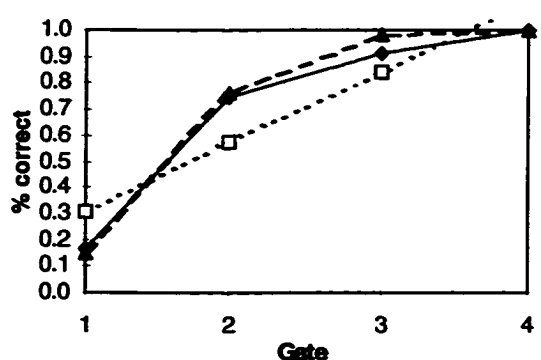
39 /tyazuke/ [tja] L: 0.497 O: 0.069 m: 3-4

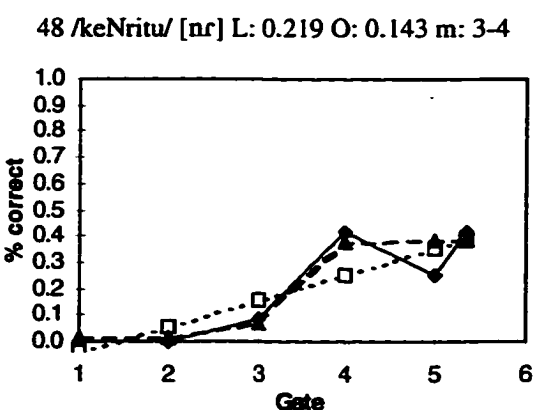
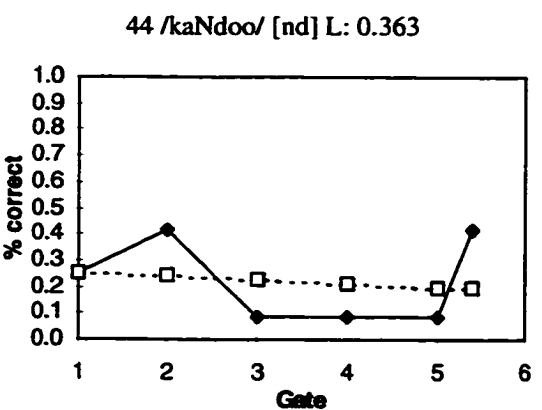
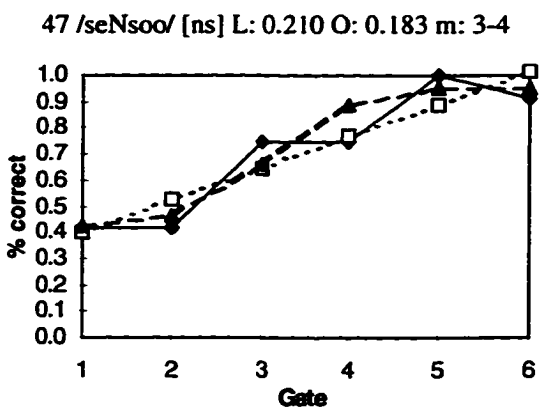
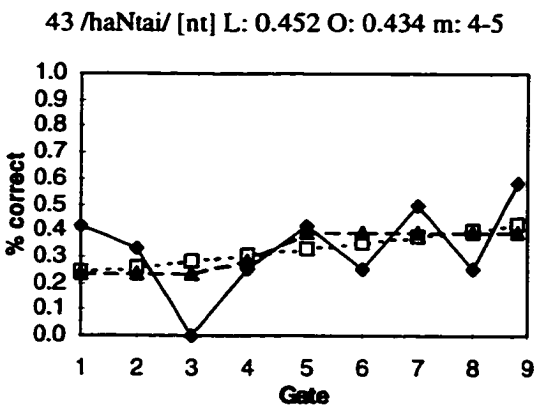
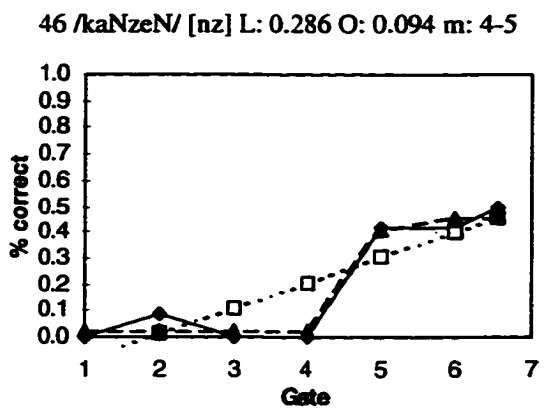
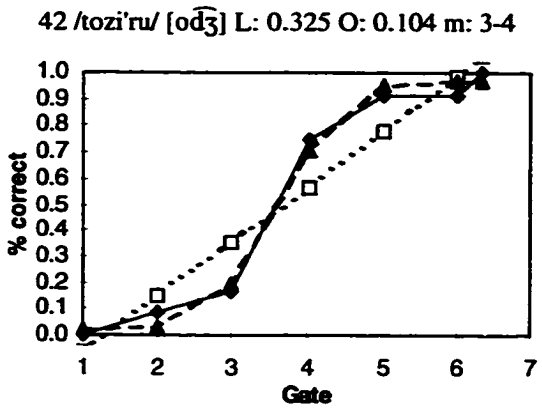
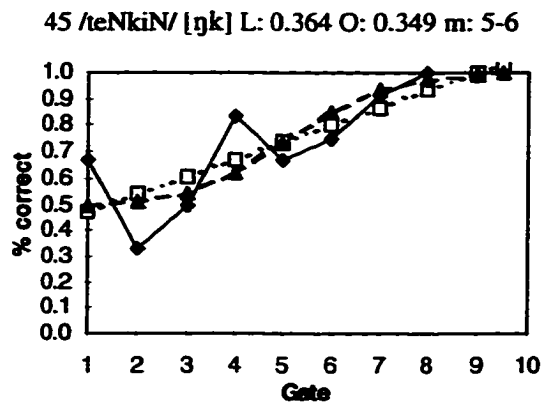
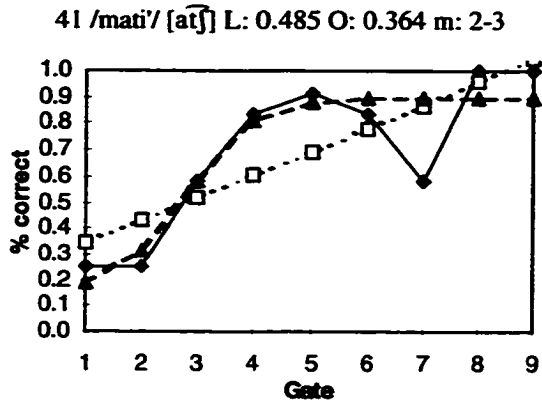


36 /kara'i/ [ar] L: 0.369 O: 0.063 m: 3-4

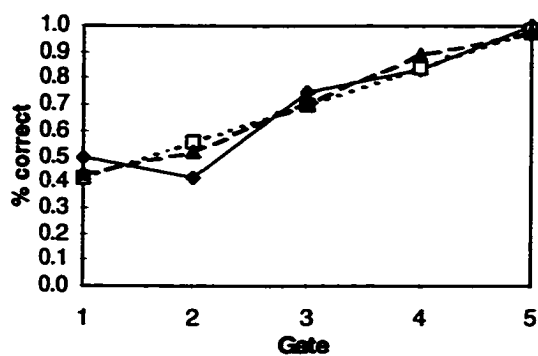


40 /zyokyo'ozyu/ [dzo] L: 0.261 O: 0.069 m: 1-2

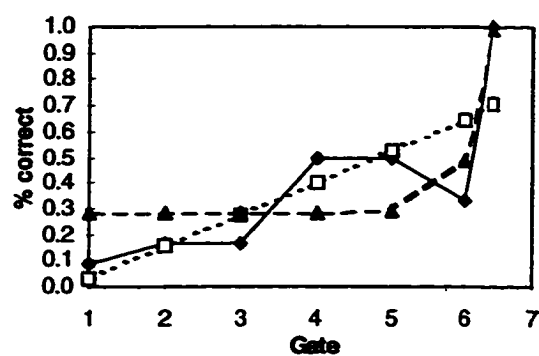




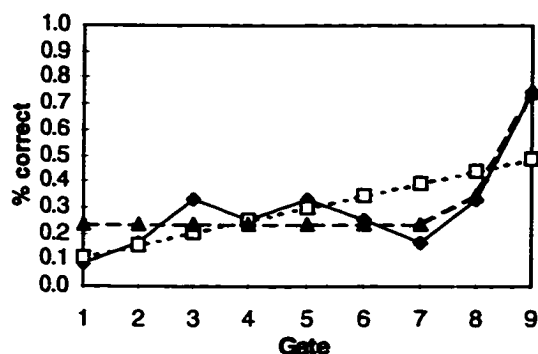
49 /koNyaku/ [ɲj] L: 0.173 O: 0.145 m: 3-4



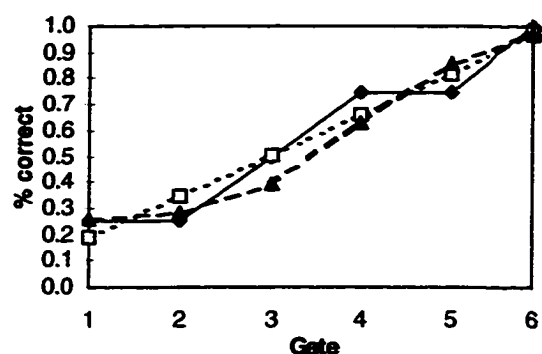
53 /kitamuki/ [kʲt] L: 0.461 O: 0.429 m: 6-7



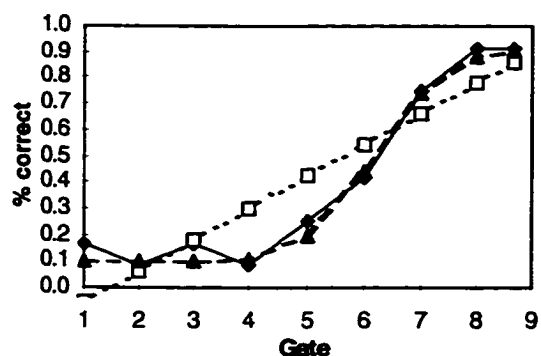
50 /kiNtyoo/ [nt̚] L: 0.399 O: 0.230 m: 8-9



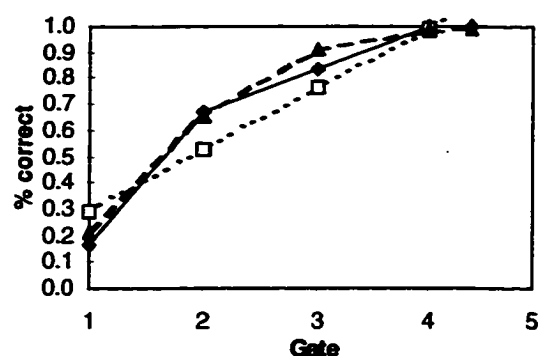
54 /kokutetu/ [kt] L: 0.162 O: 0.201 m: 4-5



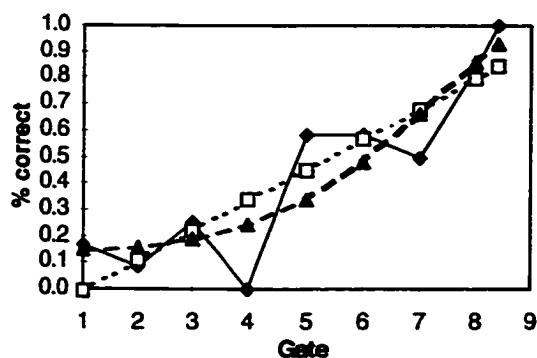
51 /sukunai/ [sk] L: 0.416 O: 0.122 m: 6-7



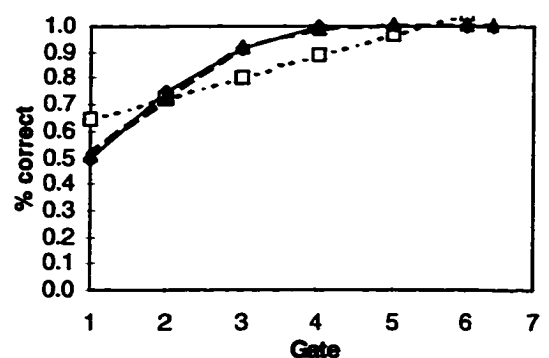
55 /kyaku/ [kj] L: 0.220 O: 0.092 m: 1-2



52 /sikaku/ [ʃʲk] L: 0.471 O: 0.413 m: 7-8

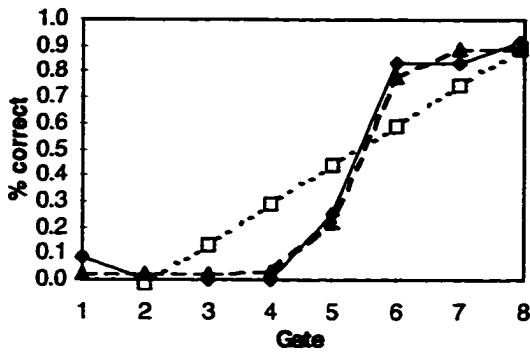


56 /dakyoo/ [kj] L: 0.238 O: 0.029 m: 1-2

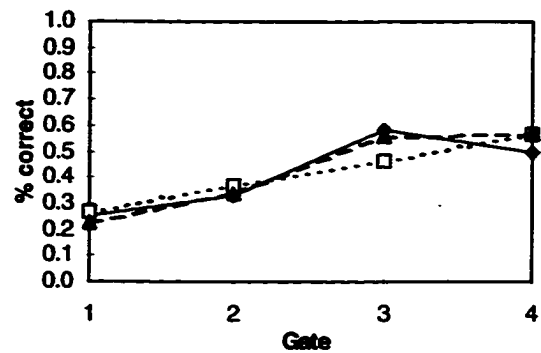




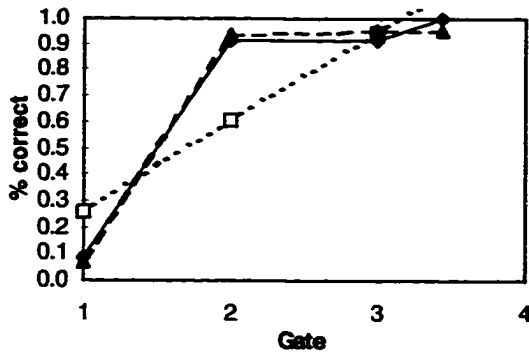
57 /hyoo/ [ɕj] L: 0.518 O: 0.115 m: 5-6



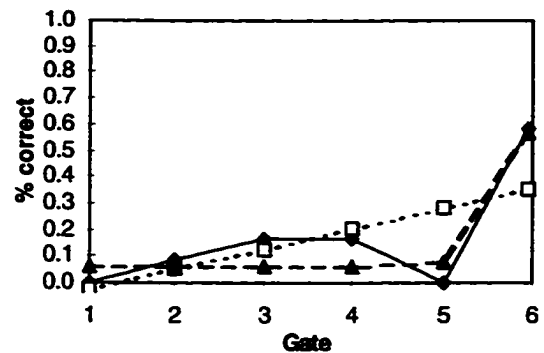
61 /sassoku/ [ss] L: 0.139 O: 0.075 m: 2-3



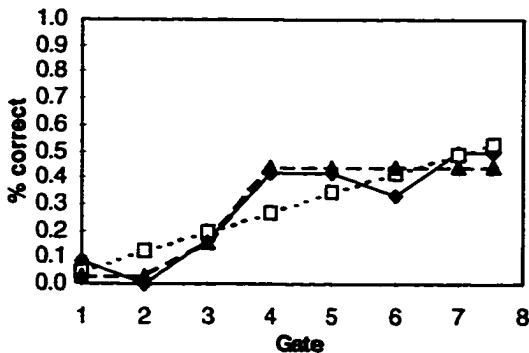
58 /ryokaN/ [ɾj] L: 0.375 O: 0.064 m: 1-2



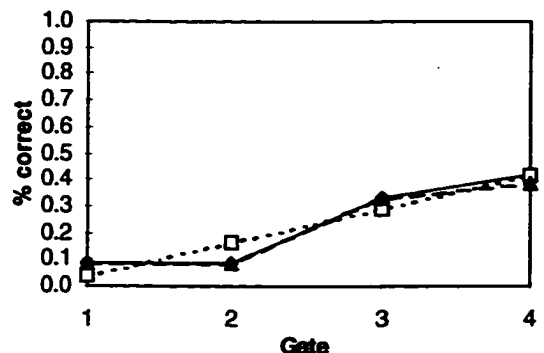
62 /hassya/ [ʃʃ] L: 0.369 O: 0.182 m: 5-6



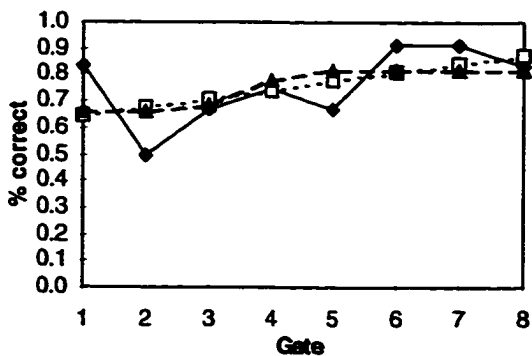
59 /mottaina'v/ [tt] L: 0.228 O: 0.153 m: 3-4



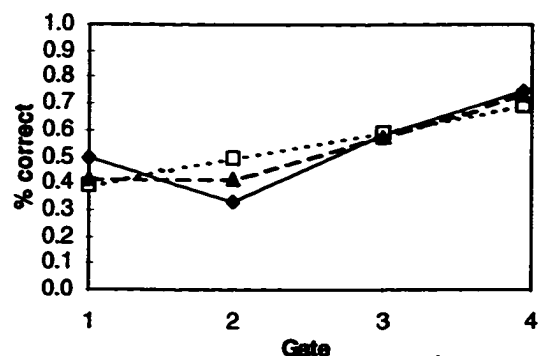
63 /teNmetu/ [mm] L: 0.102 O: 0.034 m: 2-3



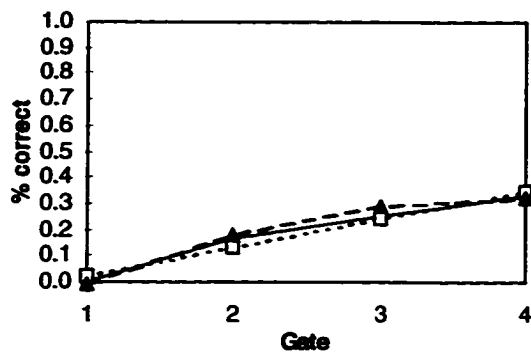
60 /sakka/ [kk] L: 0.316 O: 0.314 m: 3-4



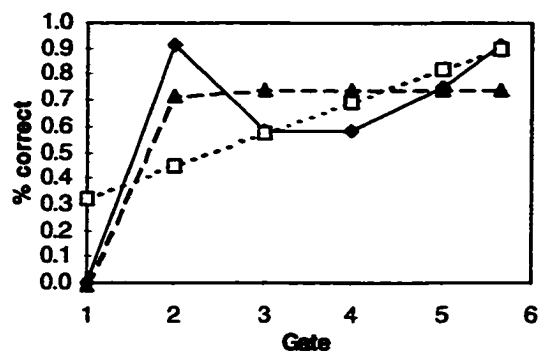
64 /aNnaizyo/ [nn] L: 0.202 O: 0.120 m: 3-4



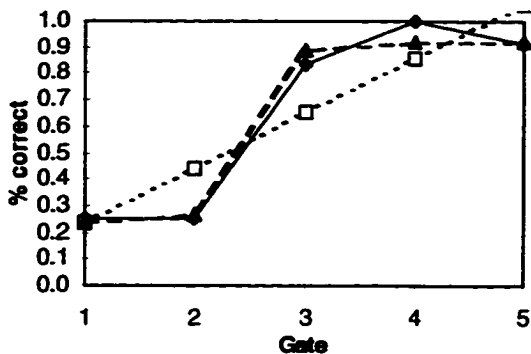
65 /tootyaku/ [oo] L: 0.046 O: 0.042 m: 1-2



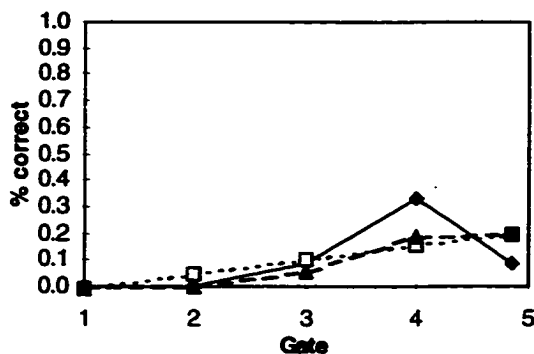
69 /siatu/ [ja] L: 0.585 O: 0.347 m: 1-2



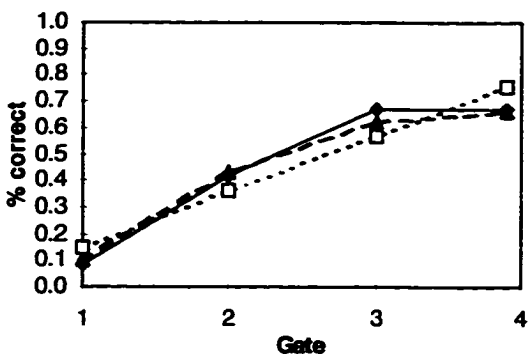
66 /keigo/ [ee] L: 0.336 O: 0.102 m: 2-3



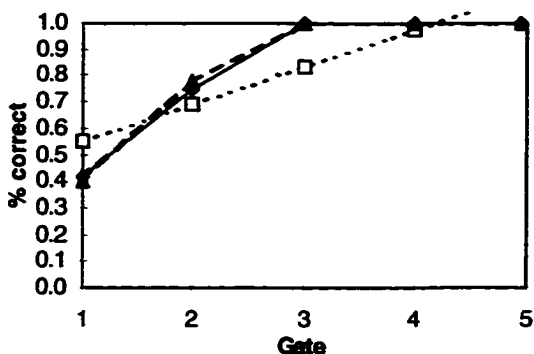
70 /kaeri'miti/ [ae] L: 0.220 O: 0.190 m: 3-4



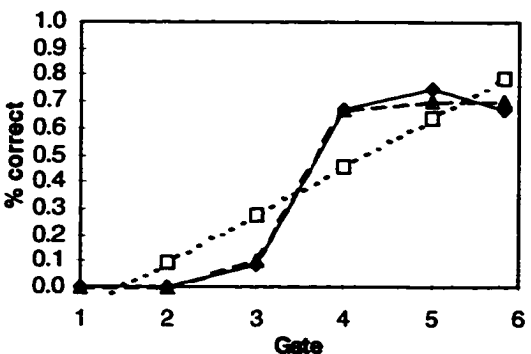
67 /syuukaN/ [uu] L: 0.160 O: 0.052 m: 1-2



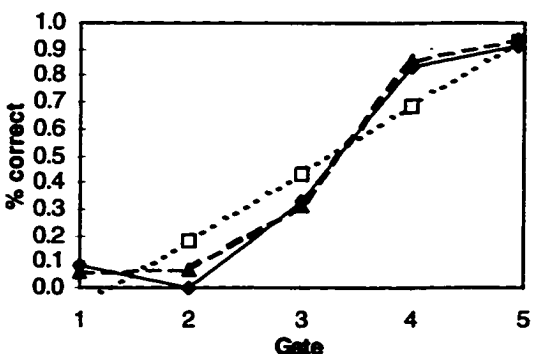
71 /taiko/ [ai] L: 0.248 O: 0.036 m: 1-2



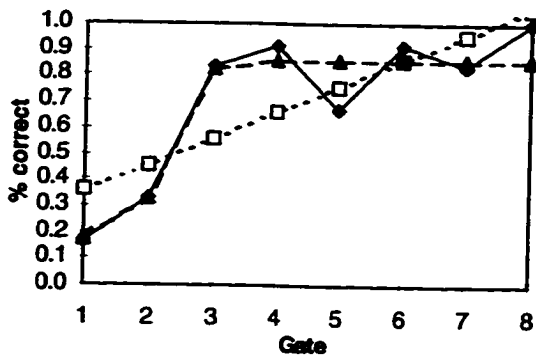
68 /haori/ [ao] L: 0.355 O: 0.061 m: 3-4



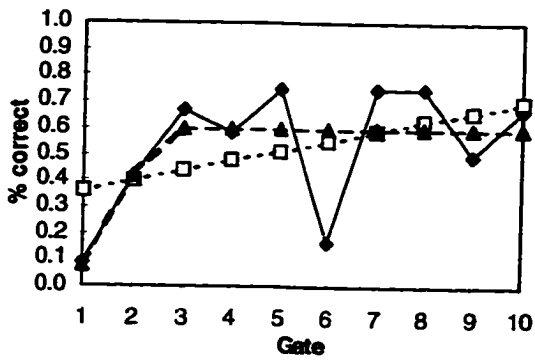
72 /koibito/ [oi] L: 0.298 O: 0.083 m: 3-4



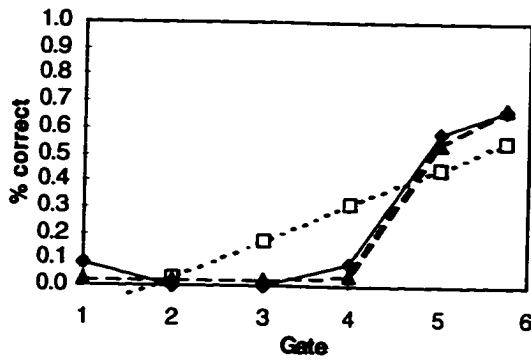
73 /teNiN/ [ēi] L: 0.475 O: 0.254 m: 2-3



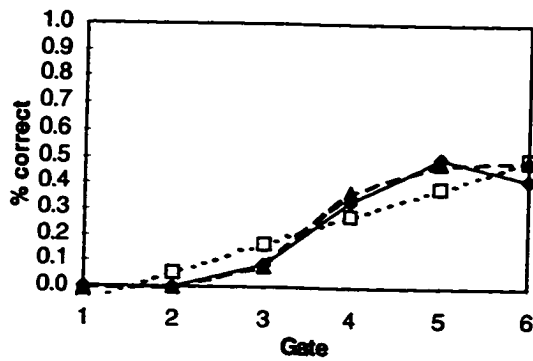
74 /hiN/ [iŋ] L: 0.644 O: 0.524 m: 1-2



75 /maNne'Nhitu/ [an] L: 0.386 O: 0.089 m: 4-5



76 /seNmoN/ [em] L: 0.188 O: 0.081 m: 3-4



### Appendix C: Graphs of results for word initial stops

The following pages show graphs of the data for each word initial stop for the percent correct measure (%Corr). Each graph shows the data points (solid line with diamond markers), the best linear fit (dotted line with unfilled square markers), and the fitted ogival curve (dashed line with triangular markers). No data was excluded because of an anomalous slope. Data which was better fit by a line than by an ogival curve does show both the linear fit and the unsuccessfully fitted curve.

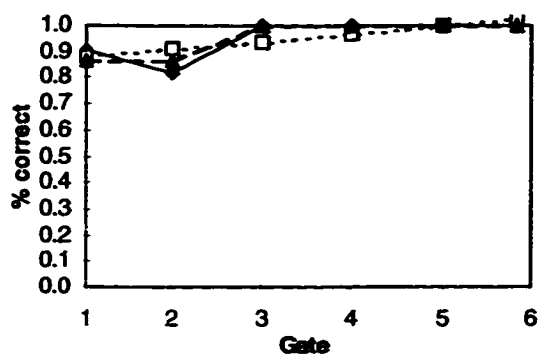
Above each graph is the number and name of the word (numbers keyed to the tables in Chapter 4), and the word initial transition. The number after the capital L is the least squares error of the linear fit, the number after the capital O is the least squares error of the ogival fit, and the numbers after the small m are the area of maximal slope of the fitted curve, in gate numbers.

For each graph, the x-axis represents gate number (time of endpoint of gate), and the y-axis is the percent of responses with the word initial stop correct.

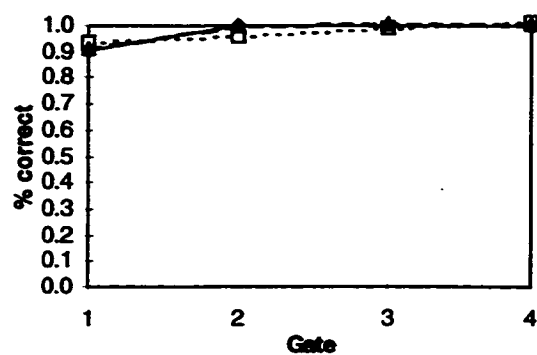
The graphs appear in the following order:

1. English word initial stops
2. Japanese word initial stops

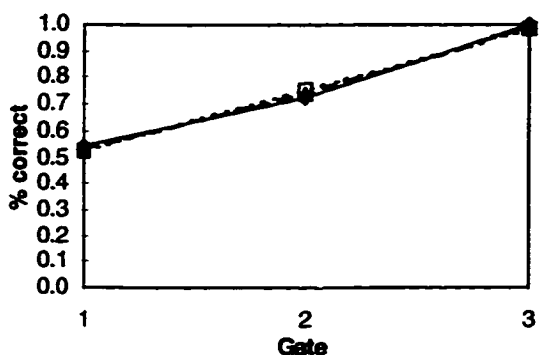
1 tip /tɪ/ L: 0.120 O: 0.065 m: 2-3



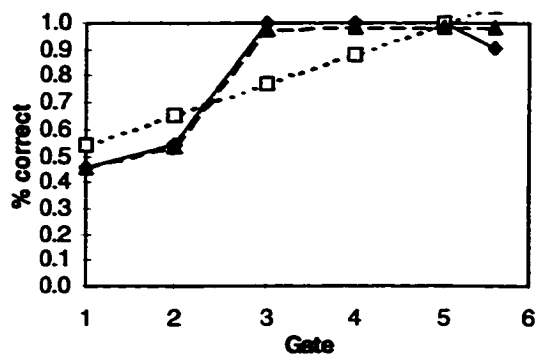
11 duck /dʌ/ L: 0.050 O: 0.015 m: 1-2



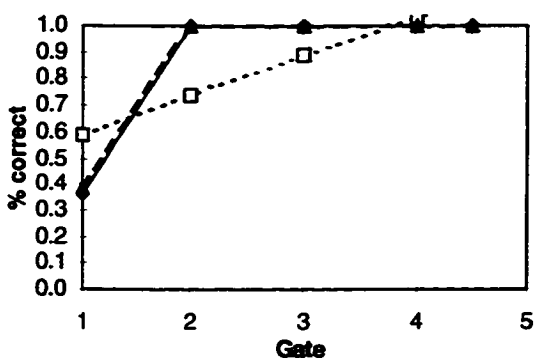
3 Tibet /tɪ/ L: 0.037 O: 0.019 m: 2-3



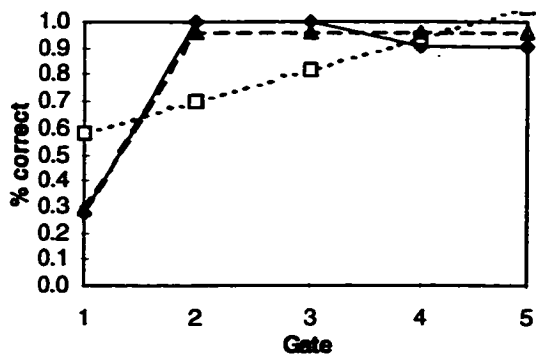
68 train /tɹɪ/ L: 0.333 O: 0.082 m: 2-3



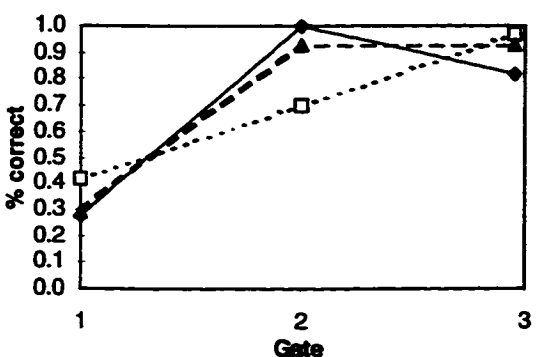
6 custom /kʌ/ L: 0.382 O: 0.019 m: 1-2



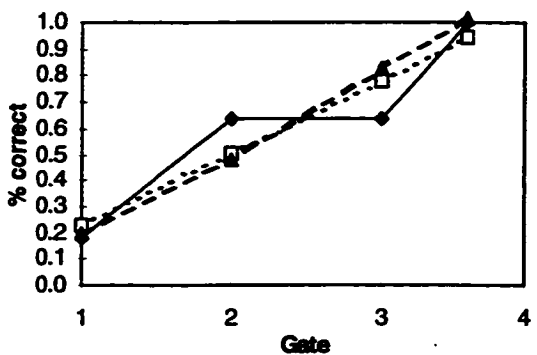
71 crops /kr/ L: 0.496 O: 0.093 m: 1-2

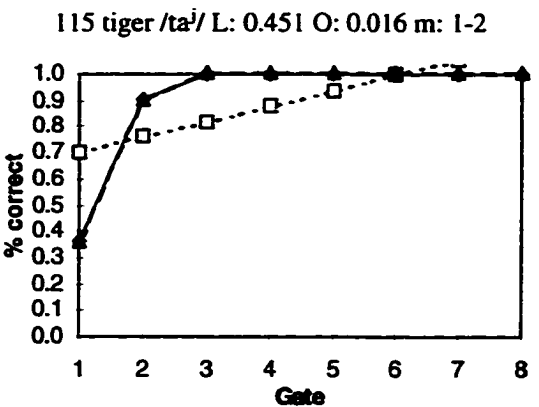
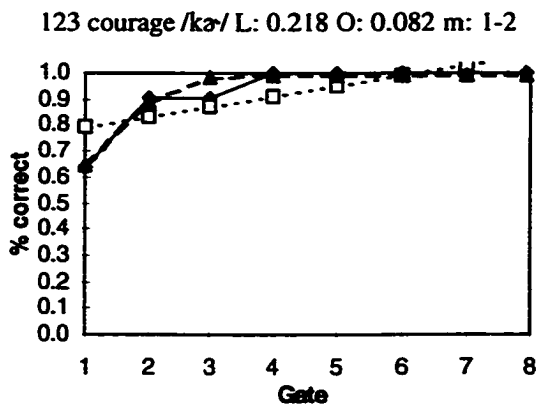
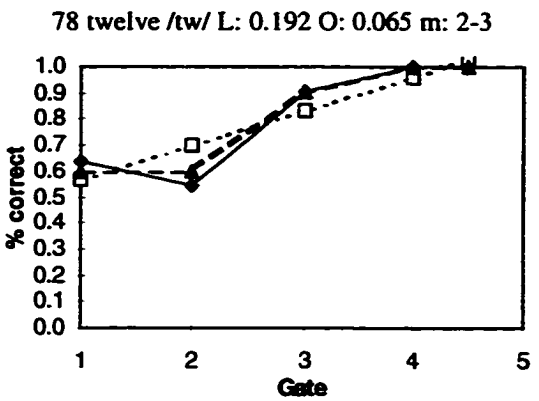
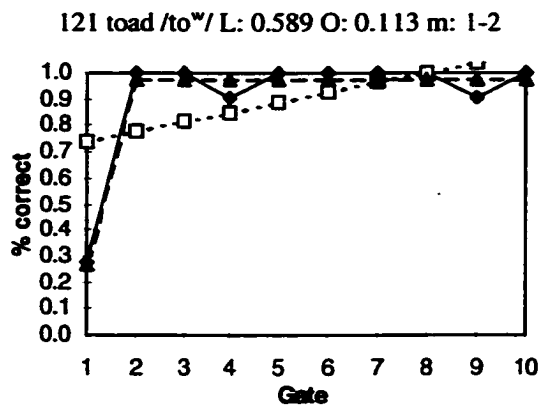
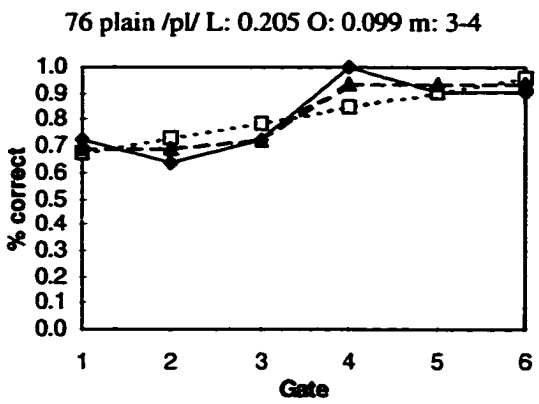
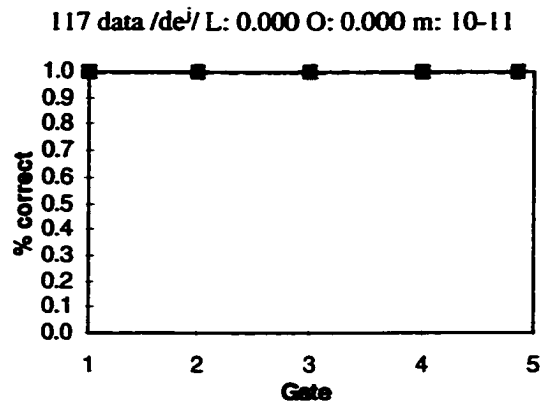
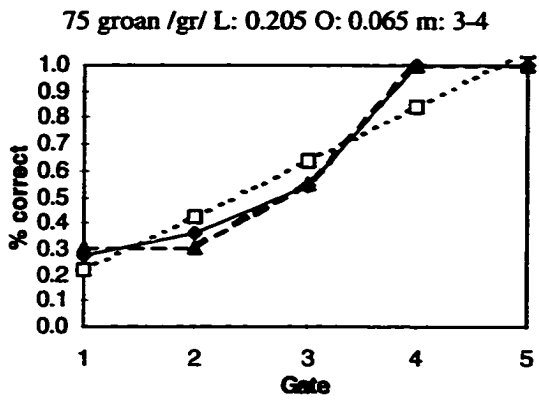


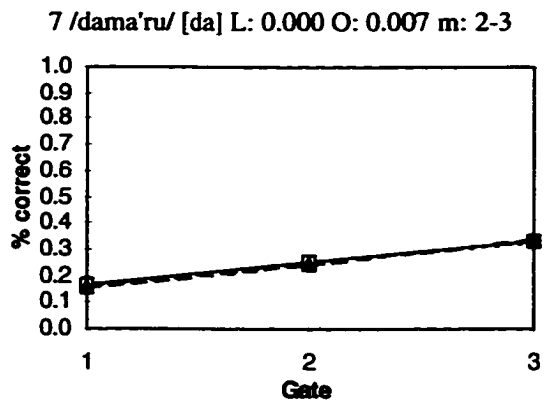
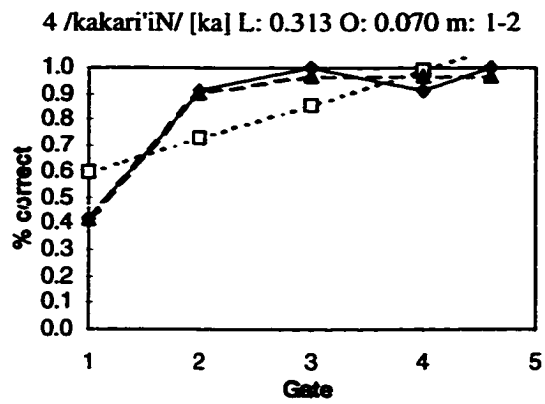
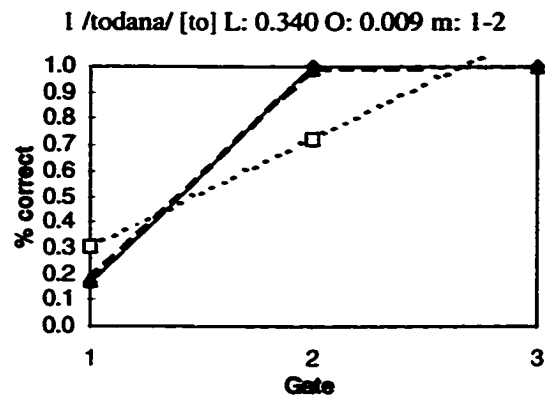
9 caboose /kə/ L: 0.365 O: 0.134 m: 1-2



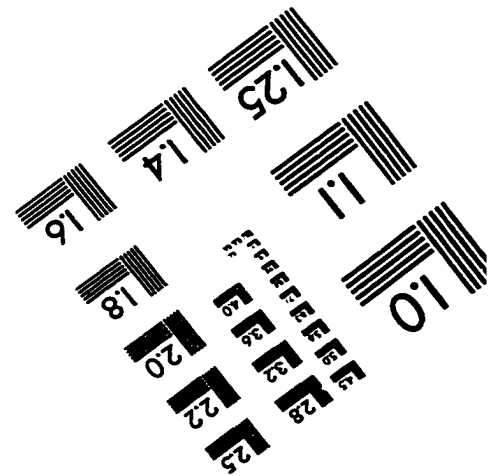
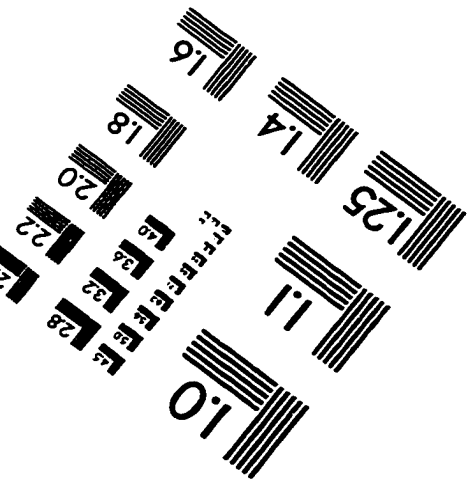
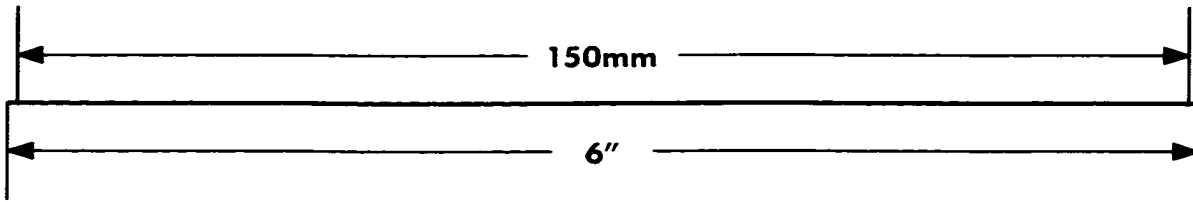
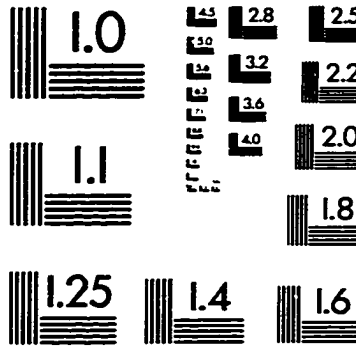
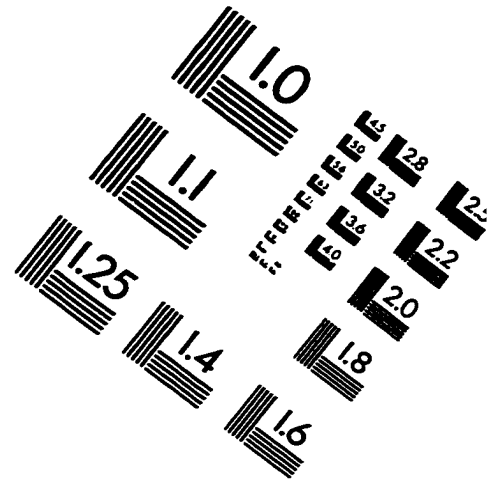
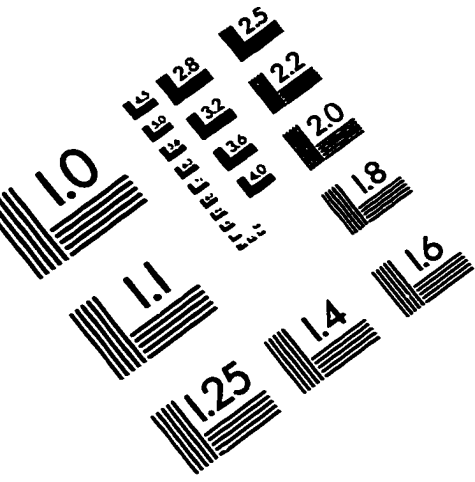
74 drop /dr/ L: 0.208 O: 0.251 m: 2-3







# IMAGE EVALUATION TEST TARGET (QA-3)



APPLIED IMAGE, Inc  
1653 East Main Street  
Rochester, NY 14609 USA  
Phone: 716/482-0300  
Fax: 716/288-5989

© 1993, Applied Image, Inc., All Rights Reserved