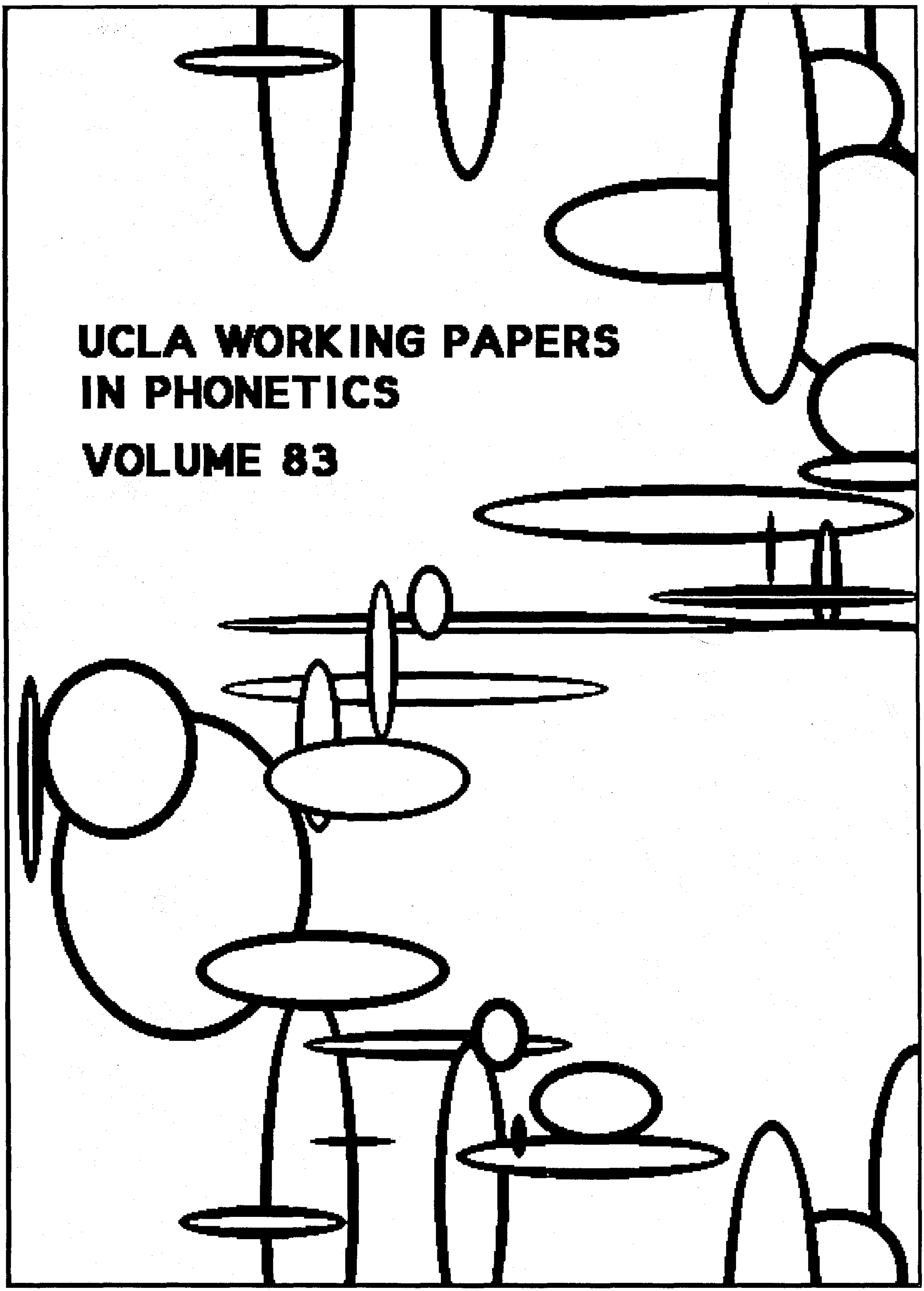
UCLA WORKING PAPERS IN PHONETICS

VOLUME 83



UCLA WORKING PAPERS IN PHONETICS, VOL. 83

February, 1993

Table of Contents

THE STRUCTURE OF SEGMENT SEQUENCES*

Ian Maddieson

INTRODUCTION

Several influential phonetic theories, such as the Quantal Theory proposed by Stevens [1] and Lindblom's Theory of Adaptive Dispersion [2], suggest that there are significant ways in which the sound structure of language is affected by optimizing the sequencing of sounds. That is, differences between possible segment sequences with respect to what might be broadly termed 'efficiency' have played an important role in shaping the syllabic and segmental structure of languages. Despite the important role that is assigned to sequential preferences in these theories, relatively little is actually known about the cross-linguistic frequency of segment sequences in the lexicon. For that matter, we do not even know much about the relative lexical frequency of individual segments in a broad sample of languages. In order to remedy these deficiencies, the lexical frequency of individual segments and of each particular syllable is being counted in a sample of languages constructed so as to draw from all the major areal and genetic groupings of the world's extant languages. The set of languages includes those shown in (1), which as you can see represents a diverse collection of languages.

(1) **Asia:** Standard Chinese, Tibetan, Korean, Thai, Jeh, Wa, Nyah Kur, Sora, Kannada, Darai
The Americas: Comanche, Totonac, Quechua, Navaho, Kwakw'ala, Yupik, **Shipibo, Pirahã**
Africa: !Xóõ, Ngizim, Maninka, Gbaya, Kanuri, Igbo,
Europe: Turkish, Polish, Finnish
Pacific: Fasu, **Hawaiian, Kadazan, Rotokas**

METHOD

To collect reasonably comparable data on such a varied set of languages a rather diverse range of sources of information have been used. Generally the sources are printed or computer-readable dictionaries or which provide a phonetic or phonemic transcription of entries, or employ an orthography that is interpretable in such terms. Vocabulary size is at least 2000 lexical entries. Recent loan words, especially those of wide international currency relating to modern technological, political or cultural concepts ('telephone', 'democracy', 'football'), have been excluded wherever recognizable. All remaining items are coded in an ASCII-based segmental transcription in which the division into syllables, and the division within the syllable into onset, nucleus and coda (if any) is represented. A relatively simple program, incorporating some error-checking, is used to count the number of occurrences of each onset, nucleus, and coda type, as well as each front-end demisyllable (onset-nucleus combination) and each back-end demisyllable (onset-coda combination). Complete syllables can also be counted. The result is a survey showing segment and sequence frequencies in the lexicon of a sample of languages that is free of particular areal biases.

The ruling assumption underlying this work is that examining the relative frequency of sequences provides a tool for drawing inferences about the nature of the selectional preferences that shape the sound structure of languages. Whereas much previous analysis of segment sequences has focussed on permitted **sequences** of **classes** of consonants, frequently summarized in terms

* Paper presented at the Second International Conference on Spoken Language Processing, October 12-16 1992 in Banff, Alberta, Canada.

of a scale of 'sonority' or consonant strength, the present project focusses most attention on consonant-vowel pairings. Moreover, rather than looking for simple prohibition or permissibility of the occurrence or sequencing of particular features, it looks at relative frequency. It seems likely that examining a continuum from more acceptable to less acceptable (of which the end-point is prohibition) will reveal more subtle detail. CV is the universally-present syllable type, and, by hypothesis, the same forces that shape preferences within this type of syllable also apply to more elaborated syllables with structures CVC, CCV, etc. For this reason our first analyses have concentrated on the interpretation of patterns in CV sequences.

An initial analysis, published earlier this year [5], of the five languages in the list shown earlier that are printed bold— those which constitute a subset with small segment inventories and simple phonotactics — indicated that the frequency of particular syllable types can be better predicted from the overall segment frequencies than from the assumption that the adjunction of articulatory similar segments is preferred, as had been suggested by Janson. However, analysis of a larger number of languages is now beginning to confirm that some quite strong effects of 'articulatory economy' can be seen, especially in relation to the pairings of vowels and dorsal consonants. At the same time certain strong effects that support theories giving greater weight to auditory/acoustic factors can also be seen. A pattern suggestive of a complex interplay between articulatory and auditory factors as proposed by Lindblom is emerging. In addition, cognitive factors seem to favor more equal usage of segmental sequences than might otherwise be expected, and certain segments seem to have inherent advantages regardless of their environment. There will only be time for some of these issues will be touched on in this paper, which forms one of a series of reports on the ongoing analysis of this data.

ARTICULATORY ECONOMY

We will describe the strongest indication found so far in our data that articulatory economy affects frequency of segments sequences. This concerns the interaction between vowel features and dorsal consonants. Dorsal consonants are defined as consonants in which the primary articulation involves a raising of the body of the tongue. These are velar, uvular and palatal (including palato-alveolar) consonants. Most of the world's languages have velar consonants, whereas fewer have palatals or uvulars; we will therefore focus on velars. A drive to economize on articulatory effort could be expected to result in favoring the combination of velar consonants with back vowels, especially high back ones, or to result in modifying the velar consonants when front vowels occur. In some of the languages surveyed, such as Standard Chinese and Polish, palatal or palatalized consonants occur before /i/ to the exclusion of velars. The standard historical interpretation is that earlier velars were modified in this environment, resulting in the present distributional pattern. Our sources do not represent smaller degrees of coarticulatory accommodation between consonants and vowels, such as that seen in English word pairs such as '*key*' and '*coo*'. However, if we count the relative frequency of demisyllables consisting of a voiceless velar stop followed by a high back vowel and of a voiceless velar stop followed by a high front vowel across a set of languages, a preference for the former rather than the latter can be demonstrated. Such preferences can be shown by the following calculation. The overall probability of C_i in onset position of a demisyllable and the overall probability of the V_i as nucleus are calculated by dividing the number of observed occurrences into the total number of syllables counted in the given language. These probabilities are multiplied together to predict an expected number of occurrences of C_iV_i under the assumption that the segments in CV sequences are independent of each other. The expected and observed frequencies of a given CV sequence are then compared. A cross-linguistic preference for a given sequence is indicated by a consistent trend for the observed frequency to be more than 100% of the expected; avoidance is indicated by a trend for the observed to be less than the expected. The comparison of *Ku and *Ki frequency in the lexicon of sixteen of the languages studied is given in Table 1. The numbers here are the percentages just described, that is, the observed number expressed as a percentage of the expected number, for voiceless velar stops followed by /u/ (or the next highest back vowel if no /u/ occurs) (represented *Ku, where *K represents /k/ plus sometimes other of the velar stops occurring in the language) and followed by /i/ (*Ki).

Table 1. Velar stop plus high vowels (observed as percentage of expected).

	*Ku	*Ki	
Std Chinese (+ /k^h/)	202	(0)	
Ngizim	170	113	(+ /I/)
Comanche	159	66	
Shipibo (/o/)	139	47	
Gbaya	131	95	
Turkish	130	71	
Rotokas	116	41	
Kadazan	111	78	
Wa (+ /k^h/)	110	103	
Korean (all velars)	109	102	
Kannada	100	45	
Polish	100	(0)	
Igbo	95	38	
Fasu	71	71	
Thai (+ /k^h/)	67	75	
Darai	36	39	

Note: additional voiceless velars besides /k/ are indicated in parentheses after language names.

As the table shows, there isn't a uniform preference for *Ku sequences; the percentages in thr first column vary quite widely, and the last four languages even show numbers less than 100%. Nonetheless, a predominant proportion of the languages show more than the expected number of *Ku sequences. In sharp contrast, the majority of languages show a dispreference for velar stops before high front vowels. In the two cases of Standard Chinese and Polish this is a straightforward phonological prohibition, but in most the sequence is permitted but is just exploited more rarely in the vocabulary of the language than free combination would predict.

A simple statistical test of the strength of this effect can be carried out by comparing the mean *Ku and *Ki scores with the 'expected' value of 100%. The mean of the first column in Table 1 is just over 115%, which is not significantly different from 100% by a one-tailed t-test (t (15) = 1.507, p = .076) though there is a trend in the direction of favoring *Ku. The 'avoidance' of *Ki is shown clearly in the second column where the mean, treating the two zeroes as missing values, of just over 70% is highly significant (t (13) = -4.304, p = .0004). Since velar consonants are produced with rearward motion of the tongue body, and the front/back vowel contrast also involves tongue body motion, it is natural to interpret this pattern as resulting from a tendency to enhance articulatory efficiency. The shorter articulatory movement for the tongue body between back consonant and back vowel is favored over the greater articulatory movement from a back position for the consonant to a front position for the vowel.

In contrast to dorsal consonants, the prime coronal consonants, such as alveolar or dental stops, do not involve an essential position of the tongue body, except to provide a 'platform' for tip or blade raising. That is, though a fronting movement might be required to reach articulatory contact, the body of the tongue is not raised. Therefore coronals might be expected to be much more independent — from the articulatory point of view — of the vowel environment. An approach to investigating this point is reflected in the data shown in Table 2, which gives percentages of the 'expected' occurrence of the sequences of voiceless alveolar or dental stops before high front and high back vowels, symbolized as *Ti and *Tu respectively, in 17 languages.

Table 2. Coronal stops plus high vowels (observed as percentage expected)

	*Ti	*Tu
Shipibo (/o/)	239	36
Thai (+ /t^h/)	185	100
Darai	152	45
Rotokas (/t/ > [s] _ i)	134	82
Gbaya	130	146
Std Chinese (+ /t^h/)	113	133
Kannada	120	69
Kwakw'ala (+ /t'/)	93	133
Kadazan	92	140
Turkish	92	2
Igbo	90	125
Wa (+ /t^h/)	88	100
Fasu	86	117
Ngizim	21	223
Korean (all coronals)	13	56
Comanche	8	66
Polish	(0)	57

Here there is no favoring or disfavoring of the combination of these consonants with either front vowels or back vowels. The mean of neither of these columns is significantly different from 100%. The mean of the first column is 103.5%, again treating Polish as a missing value since distinctively palatalized consonants occur before /i/, (t (15) = .232, p = .4098). The mean of the second column is 95.9% (t (16) = -.322, p = .3759). Both columns show wide variations, and besides the fact that the means are not significantly different from 100%, there is not even persuasive evidence of a general pattern for the values in one column to be typically larger or smaller than those in the other on a language-by-language basis. Compare this with the velar data, where of the 14 meaningful comparisons 11 show a lower percentage for *Ki than for *Ku.

The larger number of languages now studied strengthens the conclusion of Maddieson and Precoda [5, 1992] that these CV frequency biases don't demonstrate a "general strategy to reduce articulatory trajectories between adjacent segments" but rather reveal the influence of a more limited tendency to economize tongue body movements. This in turn strengthens the case of those who argue that there are separate articulatory 'control channels' for the tongue body, the tongue tip and the lips (as in Browman & Goldstein's 'Articulatory Phonology' model). Rather than viewing coronal consonants as having a similar articulatory position to front vowels (as Janson [6, 1986] does) or as formally defining front vowels as coronal in place (e.g. as Clements [8, 1991] does), and concluding that articulatory economy is ineffective here, it may be more insightful to consider that articulatory economy operates mainly within certain functionally defined articulatory subsystems as far as its role in shaping the phonological structure of languages is concerned.

AUDITORY CONTRAST

On the other hand, it appears that overly-minimal articulatory changes between an onset consonant and a following vowel may result in a sequence that lacks sufficient auditory contrastivity to be linguistically desirable. In the languages of our sample, sequences of a vocalic glide and the cognate high front or high back vowel — that is principally the sequences /ji/ and /wu/ — are quite consistently either avoided entirely or sharply lower in occurrence than the frequency of the component segments themselves would predict. In the same languages, the sequences /ju/ and /wi/ can be of high frequency. The data illustrating this point will be discussed more fully elsewhere as there is insufficient time on this occasion. What is obvious about this pattern is that articulatory economy cannot be its explanation, since the movement within /ji/ and /wu/ syllables is very economical. Moreover, it clearly involves the dorsal 'control channel' referred to above in connection with the disparity between *Ki and *Ku sequences. In such a case the needs of a listener for auditorily distinctive sequences take precedence over the speaker-driven preference for a measure of articulatory economy.

SEGMENT PREFERENCES

Yet other patterns in the cross-linguistic survey seem best attributed to inherent properties of certain segments which make them better or worse elements to include in a wide range of structures. Considerable work has been done on developing explanatory models to account for the prevalence of 'triangular' vowel systems that include the vowels /i, a, u/ at the corners. Models emphasizing dispersion [10, 2] and models relying on quantal regions of stability in articulatory/acoustic relations [11, 1] agree closely in predicting the primacy of these three vowel types. Yet there is a broad cross-language tendency for low central vowels of the /a/ type to be more frequent than any other vowel, even the other 'corner vowels' of the familiar triangular systems. This preference can be clearly seen in the seventeen languages included in Table 3, which lists the percentage of all the nuclei counted that are /a/, and provides for comparison with this number the percentage of nuclei consisting of the next most frequent vowel in the lexicon. Note that in this case the numbers are simply the observed percentages not adjusted in any way to reflect expectations.

Table 3. Percentage of syllabic nuclei which are /a/, & next most common vowel.

	% of nuclei consisting of /a/	% of nuclei consisting of second most popular vowel
Fasu	49.8	14.6 (/e/)
Kwakw'ala	37.4	29.6 (/e/)
Ngizim	36.4	20.4 (/u/)
Kadazan	34.4	27.5 (/o/)
Turkish	32.6	22.2 (/e/)
!Xóõ	30.6	14.3 (/u/)
Kannada	29.9	22.0 (/u/)
Polish	28.4	21.0 (/o/)
Hawaiian	27.5	14.4 (/i/)
Rotokas	25.7	18.9 (/o/)
Korean	25.4	17.4 (/i/)
Igbo	21.8	11.8 (/e/)
Gbaya	21.7	14.3 (/i/)
Jeh	16.4	9.1 (/ă/) (7.7 /ə/)
Std Chinese	19.1	18.3 (/i/ + [i̟])
Darai	21.1	20.2 (/i/)
Pirahã	41.6	42.3 (/i/)

Fourteen of the languages in this table show a very marked excess of /a/ over the next vowel (and, incidentally, the next most frequent vowel is not unlikely to be a mid vowel rather than high /i/ or /u/). The final three languages show essentially equal frequency of /a/ and the next vowel ('fricative vowel' allophones of /i/ after sibilants in Chinese have been included with other realizations of /i/). Only two languages so far analyzed show a different nucleus than /a/ to be most frequent. These are languages in which the frequency of one particular syllable accounts for this fact: in Shipibo the (demi-)syllable /ti/ is so frequent in the lexicon that the vowel /i/ is the most common nucleus, and in Comanche the (demi-)syllable /tʉ/ is so frequent that the vowel /ʉ/ is the most common nucleus. A similar 'distortion' due to a single very frequent syllable — in this case /ka/ — can also be considered responsible for the exceptionally high overall frequency of /a/ in Fasu. Despite these aberrant cases, it is clear that /a/ enjoys a disproportionate popularity that dispersion and quantal models don't predict. This parallels the greater chance that this same vowel has of being included in vowel inventories to begin with [12]. It seems reasonable to propose that these preferences follow from the fact that /a/ has greater amplitude than any other vowel, due to the proximity of F1 and F2 and the absence of any attenuation resulting from a narrow outlet. An inherent property, rather than one affected by the relations between vowel and context seems best to account for the disparity in the distribution of /a/ relative to other vowels.

CONCLUSION

The present study is providing evidence for some influence of both articulatory economy and auditory/acoustic distinctiveness in shaping the preferences for sequences, rather than for the dominance of either production-driven or listener-driven factors over the other. The evidence for these processes only emerges from considering a large quantity of data drawn from a sample of languages suitably structured to provide sufficiently independent witness: Language-particular deviations from the general preferences can be quite marked, and any general patterns will only emerge from large samples. Besides sequential preference patterns, preferences for particular segment types are also apparent. This may suggest that explanations for phonological patterning should incorporate an evaluation of the advantages and disadvantages inherent in individual segments, as well as in the transitions between segments in the spoken stream.

REFERENCES

[1] K. N. Stevens. "On the quantal nature of speech." *Journal of Phonetics* 17: 3-45, 1989

[2] B. Lindblom. "Models of phonetic variation and selection." *PERILUS (University of Stockholm)* 11, 65-100. 1990

[3] T. Vennemann. *Die neuere Entwickenlungen in der Phonologie.* Helmut Buske, Hamburg. 1986.

[4] G. N. Clements. "The role of the sonority cycle in core syllabification". In M. E. Beckman and J. Kingston *Papers in Laboratory Phonology I.* C.U.P., Cambridge. 1991.

[5] I. Maddieson & K. Precoda. "Phonetic models and syllable structure" *Phonology* 9, 45-60. 1992

[6] T. Janson. "Cross-linguistic trends in CV sequences." *Phonology Yearbook* 3, 179-196. 1986.

[7] C. Browman & L. M. Goldstein. "Tiers in articulatory phonology, with some implications for casual speech." In M. E. Beckman and J. Kingston *Papers in Laboratory Phonology I.* C.U.P., Cambridge. 1991.

[8] G. N. Clements. Place of articulation in consonants and vowels: a unified theory". Working Papers of the Cornell Phonetics Laboratory 5: 77-123, 1991.

[9] I. Maddieson. “Universals of segment sequences: a cross-linguistic lexical survey”. Paper presented at the Seventh International Phonology Meeting, Krems, July 4-7 1992. To appear in *Phonologica 1992*.

[10] B. Lindblom. “Phonetic universals in vowel systems”. In J.J. Ohala & J.J. Jaeger *Experimental Phonology*. Academic Press, Orlando. 1986.

[11] K. N. Stevens. “The quantal nature of speech: evidence from articulatory-acoustic data.” In E.E. David & P. B. Denes *Human Communication: A Unified View* . Academic Press, London. 1972

[12] I. Maddieson. *Patterns of Sounds.* Cambridge University Press, Cambridge. 1984.

Splitting the mora*

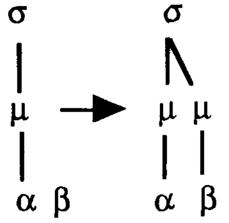
Ian Maddieson
University of California, Los Angeles and Berkeley

Introduction

Different aspects of the control of the timing of speech are reflected by mora count and phonetic duration. Moras are abstract units of timing that express the categorial equivalences between different types of long and short segments, heavy and light syllables, and other metrical structures (Hyman 1985, 1992, McCarthy & Prince forthcoming, Hayes 1989). Within any such category, phonetic duration may be quite variable. For example, a long vowel may be assigned two moras, but the actual duration of any particular vowel will depend on its height, its tonal or accentual properties, its position in the word, the nature of the adjoining segments and many other factors. However, moraic structure and duration show a general relationship. Where moraic structures differ we expect to be able to confirm this by measurements of phonetic duration, provided other factors are equal, and where durations differ we might ask if a difference in moraic structure is responsible.

It is frequently, though sometimes tacitly, assumed that syllables may contain either one or two moras but no more, reflecting the fact that quantitative contrasts in languages seem almost always to function in a binary fashion: Languages have long and short vowels, single and geminate consonants. Following Hayes (1989) and many others, a short vowel may be posited as having one mora, a long vowel two, a single consonant none and a geminate consonant one. In many languages, such as Hausa, a syllable with a consonant in coda position functions for many purposes like a syllable with a long vowel, i.e. both CVV and CVC are heavy syllables and have two moras. In others — of which Lardil is a frequently cited example — CVC syllables are light syllables and pattern like CV. Hayes (1989) proposes that languages of the Hausa type have a "weight by position" rule that assigns a mora to coda consonants (or a subset of them) when they are adjoined to the syllable. The rule is formulated as in (1).

(1) Weight by position (Hayes 1989)



where σ dominates only μ

Because weight by position only adjoins β to a syllable node dominating a single mora, it produces syllables with a maximum of two moras. There may be a few cases where this restriction is too stringent - Hayes cites Estonian - but they are few. We will assume that very strong evidence is required to overturn this assumption for a given language.

Hayes (1989) argues that the process of compensatory lengthening provides good illustrations of the advantages of a moraic theory over several competing alternatives, in particular, predicting that loss of an onset consonant may not create compensatory lengthening since an onset is not moraic. Hayes expresses this prediction in the law of Moraic Conservation in (2).

(2) Compensatory Lengthening processes conserve mora count.

In the present paper we will discuss one of the compensatory lengthening processes in Luganda and Sukuma in the light of Moraic Conservation and sets of measurements on the phonetic durations of affected segments.

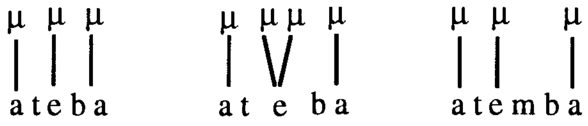
Compensatory lengthening of vowels before prenasalized stops occurs in many Bantu languages: A structure such as CVNSV appears as CV:NSV. In a moraic account, adapted from the analysis of Luganda offered in a different framework by Clements (1986), this process involves reassigning a mora originally associated with the nasal element to the preceding vowel. The nasal element is simultaneously incorporated into the onset of the following syllable. Since tautosyllabic postvocalic consonants generally add weight to the syllable they belong to (supporting the original assignment of a mora to the nasal element), and onsets, even if complex, generally do not add weight to the syllable (supporting the delinking of the mora of the nasal when it is incorporated into the onset) this analysis seems well-motivated. A long vowel resulting from compensatory lengthening receives the same surface representation as a lexical long vowel: both have two moras.

In more detail the process is as follows. Consider the forms in (3). These are verbs preceded by the third singular (class 1) subject marker /a-/.

(3)	a)	/ateba/	"he foils"
	b)	/ateeba/	"he shoots"
	c)	/atemba/	"he climbs"

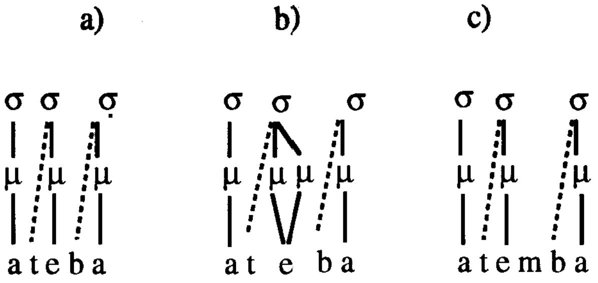
Given that there is an underlying vowel length contrast, vowels must be assigned either one or two moras as appropriate, as in (4).

(4) a) b) c)



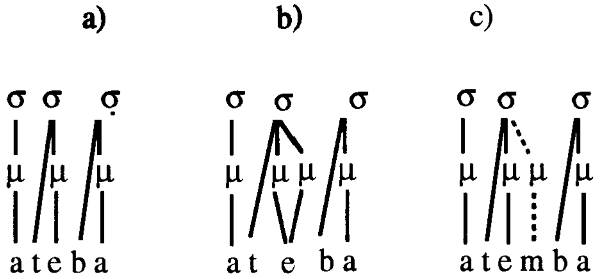
In the first stage of syllabification, a syllable node is created for each vowel. Subsequently, any single consonants preceding vowels are adjoined to syllable nodes, creating onsets, as in (5).

(5)



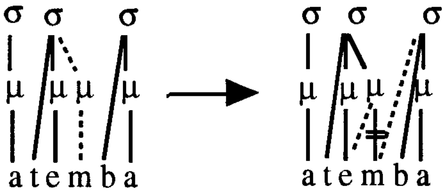
Leaving aside geminate consonants, in Luganda the only consonants not syllabified so far will be nasals, as in /atemba/ "he climbs" in 5 (c). By the ordinary rules of syllabification, such nasals will form codas to the preceding syllable. Luganda has a "weight by position" rule (or its equivalent), so a mora will be projected for the nasal before it is adjoined to the syllable, giving (6).

(6)



Compensatory lengthening involves delinking the nasal from its mora as it is incorporated into the onset of the following syllable. (We will not discuss here the issues involved in this incorporation, but most likely [mb] becomes a single phonological entity). The mora is then stranded, a situation which can only be remedied by linking it to the preceding vowel (otherwise association lines are crossed). Thus we have (7).

(7)



The result of this is that the surface representation corresponding to the second syllable — the /te(e)/ portion — of both 3 (b) /ateeba/ "he shoots" and 3 (c) /atemba/ "he climbs" is the same. They both have two moras linked to the vowel. Both are distinguished from the one-mora vowel in 3 (a) /ateba/ "he foils". Moreover, the constraints of the theory predict that no more than these two phonologically distinctive lengths can occur.

If life was simple, we might expect that underlying long vowels of the type found in 3 (b) and long vowels derived from the compensatory lengthening illustrated in (7) have the same phonetic duration, if other things were equal. To a rough approximation this appears true in Luganda, as was shown in some measurements published without many details by Herbert in 1975. I have done additional measurements on sets of words similar to those in (3). The items used, all verbs, were suggested by Larry Hyman and recorded by Francis Katamba. The verbs were placed in a frame in which tonal distinctions are levelled. Durations were measured from waveform displays, with doubtful segmentations verified by examination of spectrograms. The results are shown in table 1 for underlying short vowels, compensatorily lengthened vowels, and short vowels.

Table 1. Vowel duration measurements in LuGanda (in milliseconds).

	mean	s.d.		difference
Short vowel	73	27.1		
(n = 30)			}	118
Lengthened vowel	191	43.2		
(n = 22)			}	46
Long vowel	237	29.2		
(n = 25)				

In these new measurements, these three groups of vowels are all significantly different from each other at at least the .01 level (by Fisher's PLSD, a post-hoc test conducted after analysis of variance had shown a main effect of class: F (2, 74) = 183, $p <$ .0001). However, the compensatorily lengthened vowels are much closer to the duration of the underlying long vowels than to that of the short vowels. Both lengthened and long vowels are well over twice as long as the short vowels, whereas a lengthened vowel is only 40 ms shorter than a long vowel and has 80% of its duration.. Although there are clearly details to be accounted for in the surface durations, there is nothing here to fundamentally challenge the appropriateness of the phonological account of compensatory lengthening given above. (The long/lengthened vowel difference may well reflect durational adjustments made within the phonological word. The complex prenasalized consonant has a duration about 35 ms longer than a single nasal, and is considerably longer than a single /b/ which is usually lenited to [β] and somewhat longer than the other obstruent consonants measured.)

Sukuma has a similar process of compensatory lengthening before NC sequences (Batibo 1977). At first, it would appear that the same account given for Luganda would be appropriate for Sukuma, apart from additional rules needed to derive 'aspirated nasals' from underlying nasal + voiceless stop. However, the surface durational patterns are different. Measurements were made of sets of words, all verb infinitives, similar to those given in (8). These words were suggested by Herman Batibo, and read by him. I made measurements of the segment durations from waveforms, with assistance from Gary Holmes.

(8)

kʊkhaba	“to faint”
kʊkhama	“to wring out”
kʊkhaaba	“to exchange”
kʊkhamba	“to roof”

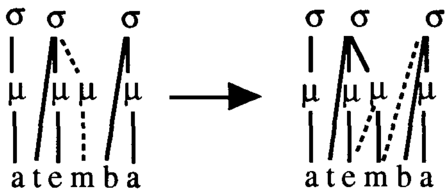
The results are shown in table 2. Compensatorily lengthened vowels are significantly shorter than lexical long vowels, as they are in Luganda, and the differences between all three means are again significant, each pairwise comparison at least reaching the .01 level (by Fisher's PLSD following analysis of variance which showed a significant main effect, F (2, 51) = 44.2, p < .0001). The difference is in the pattern formed by the durations relative to each other. The lengthened vowels fall almost halfway between the duration of the long and short vowels, in fact, the mean for lengthened vowels is slightly closer to the duration of short vowels. The mean duration of the compensatorily lengthened medial vowel in words such as *kʊkhaːmba* “to roof” is 200 ms, whereas that in short vowel words is 129 ms and that in long vowel words is 280 ms. Long vowels are over twice the length of short vowels in this data, but lengthened vowels are only about one and half times the length of short ones. The difference between long and lengthened vowels is double that found in Luganda. (All durations are somewhat longer in the Sukuma data, perhaps reflecting an overall speech rate difference. This is not likely to affect the pattern of differences within each language, but means that direct comparisons between the languages must be handled with care. In other respects—the type of data, the speech style, the status of the speakers, etc, the two data sets are very similar.)

Table 2. Vowel duration measurements in Sukuma.

	mean	s.d.		difference
Short vowel	129	58.2		
(n = 18)			}	71
Lengthened vowel	200	40.1		
(n = 18)			}	80
Long vowel	280	38.3		
(n = 16)				

Rather than the process represented in 7 above, these duration measurements make it seem as if only half the moraic value attributed to the nasal when it is in coda position has reassociated to the preceding vowel, and the nasal is only partially incorporated into the onset of the following syllable. The process can be represented as in (9), where the original syllabic and moraic affiliations of the nasal portion are changed by linking to the right and the left respectively as before but without delinking the mora originally created by the weight by position rule from the nasal. This mora is now split between vowel and nasal.

(9)



This produces a configuration unlike those previously entertained. Note that admitting such structures permits the representation of three distinct vowel lengths, but no more: a vowel linked to one mora, a vowel linked to two moras, and a vowel linked to, as it were, one and a half moras, as in the output of (9). The nasal is represented as a 'semi-geminate', functioning in two syllables but having only a shared mora. Moraic conservation is not violated, and onsets still contribute no weight. Furthermore, the nasal portion has a mean duration of about 120 ms, compared to about 75 ms for a single intervocalic nasal, as shown in table 3. The 'semi-geminate' nasal is thus almost 50 ms longer than the intervocalic nasal. In Luganda this difference is only about 20 ms. Given that the duration of the non-nasal portion of the prenasalized stops is much the same in the two languages, this supports the idea that any representational difference should concern the nasal portion. These further durational comparisons are also shown in table 3.

Table 3.

	Luganda		Sukuma	
Nasal in VCV	62		73	
		} 23		} 46
Nasal portion of NC	85		119	
		} 24		} 25
Total duration of NC	109		144	

I contend that the representation in (9) is appropriate for Sukuma, and that it is necessary to represent Sukuma differently from Luganda. I will consider two alternatives. One is that Sukuma has the same compensatory lengthening rule as Luganda, i.e. (7), and the differences are only matters of phonetic realization. The other is that Sukuma differs from Luganda in lacking a compensatory lengthening rule, but phonetic processes produce the intermediate length before a nasal + stop sequence. Let us take the second one first. If the vowel is assumed to be a phonologically short vowel before a nasal + stop cluster its intermediate length is not predictable from any of the familiar phonetic processes that affect phonetic vowel duration. Vowels are not in general lengthened before nasal + consonant. In fact, vowels are usually shortened when followed by a consonant cluster, e.g. in English (Klatt 1979). This is generally assumed to be because the first element of the cluster is within the same syllable as the vowel preceding it and "borrows" some of its duration in the process called "Closed Syllable Vowel Shortening" in Maddieson (1985). The same holds for geminates. Except if a phonological process such as compensation for a moraic reorganization occurs, a coda consonant can be expected to shorten a preceding vowel. If on the other hand we assume Sukuma has short vowels before prenasalized stops that are are single underlying complex consonants, there is again no general process which leads us to expect vowels to be lengthened in this environment. For example, no lengthening is found in Fijian (Maddieson 1990, Maddieson & Ladefoged forthcoming) where prenasalized stops are clearly single consonants from the outset. There are some smaller types of effects of the nature of the following consonant on preceding vowel duration, such as the voicing contrast in stops, but nothing of the magnitude and nature of the lengthening that would have to be assumed in this case.

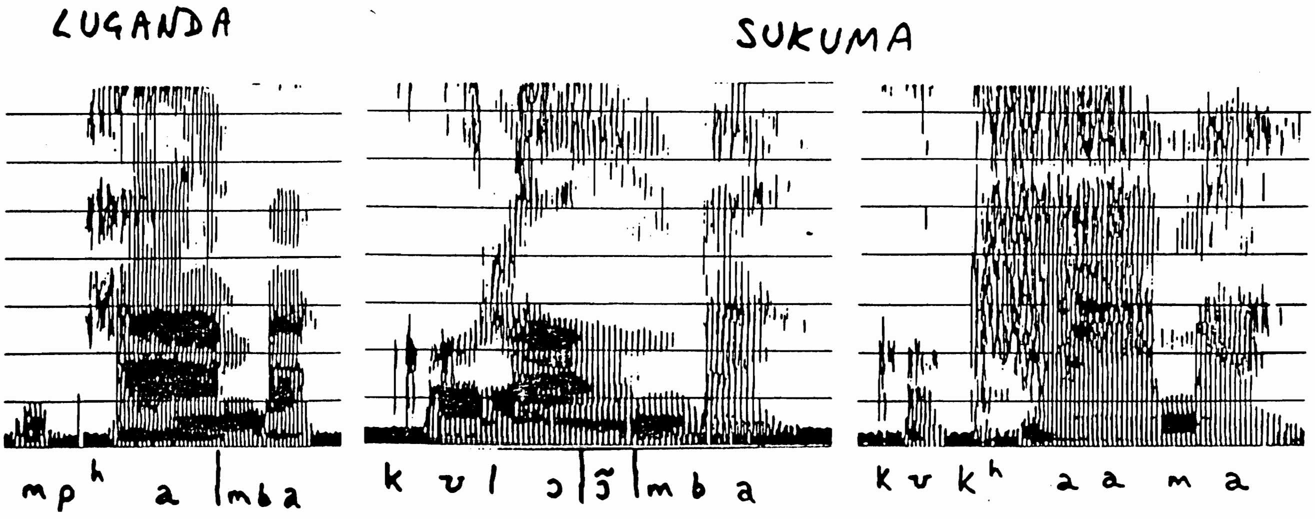
It might therefore be posited that these intermediate length vowels are actually underlying long vowels that have been shortened by the regular phonetic process affecting vowels before tautosyllabic consonants. That is, the contemporary process in Sukuma is wrongly labeled 'compensatory lengthening' (even though that is the correct historical account). Many words always have intermediate length, and this account might seem feasible if only these items were considered. However there is considerable evidence to show that we are dealing with an operation that lengthens short vowels, rather than one that shortens long vowels. Consider the alternations in a prefix such as the comitative *na*. This has a short vowel except before prenasalized stops, compare *namuunho* "with a person" and *naːmbazu* "with ribs".

It therefore seems fairly clear that we cannot dispense with a compensatory lengthening rule in Sukuma. The argument for providing distinct compensatory lengthening rules for Sukuma and Luganda, rests on much the same grounds of lack of phonetic generality of the phonetic realization rules that otherwise would be required to generate quite distinct lengths from the same output structure, namely a vowel linked to two moras. There is also another phonetic hint that the structures are different in the two languages. In the Luganda data to hand, the nasal element is generally not anticipated by nasalization of the lengthened vowel that precedes it. On the other hand, in Sukuma approximately half the vowel preceding the nasal element is nasalized. Compare the first two spectrograms in (10). In the Luganda word in (10a) there is no indication of nasal coupling during the vowel preceding /mb/, whereas in the Sukuma word in (10b) a dramatic change in the resonance pattern of the vowel is apparent about midway through its duration. The bandwidths of the formants broaden, resulting in much less clear definition of their frequencies in the spectrogram. Other work with the same Sukuma speaker confirms directly from nasal airflow measurements the anticipation of nasality in such positions (Maddieson 1991). Simple intervocalic nasals have a different pattern for this speaker, as shown in (10c). Here, no change in the formant pattern is observed during the long vowel /a:/ preceding /m/, indicating that this nasal is affiliated with the following syllable alone.

In conclusion, then, intermediate length is an aspect of the phonology, not the phonetic realization, of Sukuma, requiring a structural representation distinct from that of a long vowel and that of a short vowel, but intermediate between them. If the mora originally associated with the nasal is not delinked but is split between the nasal and the preceding vowel when the resyllabification occurs, such a structure is generated. This analysis predicts that some 'weight' remains associated with the prenasalized stop; in fact, Sukuma postvocalic prenasalized stops are substantially longer than simple nasals or stops in the same position, and longer than those in Luganda. Rather than lacking a compensatory lengthening rule, Sukuma shows an intermediate historical stage on the way to the Luganda situation.

* This is a very slightly expanded version of a paper presented at the LSA annual meeting held in Philadelphia, January 9-12 1992. Many thanks to Herman Batibo, Francis Katamba and Larry Hyman for the data that made this study possible. They are not responsible for the interpretation of the data. For presentational effect it is assumed that the data reflect valid samples of the Luganda and Sukuma speaking populations; however, appropriate caution should be exercised in generalizing from single speakers, and further work is needed to confirm the general applicability of the results.

(10) (a) (b) (c)



References

Batibo, Herman. 1976. A new approach to Sukuma tone. In L.M. Hyman (ed) *Studies in Bantu Tonology*. University of Southern California, Los Angeles: 241-257.

Batibo, Herman. 1977 [1985]. *Le Kesukuma, Langue Bantoue de Tanzanie: Phonologie, Morphologie*. Éditions Recherche sur les Civilisations, Paris.

Batibo, Herman. 1991. The two-directional tone melody spread in Sukuma. Paper presented in special session on African languages, Berkeley Linguistics Society.

Clements, G.N. 1986. Compensatory lengthening and consonant gemination in Luganda. In L.Wetzels & E.Sezer (eds) *Studies in Compensatory Lengthening*. Foris, Dordrecht: 37-78

Goldsmith, John. 1985. On tone in Sukuma. In D.L. Goyvaerts (ed) *African Linguistics: Essays in memory of M.W.K. Semikenke*. John Benjamins, Amsterdam: 167-187.

Hayes, Bruce. 1989. Compensatory lengthening in moraic phonology. *Linguistic Inquiry*. 20: 253-306.

Herbert, Robert K. 1975. Reanalyzing prenasalized stops. *Studies in African Linguistics* 6: 105-124.

Herbert, Robert K. 1986. *Language Universals, Markedness Theory and Natural Phonetic Processes.* Mouton de Gruyter, Berlin.

Hyman, Larry M. 1985. *A Theory of Phonological Weight.* Foris, Dordrecht.

Hyman, Larry M. 1992. Moraic mismatches in Bantu. Paper presented at LSA Annual Meeting, Philadelpia, January 1992.

Klatt, Dennis H. 1979. Synthesis by rule of segmental durations in English sentences. In B. Lindblom & S. Öhman (editors) *Frontiers of Speech Communication Research.* New York: Academic Press, 287-300.

Maddieson, Ian. 1985. Phonetic cues to syllabification. In V.A. Fromkin (ed) *Phonetic Linguistics*. Academic Press, New York.

Maddieson, Ian. 1991. Articulatory phonology and Sukuma 'aspirated nasals'. 17th Berkeley Linguistics Society Meeting, special African session, *Proceedings* (ed K. Hubbard): 145-154.

Maddieson, Ian & Peter Ladefoged. Forthcoming. The phonetics of partially nasal segments. In M.K.Huffman & R.A.Krakow (eds) *The Feature Nasal: Phonetic Bases and Phonological Implications*. Academic Press, Orlando.

McCarthy, John & Alan Prince. Forthcoming. Prosodic Morphology.

Appendix. List of Sukuma words.

Short V		Long V		Lengthened V	
oral C	nasal C	oral C	nasal C	prenasalized C	aspirated nasal
kʊkhaba *faint*	kʊkhama *wring out*	kʊkhaaba *exchange*	kʊkhaama *be bewildered*	kʊkhamba *to roof*	
kʊlaba *twinkle*	kʊlama *live long*	kʊlaaba *cultivate*		kʊlamba *to become expensive*	kʊlaɲha *to stay on on a tree*
kʊsaba *be rich in cows*		kʊsaaba *dig up*	kʊsaama *migrate*	kʊsamba *groan*	
kʊseba *boil up*		kʊseeba *dig out*		kʊsemba *take off*	kʊʃemha *to milk*
kʊʃiga *leave*		kʊʃiiga *remain*		kʊʃiŋga *stay*	giŋhi *owl*

The phonological representation of laterals[1]

Joyce McDonough

0. Introduction

This paper outlines an argument that laterals - lateral plosives, fricatives and approximants- are phonologically represented in feature geometry as a kind of root node, an aperture or A-position, as in (Steriade 1991, 1992). The central idea underlying this proposal is that lateral is best understood and represented if we return to the idea that lateral is a major class feature (Ladefoged 1972, 1993) distinguishing central from lateral airflow.

1. Lateral contours

In the standard representation of feature geometry (Clements 1985, Sagey 1986, McCarthy 1988), contour segments are single segments that display opposing values for a given feature: [+/-continuant] for affricates, [+/-nasal] for pre- or post- nasalized stops.

(1) Standard representation of contour segments in feature geometry:

Rc = consonantal root

-x **+x**

where x = [cont] affricates
[nasal] pre/post nasalized stops

A third kind of contour segment is the lateral affricate found in languages like Navajo (Athabaskan), Flathead (Montana Salish), Dagara (Niger-Congo). A lateral contour segment must contain in its representation a lateral and a non-lateral substring. We can then , given standard representation, expect lateral affricates to be represented in a feature geometry as in (2), as a contour segment with opposing values for lateral.

(2) Lateral affricate as a contour segment ' tl ' :

Rc

-lat **+lat**

There are a number of problems with this approach. [-Lateral] is a vacuous feature that describes no natural class of segments and has no relevance outside of this use. There are, for instance, no rules of the sort found in (3), that would nasalize all [-lat] segments after a nasal.

(3) Lateral as privative feature:

[1] This paper is a talk presented at the LSA in January 1993 in Los Angeles. I would like to thank Ken DeJong, Keith Johnson, Peter Ladefoged, Ian Maddieson, Donca Steriade, Richard Wright and the members of the UCLA Phonetics Lab. This work is supported by an NIH grant.

No rules of the sort:

* [-lat] ----> [+nas] / [+nas] _____

Another issue is the lack of agreement in the literature on the position and/or status of lateral in the feature hierarchy as either a place or a manner feature.

2. Lateral as a place or manner feature

Levin (1988) has argued that all laterals are coronals. She proposes a representation as in (4), with lateral as coronal dependent.

(4) Lateral as place feature (Levin 1988):

```
[lateral]
  COR
   |
 PLACE
```

In this view lateral is a place feature. One problem for this view, as she notes, is the existence of velar laterals discussed in Ladefoged, Cochran, Disner, 1983, and Ladefoged and Maddieson (1986).

(5) Non-coronal laterals (Ladefoged, Cochran, Disner, 1983, Ladefoged and Maddieson 1986):

(a) velar laterals [ʟ] (Melpa, Mid-Waghi, Kanite)
velar lateral affricates [k͡ʟ’] (Zulu)

(b) <u>Zulu</u>: velar lateral affricate

[k͡ʟ’ìná]
‘be naughty’

(c) <u>Mid-Waghi</u>: veelar lateral in contrast with coronal laterals

dental	alveolar	velar
al̪a al̪a	alala	aʟaʟe
‘again and again’	‘to speak improperly’	‘dizzy’

Addressing this, she argues that non-coronal laterals and lateral affricates would be represented as in (5), as doubly articulated segments containing a primary coronal constriction.

(6) Velar laterals (not true velars):

```
[lateral]
  COR        DORSAL

       PLACE
```

However evidence is, that for velar laterals, the primary contact is in the velar region and with lateral airflow:

3. Central versus lateral articulation

I propose that we take a different approach to laterals and return to the idea (Ladefoged 1964, Catford 1977) that lateral marks a major category distinction, between central vs lateral airflow. Laterals are articulated so that the oral airflow channels are formed along the sides of the tongue, central consonants have airflow channels along the central part of the tongue.

3. 1 Lateral contrasts

The consonant systems of the Athabaskan languages make use of the central/lateral contrast in their inventories. Navajo, an Athabaskan langauge, has five lateral consonants, three lateral affricates and a voiced and voiceless lateral fricative. All the laterals are coronal. The inventory is reproduced below from Young and Morgan (1986). The laterals are in boldface.

(9) Segment inventory: Navajo consonants (IPA):

(p)	t^{h}	t	t'				k^{h}	k	k'	ʔ
							kw^{h}	kw		
	tɬ^{h}	**tɬ**	**tɬ'**							
	ts^{h}	ts	ts'	tʃ^{h}	tʃ	tʃ'				
	s	z		ʃ	ʒ		x	ɣ		h
	ɬ	**l**								
(m)	n									
w				j						

There are two main groups of consonants in Navajo, obstruents with oral closure (plosives) and all others (non-plosives). Plosives exhibit a three way series: aspirated, unaspirated and glottalized. Fricative voicing is predictable and labials consonants are rare.

Navajo also has a root constraint blocking a root from bearing opposing values for the feature anterior as in (10a). Roots can be found as in (10b) column I and not as in column II:

(10) Navajo root constraint

(a) No roots of the kind: *[C .... C]

$$\begin{array}{cc} \multicolumn{2}{c}{*[\ C \ \ldots\ C\ ]} \\ \vdots & \vdots \\ \alpha\text{ant} & \beta\text{ant} \end{array}$$

(b)

	I		II
	tsį́į́s	‘pull, tug’	*tʃiis
	ʒóóʃ	‘lie parallel’	*zóóʃ

Laterals neither invoke nor block this constraint. They can appear with consonants having either value of anteriority as in (11).

(11) Laterals allowed with both values of anterior

tɬóóʃ	‘move on all fours’	łííʃ	‘leak’
tɬ‘is	‘harden’	lóós	‘lead’

A harmony system is in effect in Navajo that mimics the root constraint by extending the constraint into a feature changing rule (McDonough 1990).

Given this system of contrasts and constraints we need to account for three things in Navajo: 1) the representation of lateral affricates as contour segments, 2) the contrast between lateral plosives and non-plosives and 3) the contrast between lateral fricatives and the more common lateral approximants. As we have seen, current representation of lateral as a feature in the hierarchy below the level of root as either a manner feature off the root (or supralaryngeal node in Sagey (1986)), or as a coronal dependent place feature (Levin 1986) has serious problems.

3.2 Lateral as a root feature

McCarthy (1989) proposed, in a widely accepted tact, that the features [consonant] and [sonorant] occupy the root node, as major class features. We can identify the central /lateral opposition as a major class opposition and widen the scope of features associated to the root node to include this opposition. In the representation below the feature [lateral] is in the root node of lateral segments.

(12) laterals:
root = [cons]
[son]
[lat]

Those without lateral specification are central consonants:

(13) central consonants:
root = [cons]
[son]

It is not clear, however, how to incorporate lateral into the root node while accomodating contour segments and the various dependencies of place and stricture associated with laterals.

3.3 Aperture theory

One solution before us lies in Aperture Theory of Steriade (1992, 1993). Steriade takes up the idea that released plosives consist of a closure and a release phase. These phases are represented by aperture or A-positions. Stops are bipositional, with a closure position and a release position. Non-plosives or continuants have one A-position. The release position of plosives is equivalent to the closure of continuants.

(14) Steriade (1992)- *Aperture Theory*:

(a) Released plosives have two positions - closure/ release
Continuants have one - closure

(b) Release of plosives = closure of continuants

In this view then there are three kinds of apertures represented as A-positions in (14): a closure and two kinds of release, or continuant closure; fricative and approximants:

(15)	A_0	A_f	A_{max}
	closure of stops	fricative stricture	approximant stricture

Combining these will result in four kinds of segments: fricatives, approximants, stops with a closure and an approximant release, and stops with a fricated release, or affricates.

(16)		
	stops:	A_0 A_{max}
	affricates:	A_0 A_f
	fricatives:	A_f
	approximants:	A_{max}

This is in accord with general consensus concerning consonantal types.

3.4 Lateral as an A-position

If lateral is a major class category then all segments fall into one of two categories: ones with central airflow and ones with lateral airflow. The three kinds of consonantal strictures agreed apon by phonologists and phoneticans are stops, fricatives and approximants. They are by default central consonants. Thus all of the types of segments in (16) are central. If we take the central/lateral distinction to be operative in a grammar, lateral as a category stands in opposition to these three types.

We can represent this opposition as in (17), by adding to the inventory of A-positions a lateral A-position:

(17) Lateral A-position

laterals: A_l

This A-position has the same status as the other A-positions; as a kind of root node with content, a position that serves as a docking site which mediates between features and the prosodic structure.

How do we then resolve the central/lateral opposition, stipulated as an A-position contrast, with the existence of lateral affricates, fricatives and approximants?

The existence of a lateral A-position neatly explains the lack of a contrast between lateral stops and affricates. All plosives are bipositional. All lateral plosives are affricates. They can be represented as in (18): as a segment with a closure and lateral release.

(18) Lateral plosives contain a stop position:

lateral affricates: A_0 A_l

The contour nature of lateral plosives is accounted for in the same way affricates are accounted for in Steriade's Aperture Theory: they are contour segments because they are bipositional. In the case of lateral affricates, their release is lateral.

By adding to the inventory of A-positions we predict the existence of three non-plosive consonant types:

(19) Non-plosives:

fricatives	approx.	laterals
A_f	A_{max}	A_l

The status of laterals in this view is equivilant to, and in opposition to, the A-positions for fricatives and approximants.

Note however, as the distinction between fricatives and approximants is represented by A-positions, if lateral is also an A-position, we are left with accounting for the well-documented distinction between lateral fricatives and lateral approximants.

3.5 Contrasts among lateral non-plosives

There are two aspects to this problem: the existence of lateral fricatives and lateral approximants and reported voicing contrasts among the lateral fricatives and approximants.

It is a well-know fact that both lateral fricatives and lateral approximants exist. Maddieson (1984) reports that of the laterals found in languages in the UPSID database, 79.7% are approximants and 10.8% are fricatives. In addition there are languages in which both type exist (Zulu, Hupa, Nez Perce, Kwakw'ala, Quileute), thus the contrast is phonologically significant.

I propose that the distinctive property of the lateral fricative/approximant contrast is voicing. Voicing governs the presence or absence of fricative noise in laterals. The basis of this proposal lies in the distribution of voicing among lateral fricatives and approximants. Pike (1943) and Catford (1977) have pointed out that, phonemically, voiceless laterals are fricatives, never approximants. In a study of voiceless laterals Maddieson and Emmory (1984) state that while voiceless lateral approximants vs fricatives are phonetically distinct, that is, four way contrasts exist between voiced and voiceless fricatives and voiced and voiceless approximants, the difference is not exploited as a contrast in any language. This distribution fact is explained if voicing were the distinctive property between lateral fricatives and approximants. Voiceless laterals are fricatives by default.

Evidence for this approach is in rules such as the ones found in Melpa in (20), voiced lateral approximants are devoiced, and, crucially, fricated, in final position:

(20) Devoicing and frication of final laterals in Melpa (Ladefoged and Maddieson 1986)

dental		alveolar		velar	
kial̪t̪ɨm	'fingernail'	lola	'speak improperly'	paʟa	'fence'
wał̪	'knitted bag'	bał	'apron'	rał	'two'

This lateral voicing hypothesis also predicts that cases where voicing is the distinctive property of a contrast between lateral segments, as in Navajo where fricative voicing is predictable and the unmarked laterals are fricatives, that the voiced version will be an approximant. Ladefoged and Maddieson (1986) have reported the voiced lateral non-plosive of Navajo is an approximant. Thus in Navajo, while laterals participate in the phonology of voicing that applies to all fricatives, how they do it is different. All fricatives exhibit voiced/voiceless pairs in the phonology, but for the lateral as opposed to the central fricative, its voiced reflex is an approximant.

If voicing is the distinctive feature of this contrast, we can represent lateral approximants as in (20). Lateral A-positon marked for voicing.

(21) Lateral approximant:

Root	Al
	|
Laryngeal	°
	voi

This phenomena of the sensitivity of stricture to voicing is also known to us in the form of fricative/glide alternations invoked by intervocalic voicing[3].

3.6 'Voicing' in fricated laterals

The voicing contrast found among the lateral fricatives and approximants is at first glance harder to capture. As noted, there are no occuring contrasts found between voiceless fricatives and approximants. However, languages such as Flathead (Montana Salish) and Zulu have a three way contrast between voiced and voiceless lateral fricatives and a voiced lateral approximant. The following data from Ladefoged, Cochran and Disner (1977) exhibits this contrast.

(22)

Zulu:		
voiced approximant	lálà	'sleep'
voiced fricative	ɮálà	'play'
voiceless fricative	ɬânzà	'vomit'

If voicing is the contrastive property distinguishing lateral fricatives from approximants, how do we represent the apparent voicing contrast between the two fricative types?

One approach is to view this data in (22) set as a three way contrast, reminiscent of the three way stop contrasts in Thai and Eastern Armeneian (Lisker and Abramson 1964). The contrast for Thai is described as voiced, voiceless unaspirated and voiceless aspirated, as in the data below (from the SOWL database):

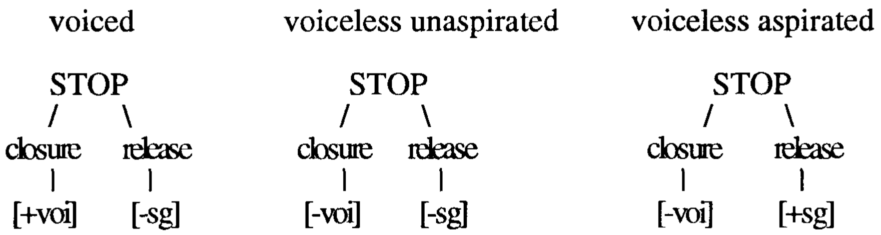
(23)

Thai:		
voiced	**bâ:**	'crazy'
voiceless unaspirated	**pâ:**	'aunt'
voiceless aspirated	p^{h}â:	'cloth'

[3] McDonough (1993) argues that there is another aperture position in Navajo, an unmarked or underspecified position 'A'. Like laterals, this segment interacts in the voicing phonology according to its type: it surfaces as a glide when voiced. This is the locus of the glide/ fricative alternations in Navajo.

Keating (1990) proposes a categorial representation of three way stop contrasts such as this using the features [voice] and [spread glottis], realized on the closure and release positions of stops respectively:

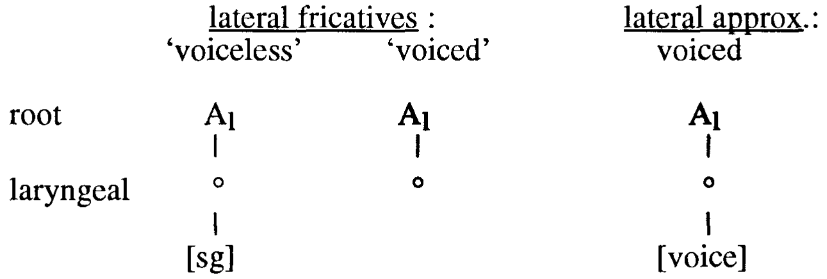
(24) Three way stop contrast (Keating 1990):



As Abramson and Lisker (1964) point out, the acoustic cues for the perception of the stop contrast in Thai are voice onset time. Keating calls this a voicing continuum, [1voice], [2voice], and [3voice], represented as above. Maddieson and Emmory (1984) report a salient difference between the voiceless lateral fricative and the voiceless lateral approximant is VOT; the voiceless approximant had a consistantly shorter VOT than the voiceless fricative. Based on this it seems reasonable to propose that a similar situation as the stop contrast in Thai entails for the contrast among the laterals we see in (22). We predict the 'voiced' lateral fricative to pattern as the voiceless unaspirated stop does in the contrast in (24), medially between the lateral approximant and the 'voiceless' lateral fricative.

Keating's representation in (24) crucially involves stops. The feature voice, representing vocal chord vibration, is realized on the closure, aspiration ([sg]) on the release. Keating (1984, 1990) notes the lack of a similar tri-partite pattern of voicing in continuants and correlates this with the structural differences between continuants and stops (one postion versus two positions) in Steriade's system. Continuants don't have a release, thus don't have a distinct docking site for [sg], and therefore can't show the three way contrast. However, this distribution property is not intrinsic to the representation. As Keating notes, the connection of [voice] or [sg] to a position is not a necessary one. The bipositionality of stops simply explains the greater richness of stop contrasts, as Steriade has pointed out. If we assume the features [voice] and [sg] to be privative, the three way contrast in (22) can be captured as in (25).

(25)



The phonological features distinguishing the three way lateral contrast in (22) are [spread glottis] and [voice]. The lateral approximant is [voice], the 'voiceless' lateral fricative is [spread glottis], the 'voiced' fricative is unaspirated, and unmarked.

It remains to justify outside a structural explanation why continuants other than laterals do not show three way contrasts.

3.7 Navajo consonantal contrasts

In (26) and (27) are represented the consonantal contrasts of Navajo, as a set of A-position distinctions. The stops and affricates are as in Steriade (1992, 1993) in (15), the laterals affricates are represented bipositionally in (23iii) as a closure with a lateral release.

(26) Navajo plosives :

i. stops:

$A_0 \quad A_{max}$

ii. affricates:

$A_0 \quad A_f$

iii. lateral affricates = laterally released stops, ('lateral plosives' Sapir 1932):

$A_0 \quad A_l$

There are three types of continuant in Navajo that correspond to the release positions of stops represented below in (26): fricatives, approximants and laterals. The two kinds of lateral continuants are distinguished by the feature voice in (24iii).

(27) Non-plosives:

i. fricatives:

A_f

ii. approximants:

A_{max}

iii. lateral fricatives:

$$\begin{array}{ccc} [\text{ɬ}] & [\text{l}] & \\ A_l & A_l & \text{root} \\ | & | & \\ \circ & \circ & \text{laryngeal} \\ | & | & \\ [\text{sg}] & [\text{voi}] & \end{array}$$

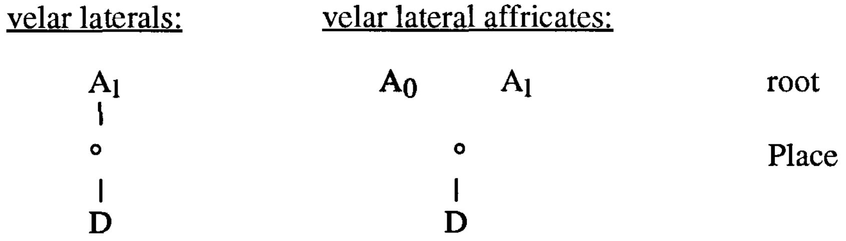
The system of consonantal contrasts in Navajo is then in part a system of symmetrical aperture contrasts: there are three continuant A-positions (A_f, A_{max}, A_l) and these combine with closure positions resulting in three sets of plosive contrasts as in (26). The richness of the lateral inventory in Athabaskan is a result of this symmetry.

3.8 Non-coronal laterals

While non-coronal laterals are extremely rare (Maddieson 1985:77), they do occur, thus the non-coronality of laterals cannot be explained away by phonological impetus.

Both velar laterals and velar lateral affricates are easily accomodated within this extended system of A-positions. In (28) are the representations of the Zulu non-coronal laterals of (5) in this view with Dorsal (D) marking velar closure.

(28) Zulu:



The propensity of laterals for coronal place of articulation, explained in Levin's (1986) proposal by intrinsic phonological coronality, is left to be explained outside of the feature heirarchy. The work of Stone (1990) and others suggests that this fact lies in the intrinsic physiology of the tongue.

4. Conclusion

Given the existence of a lateral A-position 'A_l', we can account for the representation of lateral affricates as contour segments, and the mootness of the lateral stop vs affricate distinction, as a result of the bipositionality of plosives. Bipositionality allows a stop to contain a lateral and a non-lateral substring. As lateral is an aperture postion, the only possible lateral plosive in this view is a stop with a lateral release. We can also account for the distinction between lateral plosives and non-plosives. Aperture Theory states that the release of stops is equivilant to the closure of continuants. The existence of a lateral A-position predicts the existence of a lateral continuant symmetric to the pattern of the fricative A-position; i.e. laterals and lateral releases on stops symmetric to fricatives and fricative releases on stops.

The final problem presented to us by the Navajo inventory is that of distinguishing between lateral approximants and lateral fricatives. Since the distinction between fricatives and approximants is represented as an A-position contrast and since laterals are also represented as an A-position, the distinctions between fricatives and approximants and between lateral fricatives and lateral approximants are predicted to be discrete. The proposal offered in this paper is that the lateral fricative/approximant distinction is one of voicing. This proposal is backed by evidence that lateral fricatives alternate with lateral approximants in voicing rules. In laterals then stricture and voicing are interdependent. This interdependence is familiar to us in phenomena such as fricative/glide alternations in intervocalic voicing.

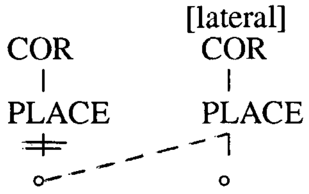
This proposal has lead us to consider the three and four way contrasts among lateral continunats. While no contrast exists between voiceless lateral approximants and voiceless lateral fricatives, three way contrasts can be found between lateral approximants and 'voiced' and 'voiceless' lateral fricatives. The claim of this paper is that this three way contrast is congruent to three way contrasts found on stops in languages like Thai, as a voicing continuum. After Keating (1990) this contrast can be represented using the features [voice] and [sg].

One prediction that this proposal makes is that there will be no spread of lateral independent of place and stricture. Clements, Levin (1986) and others have analyzed examples of lateral alternations as assimilation processes involving place and/or manner features. What is true of these cases is that there are no clear and straightforward examples of lateral assimilation such as we are familiar with from nasal place assimilation. The Klamath case, for instance, discussed in Levin (1986), which involves the very common process of place assimilation in nasals (n + p --> mp), also shows an alternation of the following type:

(29) n + l ---> ll

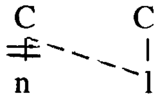
Levin analyzes this as a further case of nasal assimilation to place and evidence for her proposal that lateral is coronal dependent. We might assume a representation of this process as follows:

(30)



However it is unlike the other Klamath nasal alternations in that it also effects the feature [nasal]. The alternation in (29) involves the deletion of a nasal before a lateral and the spread of lateral. As lateral does not spread independent of any of its specifications, including apparently its default specification [-nas] here, this alternation can also be analyzed as the deletion of a nasal segment before a lateral, and resulting the spread of the lateral segment onto the empty slot.

(31)



The Selayarese examples she analyzes suffer the same problem. At best, the value of including alternations such as the one in (26) under the much more widespread cases of nasal place assimilation is open to discussion.

Bibliography

Bladon, R. A. W. (1979). "The production of laterals: Some articulatory properties and their acoustic implications." *Current Issues in the Phonetic Sciences.* Amsterdam, John Benjamins. 501-508.

Bladon, R. A. W. and A. Al-Bamerni (1976). "Coarticulation resistance of English /l/." *Journal of Phonetics.* **4**: 135-150.

Clements, G. N. (1989). A Unified Set of Features for Consonants and Vowels. In

Clements, G. N. (1985). The geometry of phonological features. In C. J. Ewen & J. M. Anderson (Eds.), *Phonology Yearbook 2* (pp. 225-252). Cambridge University Press.

Chomsky, N. and M. Halle (1968). *Sound Pattern of English.* New York, Harper and Row.

Davey, A., L. Moshi, et al. (1982). “Liquids in Chaga.” *UCLA Working Papers in Phonetics*, **54**: 93-108.

Jun, J. (1992). "The position of [lateral] in feature geometry". ms. UCLA.

Keating, P. (1990). “Phonetic representation in a generative grammar”. *Journal of Phonetics 18*, 321-334.

Keating, P. A. (1984). Phonetic and phonological representation of stop consonant voicing. *Language, 60*(2), 286-319.

Keating, P.,Linker, W., & Huffman, M. (1983). Patterns in allophone distribution for voiced and voiceless stops. *J. Phonet., 11,* 277-290.

Kenstowicz, M. (1992). *Phonology in Generative Gramma*r. ms. MIT

Ladefoged, P. (1971). *Preliminaries to Linguistic Phonetics* ((Midway reprint 1981) ed.). Chicago: University of Chicago Press.

Ladefoged, P., A. Cochran, and S. Disner. (1977). “Laterals and trills.” *JIPA* **7** (2): 46-54.

Ladefoged, P. and I. Maddieson (1986). “The Sounds of the World's Languages.” *UCLA Working Papers in Phonetics*, **64**.

Levin, J. (1988). “A place for lateral in feature geometry.” Ms., University of Texas, Austin.

Lisker, L., & Abramson, A. S. (1964). “A cross language study of voicing in initial stops: Acoustical measurements.” *Word, 20,* 384-422.

Lisker, L., & Abramson, A. S. (1971). Distinctive features and laryngeal control. *Language, 47,* 767-785.

Maddieson, I. (1984). *Patterns of Sounds.* Cambridge, Cambridge University Press.

Maddieson, I. and K. Emmorey (1984). “Is there a valid distinction between voiceless lateral approximants and fricatives?” *Phonetics* **41**: 181-190.

Maddieson, I. and P. Ladefoged (1989). “Multiply-articulated segments and the feature hierarchy.” *UCLA Working Papers in Phonetics*, **72**: 116-138.

McCarthy, J. (1989). “Feature geometry and dependency: a review”. *Phonetica* **43.**

McDonough, J. (1990). "Consonant harmony in Navajo."*The Proceedings of WCCFL X.* Stanford University Press.

McDonough, J. (1993). *The Phonetics and Phonology of Navajo*. Ms, UCLA.

McDonough, J., P. Ladefoged, H. George (1993). "Navajo vowels and phonetic universal tendencies." *Fieldwork Studies of Targeted Languages, UCLA Working Papers in Phonetics* **83.** (Also in *JASA* 92 (4, Pt 2):2416, 4aSP15.)

McDonough, J., and P. Ladefoged (1993)."Navajo stops."*Fieldwork Studies of Targeted Languages,UCLA Working Papers in Phonetics* **83.**

Rice, K and P. Avery. (1991). “On the relationship between laterallity and coronality”, *The Special Status of Coronals, Phonetics and Phonology* V 2, Academic Press.

Sagey, E. (1986) *The Representation of Features and Relations in Nonlinear Phonology*. Ph.D dissertation, MIT.

Selkirk, E. (1990). “Two root theory of geminates”, *University of Massachusetts Occasional Papers*, **14** (Dunlap and Padgett, eds.).

Shaw, P. (1990). "Consonant harmony systems" in *Phonetics and Phonology 2* (Paradis and Prunet, eds.) New York: Academic Press.

Steriade, D. (1992). “Complex onsets as single segments: the Mazateco pattern.” Ms, UCLA.

Steriade, D. (1993). “Closure, release and nasal contours.”*Nasality: Phonological and Phonetic Properties* (Huffman and Krakow, eds.) New York: Academic Press.

Steriade, D. (1986). "A note on coronal." Ms, MIT.

Stone, M. (1991). Toward a model of three-dimensional tongue movement. *Journal of Phonetics, 19*, 309-320.

Acoustic and auditory analyses of Xhosa clicks and pulmonics

Keith Johnson
University of Alabama, Birmingham

Introduction

Usually, the phonological natural classes predicted by acoustic properties are the same as those predicted by articulatory properties.[1] For example, labial sounds may be characterized either by the low frequency emphasis in their acoustic spectra or by the presence of a constriction at the lips. Similarly, the presence of turbulent noise in the acoustic wave form of a sound defines roughly the same class of sounds as those which have narrow articulatory constrictions (fricatives). In these and most other cases, the predicted natural classes are the same whether one chooses to focus on sounds or articulations. Thus, for most features phonetic implementation may be defined in either acoustic or articulatory terms.

Clicks are interesting because, although the natural classes among them predicted by acoustic and articulatory properties are not different, the acoustic properties of clicks predict different cross-classifications of click and nonclick sounds than do their articulatory properties. Therefore, we expect to find that either the acoustic or the articulatory properties of clicks correctly predict the cross-classifications with pulmonics which are needed for the statement of linguistic phenomena.

This paper reports a quantitative analysis of acoustic and auditory properties of clicks and pulmonics in Xhosa. The study focused on the cross-classification of clicks and nonclicks and extended earlier work on the acoustics of clicks (Traill, 1992a, 1992b, Sands, 1991, Ladefoged & Traill, to appear) in two ways. First, the paper reports the use of two quantitative procedures (cluster analysis and moments analysis) to discover the similarities among sounds. Theoretical and perceptual studies (Stevens & Blumstein, 1978; Traill, 1992b) have suggested certain acoustic characterizations, but the most commonly used methods for discovering acoustic properties are still quite subjective. Cluster analysis provides a rigorous basis for statements of acoustic/auditory similarity, and moments analysis provides an objective basis for the description of spectral properties. Second, the study tested the hypothesis that auditory spectra provide a better basis for the comparison of speech sounds than do acoustic spectra. This hypothesis is based on the assumption that if we can simulate the properties of the human peripheral auditory system, we can get a representation that is closer to the listener's experience of speech than is the acoustic spectrum.

Classification of clicks

(1a) shows an articulatory and acoustic classification of coronal clicks based on extensive phonetic data. Note that the feature [lateral] is redundant for the lateral click and the somewhat antiquated feature [delayed release] is used to distinguish the affricated and unaffricated clicks. An alternative representation in terms of aperture sequences (Steriade, 1992) is possible if we assume, following McDonough (1993), that there is a lateral aperture, and that frication on lateral aperture is redundant for voiceless lateral sounds. The revised classifications are shown in (1b).

[1] I am assuming that phonetic similarity among sounds "predicts" phonological natural classes. Since the sounds in a natural class typically share some phonetic property, it seems reasonable to hypothesize that if we find that a phonetic property is shared by certain sounds these sounds may function as a phonological natural class.

(1a) Articulatory and acoustic classifications of clicks. Adapted from descriptions given by Traill & Ladefoged (1984, to appear), Traill (1985), and Sands (1991).

Articulatory properties	| Dental	|| Lateral	! Alveolar	ǂ Palatal	Acoustic properties
laminal	+	-	-	+	acute
delayed release	+	+	-	-	noisy
lateral		+			

(1b) Alternative representation with aperture sequences.

Articulatory properties	| Dental	|| Lateral	! Alveolar	ǂ Palatal	Acoustic properties
laminal	+	-	-	+	acute
release aperture	A_f	A_{lat}	A_{max}	A_{max}	noisy release

From the articulatory properties in (1) we can predict some click/nonclick cross-classifications. For example, dental and palatal clicks should pattern with laminal pulmonic consonants such as dentals as opposed to apical consonants such as alveolars. Of course, since clicks are coronal they should pattern with other coronal consonants. However, it is important to note that clicks have a velar or uvular closure in addition to, and simultaneous with the anterior closures shown in (1), so clicks are specified for two articulators ([coronal] and [dorsal]) and may participate in processes involving either. The acoustic properties shown in (1) predict the same natural classes among clicks as the articulatory properties. However, cross-classification of clicks and non-clicks predicted by the acoustic properties are different from those predicted from articulation. For example, the lateral and alveolar clicks are acoustically [+grave] and thus should pattern with other [+grave] sounds.

Acoustic versus articulatory features

Traill (1992a) argues that natural classes of clicks and pulmonic consonants predicted by acoustic properties correspond to the classes at work in synchronic phonological processes and diachronic sound changes, while classes defined by articulatory properties do not. He discusses two sets of data.

In the first, a synchronic process in !Xóõ, /a/ becomes [i] when preceded by [|] or [ǂ] and followed by [i], as the examples in (2) illustrate (note that there is a partial raising and centralizing of /a/ to [ɐ] after [ʘ], [!] and [||]). Traill (1992a) also notes that 'there are a handful of cases' that show a change from /a/ to [i] "following the non-click consonants /t th s/". An earlier account of the process, outlined in Traill (1985), indicated that the class of sounds which conditions /a/-raising is "dental" and also includes /l/ and /n/.

(2) /a/-raising in !Xóõ (from Traill, 1992a).

/|ā̤i/ → [|ī̤i]	'aardwolf' (Proteles cristatus)
/ǂái/ → [ǂíi]	'steenbok' (Raphicerus campestris)
/ʘá'i/ → [ʘɐ́'i]	'abomasum' (part of ruminant's digestive system)
/!ái/ → [!ɐ́i]	'sp. of tree' (Zisyphus mucronata Willd.)
/||à̤i/ → [||ɐ̤̀i]	'old (Class.1)

These data illustrate that [|] and [ǂ] pattern together in !Xóõ and must be grouped in a natural class contrasting with [⊙], [!] and [||], a pattern which can be predicted by either the articulatory or the acoustic properties of clicks ([± laminal] or [± grave]). Traill (1992a) argued that /a/-raising in !Xóõ is evidence that clicks and pulmonics should be cross-classified according to their acoustic properties rather than their articulatory properties because [|] and [ǂ] like dental pulmonics and /i/ are [-grave]. However, the descriptions given in Traill (1985) make it clear that the dental sounds in !Xóõ, like dentals in other languages, are laminal. Therefore, the articulatory description given in (1) makes the same predictions for /a/-raising in !Xóõ as the acoustic description, and Traill's (1992a) argument does not hold.

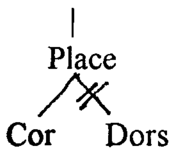
The second set of data discussed by Traill (1992a) provides stronger support for his hypothesis. These data (3) illustrate a diachronic process in Khoe dialects in which [!]-series clicks are replaced by pulmonic velars and [ǂ]-series clicks are replaced by pulmonic palatals.

(3) Diachronic click replacement in Khoe dialects (from Traill, 1992; Traill, 1986).

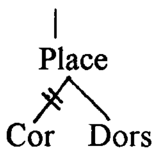
!Gwi	Ts'ixa	gloss
[!are]	[kare]	'cut into strips'
[!ŋaro]	[ŋgaro]	'chameleon'
[!ganee]	[ganni]	'chin'
[!hae]	[khae]	'pierce'
[ǂii]	[cii]	'call'
[ǂŋu]	[ŋɟuu]	'black'
[ǂgoa]	[ɟua]	'ash'
[ǂhuni]	[chuni]	'elbow'

Informally the sound change in Ts'ixa involves a change of nonaffricated clicks to pulmonics, and since clicks are specified for both [dorsal] and [coronal] we may formally express the change as a delinking process in which one of the place nodes in the click delinks. However, this formalization requires two types of delinking as shown in (4). The dorsal node delinks in the palatal click [ǂ] resulting in a non-click palatal, while the coronal node delinks in the alveolar click [!] resulting in a non-click velar. Without some inelegant stipulations this account cannot explain why coronal delinking applies to the alveolar click while dorsal delinking applies to the palatal click. Since the data in (3) indicate that at some point in the history of Ts'ixa there were rules like (4) in the synchronic grammar, these data suggest that phonological cross-classifications between clicks and pulmonic consonants can not be predicted from the articulatory properties of the sounds. However, since both the alveolar click [!] and the velar stop [k] are classified acoustically as [+grave], click replacement in Ts'ixa is consistent with Traill's hypothesis that the phonological cross-classification of click and nonclick consonants is based on acoustics, not articulation.

(4) ǂ > c



! > k



A rough sketch of an analysis based on the acoustic properties of clicks and pulmonics is shown in (5). The distinction between air stream mechanisms is represented as a difference in the root node and the process is simply that clicks become pulmonics. In this account, cross-classifications between clicks and pulmonics are based on what Jakobson, Fant & Halle (1963, henceforth JFH) called resonance features.

(5)	ǂ > c			! > k		
	click	→	pulmonic	click	→	pulmonic
	|		|	|		|
	acute		acute	grave		grave

What is missing from this argument is convincing evidence that click and nonclick consonants in a click language pattern together acoustically. It has been noted previously (Traill, 1992a, 1992b; Traill & Ladefoged, to appear; Sands, 1991) that the spectrum of the alveolar click [!] has a concentration of energy in lower frequencies (compared to the spectra of other clicks), just as the spectrum of [k] has more low frequency energy than is typically found in the spectrum of [t]. These observations suggest that [k] and [!] share a phonetic property, but this important prediction has not been tested in any previous study of the acoustics of clicks. The study reported here was designed to fill this gap.

Methods

The speaker was Ncediwe Mdunyelwa. She is a native speaker of Xhosa from Capetown, South Africa who was a visiting scholar at UCLA during the 1991-1992 academic year. The recording was made in February, 1992 at the UCLA phonetics laboratory by Sujin Yi as an illustration of a project for an introductory phonetics course. A word list illustrating most of the distinctive sounds of Xhosa was recorded twice. Both recordings of a subset of this list (shown in 6) was analyzed in the present study.[2]

(6) List of Xhosa words used in the study.

sound	orthography	transcription	gloss
p^h	phala	[p^hála]	'go fast'
t^h	thala	[t^hála]	'ledge of the rock'
k^h	khala	[k^hála]	'cry out'
ɸ	fukama	[ɸuq'áma]	'lie on'
s	salisa	[salísa]	'make remain'
x	ruzula	[xuzúla]	'pull away'
ǁ	xela	[kǁɛ́la]	'tell, say'
ǀ	cuba	[kǀúba]	'tobacco'
!	qaba	[k!ába]	'paint'

[2] These words present a stiffer than usual challenge for a classification scheme because of varying coarticulation from the different contextual vowels.

The recorded utterances were digitized (20kHz, 12 bits) and in each word a 25.6 ms (512 samples) section of the acoustic wave form was identified for further analysis. In stops and clicks the section of wave form was centered around the release burst (as in Stevens & Blumstein, 1978). In fricatives the window was centered around the peak amplitude of the frication. FFT and LPC spectra were calculated from these wave form segments (the LPC had a filter order of 22).

Following Stevens & Blumstein (1978), the LPC spectra were taken to represent the acoustic spectrum. The auditory spectra were constructed from the FFTs using an auditory model. Three important characteristics of the human auditory system were modeled. First, the (100) bandpass filters used in the model were equally spaced along a nonlinear auditory frequency scale (the Bark scale, Zwicker, 1961; Schroeder, Atal & Hall, 1979). Second, the bandwidths of the filters increased as the center frequencies increased (Patterson, 1976). Because of these aspects of the filter bank, lower frequencies were emphasized relative to higher frequencies, and small differences at high frequencies were wiped out because the high frequency filters had large bandwidths. The third aspect of the human auditory system captured by the model is the relative sensitivity of hearing at different frequencies. This was modeled by applying an equal loudness contour (Fletcher & Munson, 1933) to the filter outputs (see JFH, p.27). This increased the amplitudes of frequency components between 1000 and about 3000 Hz, relative to other parts of the spectrum.

After the acoustic and auditory spectra had been calculated, the DC offset in each spectrum was removed by subtracting the average amplitude of the spectrum from each point. (The LPC spectra were down-sampled from 512 points to 100 points, so that the acoustic and auditory cluster analyses were based on the same number of points.) With the DC offset removed the focus of comparison was on the shape of the spectrum ignoring overall amplitude differences. Then the average spectrum (of the two productions) was calculated for each of the 9 consonants. Ward's method of cluster analysis (Everitt, 1980; SAS, 1982) was used to group the spectra into hierarchical similarity spaces. Spectral moments (Forrest, et al., 1988) of the average acoustic and auditory magnitude spectra were calculated without removing the DC offset.

Results

The acoustic spectra for stops (top left panel of Figure 1) agree with the theoretical predictions outlined by Stevens & Blumstein (1978). The spectrum of [t] (drawn with the thick solid line) has more high frequency energy than those of [p] or [k], and the spectrum of [k] (drawn with a thin solid line) shows a concentration of energy between 1 and 2 kHz. The same basic aspects can be seen in the acoustic spectra of the fricatives (middle left panel in Figure 1). However, [x] (thin solid line) has a peak near 4 kHz which is unexpected. In the acoustic spectra of the click bursts (bottom left panel of Figure 1), the alveolar click [!] shows a low frequency peak similar to the peak seen in [k]. The dental [|] and lateral [||] clicks have similar looking spectra, and although the dental click has more high frequency energy than the lateral click, it does not have the same high frequency emphasis seen in [t] and [s].

One problem with these spectra is knowing what to focus on. For instance, the acoustic spectrum of [p] shows a small peak at about 1 kHz which may or may not be important for the identification of [p]. The lowest peak in the spectrum of [ф] is at 1.8 kHz which may or may not correspond to the lowest peak of [p]. The alveolar click [!] has a low frequency peak similar to that of [k], but [x] and [!] have prominent peaks at about 4 kHz. Is the similarity of [!], [k] and

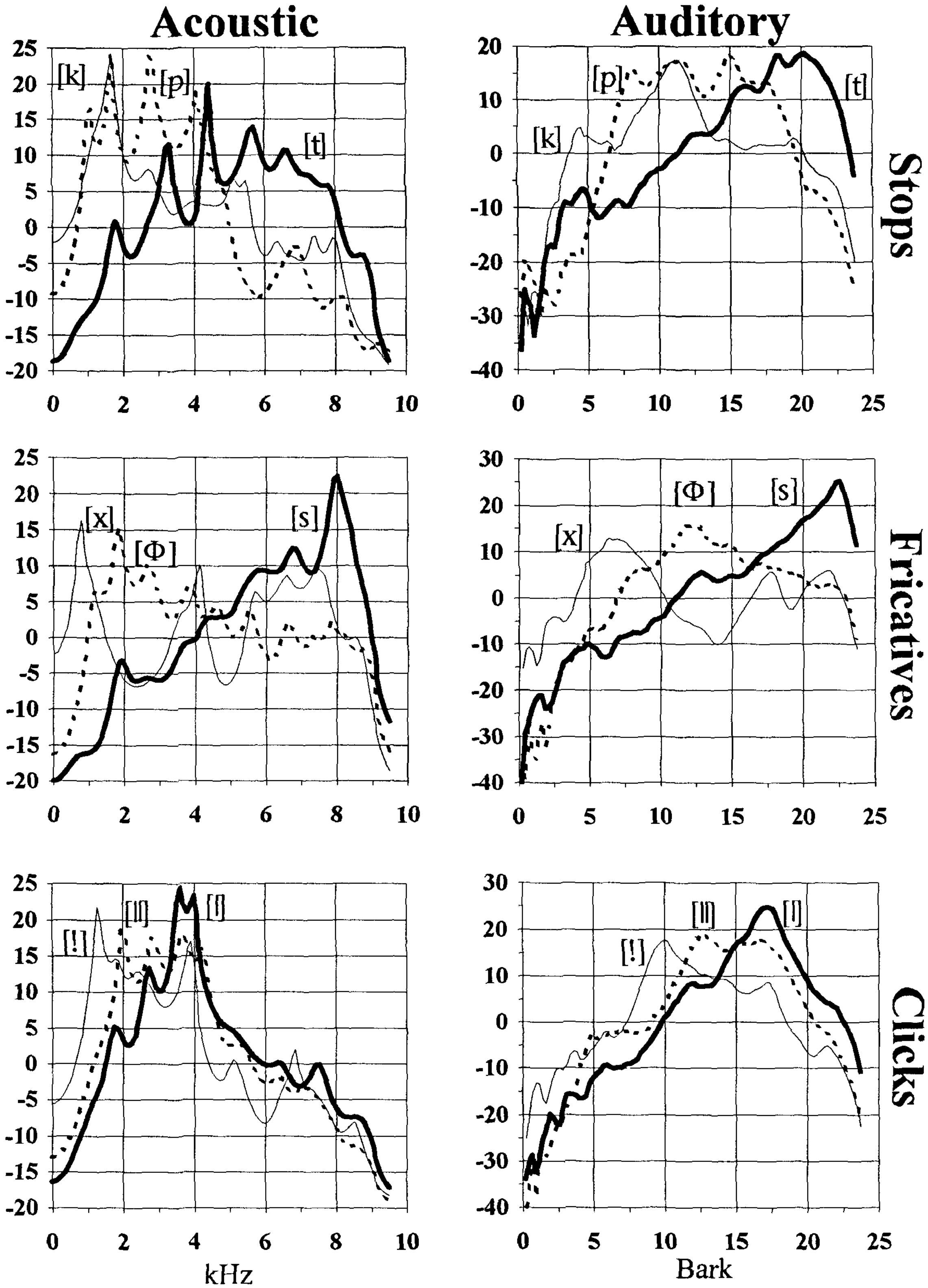


Figure 1. Acoustic and Auditory spectra of Xhosa stops, fricatives and clicks.

[x] below 2 kHz enough to offset their differences in the higher frequencies? Without *a priori* assumptions it is not clear how to assess the differences and similarities between these acoustic spectra.

The auditory spectra are shown in the right-hand panels of Figures 1. By compressing the high frequencies and expanding the low frequencies, some similarities among the pulmonic consonants are enhanced. Dentals have high frequency emphasis, labials have mid frequency emphasis, and velars show at least two broad regions of energy with one lower than the lowest peak in the labials. Additionally, the clicks show some similarities with the pulmonic consonants. The dental click [|] looks more like a pulmonic dental in the auditory spectra than it does in the acoustic spectra. Also the alveolar click [!] has a lower frequency peak than either of the other two clicks. So, some aspects of the spectra are visually enhanced in the auditory display, and the enhancement is non-arbitrary because it is the product of a psychoacoustic model of auditory processing. However, our visual inspection of these displays is still subjective. Just as with acoustic spectra, it is necessary to decide what aspects of the auditory spectra to focus on.

Figure 2 shows the results of the cluster analyses. The bottom panel shows that the auditory spectra of stops and fricatives were grouped by place of articulation. This analysis also produced an interesting picture of click/non-click cross-classifications. The alveolar click [!] was grouped with the velars as Traill (1992a) predicted. Note, however, that there was no evidence for [grave] in the JFH sense because labials did not pattern with velars. The analysis of acoustic spectra (shown in the top panel of Figure 2) is problematic. The natural classes predicted in the

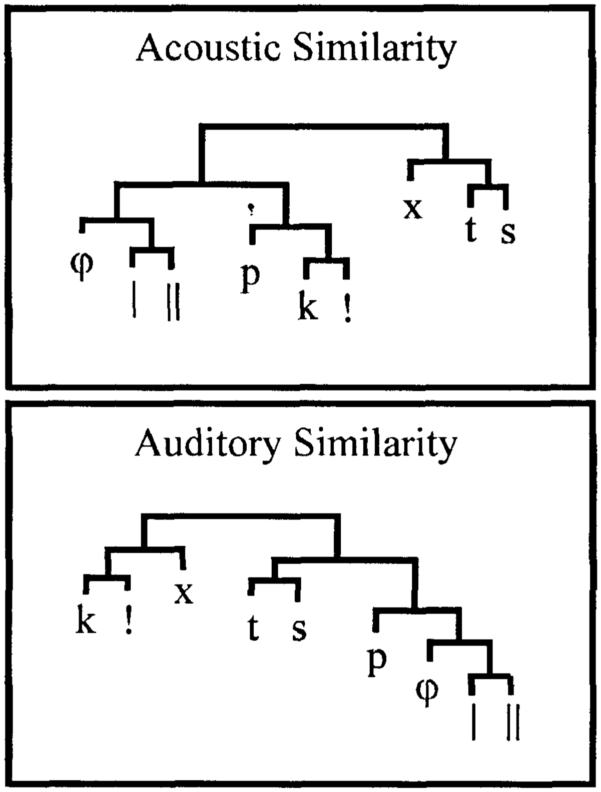


Figure 2. Results of the cluster analyses. Top panel: cluster analysis of LPC spectra. Bottom panel: cluster analysis of auditory spectra.

acoustic analysis are phonologically bizarre. The acoustic spectra of [p] and [k] were grouped together as the JFH definition of [grave] would predict, but the analysis did not result in coherent classification system because place of articulation among the pulmonic sounds was not preserved. For instance, the analysis incorrectly predicts that [x], [t], and [s] form a natural class.

Spectral Properties.

It is important to note that the cluster analysis of auditory spectra grouped the alveolar click [!] with the velar pulmonics. This supports Traill's argument that the historical association of alveolar clicks with velar pulmonic consonants in the Khoe dialects involves sound similarity. However, cluster analysis has an important possible limitation when applied to speech spectra. It focuses on spectral details rather than general (and perhaps somewhat abstract) patterns. So, spectral differences which may be due to a spectral frequency shift may affect cluster analysis as much as a difference in overall spectral shape. JFH pioneered some approaches to extracting spectral properties which reflect the general shape of the spectrum arguing that such general patterns are more linguistically relevant than absolute detailed spectral matching.

Euclidean distance matrices based on the acoustic and auditory spectra (shown in 7) support the JFH viewpoint. These matrices suggest that the similarity relations among the spectra may have been underspecified in the cluster analysis.[3] The matrix in (7a) shows that the acoustic spectrum of [ǀ] was about as similar to the spectrum of [t] as to [p], and most similar to [ɸ], while the acoustic spectrum of [ǁ] was very similar to both [p] and [ɸ]. The dental [ǀ] and lateral [ǁ] clicks had very similar spectra (especially above 4 kHz). On the other hand, although both [ǀ] and [ǁ] group with the pulmonic labials in the cluster analysis of auditory spectra, (7b) suggests that [ǀ] is more similar to the dental sounds in the auditory analysis than in the acoustic analysis. The dental click [ǀ] was about as similar to [t] (RMS=66dB) as to [ǁ] (RMS=65dB) in spite of the fact that [ǀ] was produced in the environment of a round vowel while [t] was not.

Yet cluster analysis missed the similarity between [ǀ] and the pulmonic dentals that is apparent in (7b). In order to understand this it is important to realize that initially cluster analysis uses distances as in (7), but each time a cluster is identified the elements are merged and the average spectrum is used to represent the cluster. For example, the first step in an analysis of the auditory data would be to merge [t] and [s] since the distance between their spectra is smaller than any other distance in (7b). Then a new distance matrix is constructed with the average of the [t] and [s] spectra representing the new cluster. Note in (7b) that although [ǀ] is more similar to [t] than it is to [ǁ], it is not more similar to the average of [t] and [s] than it is to [ǁ]. Therefore, cluster analysis puts [ǀ] in with [ǁ]. It seems inevitable that the distances in (7) reflect not only important phonetic differences, but also random variance which listeners disregard. So a more robust phonetic analysis may require that we pull out abstract spectral properties rather than rely on the unanalyzed spectrum. One method of extracting spectral properties which has been useful in previous research is to calculate the statistical moments of the spectrum.

The idea of using spectral moments for the quantification of spectral shape features in speech analysis appears to have originated with JFH (1963) and has been pursued more recently by Forrest et al. (1988). In moments analysis, some abstract properties of an acoustic or auditory spectrum are described by calculating the statistical "moments" of the spectrum. The first moment

[3] It should be noted that the cluster analyses used distance estimates derivable from those shown in (7).

(7) Root Mean Square (RMS) distances (in dB) between acoustic spectra and auditory spectra.

(a) Acoustic Distances

	t	k	ɸ	s	x	|	!	||
p	143.27	70.83	86.62	191.25	135.48	87.42	50.70	60.27
t		118.62	82.24	**78.51**	99.85	**86.80**	131.55	**105.80**
k			65.38	160.04	100.67	97.42	50.58	77.61
ɸ				114.83	94.41	68.51	66.74	65.19
s					115.24	**136.25**	171.55	**159.23**
x						115.24	109.42	124.86
|							85.68	**45.56**
!								63.21

(b) Auditory Distances

(b)	t	k	ɸ	s	x	|	!	||
p	146.56	99.85	79.46	171.77	155.95	108.90	72.68	75.33
t		128.56	93.36	**55.05**	138.52	**66.05**	137.12	**105.80**
k			75.98	150.38	101.25	126.44	55.59	95.48
ɸ				109.05	133.10	81.31	80.01	60.16
s					150.56	**100.85**	159.20	**136.83**
x						158.41	115.08	155.01
|							117.73	**65.45**
!								89.69

is the spectrum's weighted **mean**.[4] If the spectrum has more high frequency energy than low, the spectral mean will be high. If low frequency energy predominates, the spectral mean will be low. JFH suggested that the spectral mean is one possible correlate of the feature [grave] but they note that if we define gravity in terms of the spectral mean all vowels would be [+grave] while almost all consonants would be [-grave] and any value that the relative values of the spectral mean has in classifying sounds is lost because of these large absolute differences. The second moment is the spectral **variance**. JFH suggested the second moment as a possible correlate of the [compact]/[diffuse] distinction, but later research (Forrest et al., 1988) did not support this proposal. In the present study also the second moment did not produce a coherent division of the sounds for either the acoustic or the auditory spectra. The third moment of the spectrum is its **skew**. If the energy in the spectrum falls equally on both sides of the mean, the spectrum has a skew of 0. If there is more energy above the mean than below it the spectrum has a positive skew, and conversely, spectra with more energy below the mean than above it have negative skew. This measure is relevant for the definition of [grave] and was suggested by JFH as an alternative to the spectral mean because it is not sensitive to the absolute value of the mean. The fourth moment, **kurtosis**, is a measure of the peakedness of the spectrum which Forrest et al. suggested as a correlate of the compact/diffuse dimension instead of using variance as suggested by JFH.

[4] Jakobson, Fant & Halle (1963) used the term "center of area". Others call this the spectral "center of gravity".

Forrest et al. (1988) found that the acoustic spectrum of [t] had a higher mean than did those of [p] and [k]. That result is replicated in Xhosa (Figure 3, top panel). They also found that the acoustic spectrum of [t] had lower skew than [p] or [k], and this result was also replicated. However, the spectrum of [k] did not have greater kurtosis than the other stops as would have been predicted from the JFH description and Forrest et al.'s findings.

As in the cluster analysis of the acoustic spectra, the results of the moments analysis do not give a very coherent picture of Xhosa clicks and pulmonics. The velar fricative [x] is closer to [t] than to [k], and the labial fricative [ɸ] is closer to [k] than to [p]. Forrest et al. used spectral moments to classify fricatives (but not to cross-classify fricatives with stops) but reported only the classification results not the moments data which served as input to the analysis, so we don't know if the lack of coherence seen in the present analysis differs from their findings or not.

The results of the moments analysis of the auditory spectra also conform to Forrest et al.'s findings.[5] The auditory spectrum of [t] had a higher mean than did those of [p] and [k]. Also [k] and [p] were more positively skewed than [t] as they tended to be in Forrest et al.'s study. Interestingly, the spectral mean, skew and kurtosis of these Xhosa consonants were correlated with each other. The correlation between mean and skew was strongest ($r^2=0.938$), while the correlation between mean and kurtosis was only slightly weaker ($r^2=0.763$). This suggests that there is only one effective dimension among these measurements, and that the spectral mean, skew, and kurtosis are simply different ways of measuring the same thing for these particular sounds. It isn't clear whether this dimension is best expressed as the spectral mean (an absolute value) or skew (a relative value), or whether the same dimension will capture the same sorts of distinctions among other sounds as well.

As was the case in the cluster analysis, moments analysis of auditory spectra produced a coherent picture of the Xhosa pulmonics (stops and fricatives produced at the same place of articulation are close together in the bottom panel of Figure 3) and an interesting pattern of click/nonclick cross-classifications. However, unlike the cluster analysis, spectral moments analysis groups the dental click [ǀ] with the dental pulmonic sounds, while (as in the cluster analysis) the lateral click [ǁ] is closer to the labial pulmonics, and (as suggested by Traill, 1992a and the cluster analysis) the alveolar click [!] is quite similar to [k]. One other difference between the cluster analysis of the auditory spectra and this moments analysis is that the moments analysis resulted in a clear division between sounds that have a strong high frequency component ([t s ǀ]) and other sounds, corresponding to the JFH description of the acute/grave distinction.

This coherent division of the sounds is also interesting because the consonants were produced in different vowel environments. Sands (1991) has shown that Xhosa click spectra are influenced by labial coarticulation, so the fact that the dental click [ǀ] patterned with the dental pulmonics in the moments analysis is impressive because the click was produced before [u] while [t] and [s] were both produced before [a]. Similarly, the fricatives [ɸ] and [x] patterned with their homorganic stops in the auditory moments analysis despite coarticulatory rounding from the following [u].

[5] Forrest et al. (1988) used a Bark scale to perform an "auditory" analysis. It should be noted that they did not use an auditory model as I have here, but rather simply changed the frequency scale of the acoustic spectrum.

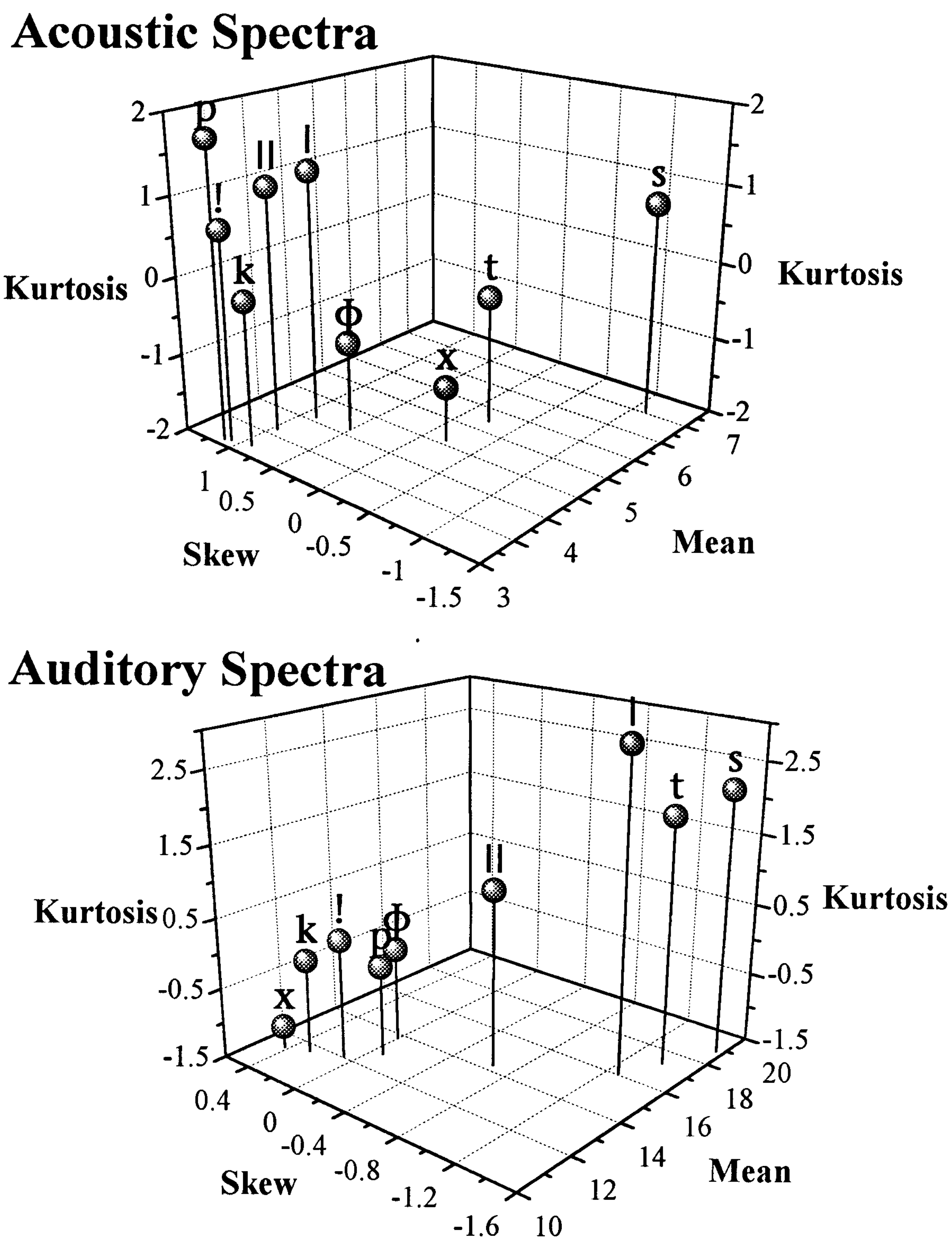


Figure 3. Results of the spectral moments analysis of the average acoustic spectra (top) and average auditory spectra (bottom).

Conclusions

The main finding of this study was that the alveolar click [!] and the velar stop [k] as produced by a native speaker of Xhosa were objectively similar to each other both acoustically and auditorily. This finding supports Traill's (1992a) contention that the proper statement of click replacement in Ts'ixa requires acoustically defined features.

An additional important finding was that analyses of auditory spectra produced different results which differed from the results of identical analyses of acoustic spectra of the same sounds. There are two reasons to prefer auditory spectra over acoustic spectra in studies that aim to predict linguistic natural classes. First, auditory spectra are preferable *a priori* because they represent a level of representation which is closer to the listener's experience of speech sounds. Second, in practical terms, this study found that auditory spectra produced phonologically coherent results in both cluster analysis and in moments analysis despite the varying vowel contexts in which the consonants had been produced, while acoustic spectra of the same sounds did not.

Finally, this study also suggested that in the auditory spectra of voiceless consonants spectral mean, skew, and kurtosis are highly correlated and seem to reflect one dimension along which these sounds differed. It was suggested that this dimension may have been most closely related to either spectral mean or skew, and that only further research can determine the potential importance of this finding.

Acknowledgments

Thanks to Ian Maddieson for his efforts in organizing the 1992 West Coast Phonetics Forum at Berkeley, CA and for asking me to make a presentation, and to Peter Ladefoged for his insightful comments on an earlier version of the paper. A truncated version was also presented to the 67th annual meeting of the Linguistic Society of America. I appreciate the comments and suggestions I received at WCPF and LSA, and also from my friends and colleagues at UCLA and UAB. My address is: Dept. of Biocommunication, UAB, Birmingham, AL 35294-0019

References

Everitt, B.S. (1980) *Cluster Analysis.* 2nd Ed., London: Heineman Educational Books Ltd.

Fletcher, H. & Munson, W.A. (1933) Loudness, its definition, measurement and calculation. *J. Acoust. Soc. Am.* **5**, 82-108.

Forrest, K., Weismer, G., Milenkovic, P. & Dougall, R.N. (1988) Statistical analysis of word-initial voiceless obstruents: Preliminary data. *J. Acoust. Soc. Am.* **84**, 115-123.

Jakobson, R., Fant, G. & Halle, M. (1963) *Preliminaries to Speech Analysis.* Cambridge, MA: MIT Press.

Ladefoged, P. & Traill, A. (1984) Linguistic phonetic description of clicks. *Lg.* **60**: 1-20.

Ladefoged, P. & Traill, A. (to appear) Clicks and their accompaniments.

McDonough, J. (1993) On the phonological representation of the feature 'lateral'. Paper presented at the 67th Annual Meeting of the Linguistic Society of America.

Patterson, R.D. (1976) Auditory filter shapes derived with noise stimuli. *J. Acoust. Soc. Am.* **59**, 640-654.

Sands, B. (1991) Evidence for click features: acoustic characteristics of Xhosa clicks. *UCLA Working Papers in Phonetics.* **80**, 6-37.

SAS (1982) *SAS User's Guide: Statistics 1982 Edition.* Cary, NC: SAS Institute Inc.

Schroeder, M.R., Atal, B.S. & Hall, J.L. (1979) Objective measure of certain speech signal degradations based on masking properties of human auditory perception. In B. Lindblom & S. Öhman (eds). *Frontiers of speech communication research,* London: Academic Press.

Steriade, D. (1992) Closure, release and nasal contours. To appear in M.Huffman and R.Krakow (eds.) *Nasality: Phonological and phonetic properties,* Academic Press.

Stevens, K.N. & Blumstein, S.E. (1978) Invariant cues for place of articulation in stop consonants. *J. Acoust. Soc. Am.* **64**, 1358-1368.

Traill, A. (1985) *Phonetic and Phonological Studies of !Xóõ Bushman.* (Quellen zur Khoisan-Forschung, 1.) Helmut Buske, Hamburg.

Traill, A. (1986) Click replacement in Khoe. In *Contemporary Studies on Khoisan in Honor of Oswin Köhler.* Rainer Vossen & Klaus Keuthmann (Eds.) (Quellen zur Khoisan Forshung, 5) Helmut Buske, Hamburg.

Traill, A. (1992a) Place of articulation features for clicks: Anomalies for universals. MS. University of Witwatersrand, Johannesburg, Department of Linguistics.

Traill, A. (1992b) The perception of clicks in !Xóõ. MS. University of Witwatersrand, Johannesburg, Department of Linguistics.

Zwicker, E. (1961) Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *J. Acoust. Soc. Am.* **33**, 248.

Spectral integration in vowel perception: Matching and discrimination studies

Keith Johnson
Marisa Fernandez
Michael Henninger
Jim Sandstrum

Introduction

This paper is about the **spectral integration** of formants in vowel perception. Chistovich and her colleagues demonstrated in the late 70's that if you change the relative amplitudes of two closely spaced formants in a two-formant vowel, the formant frequency of the best matching one-formant vowel will change, provided the formants are "closely spaced". Chistovich & Lublinskaja (1979) reported that the critical distance for this center of gravity effect is about 3.5 Bark. This estimate of the critical distance and the hypothesis that vowel formants are percepually merged has been often used but seldom studied. A fact which is all the more remarkable given that the authors served as their own listeners. We will present 3 experiments which were designed to investigate the spectral integration of vowel formants.

The experiments test the hypothesis that the center of gravity effect is caused by spectral integration, and so when two formants are within some critical distance they merge into a single perceptual formant.

Experiment 1

The first experiment was a matching study. Listeners heard synthetic stimuli presented in pairs. The first stimulus in each pair had two formants, the second had only one. The 5 naive listeners were asked to adjust the frequency and bandwidth of the formant in the one-formant stimulus, until the two stimuli sounded as similar as possible.

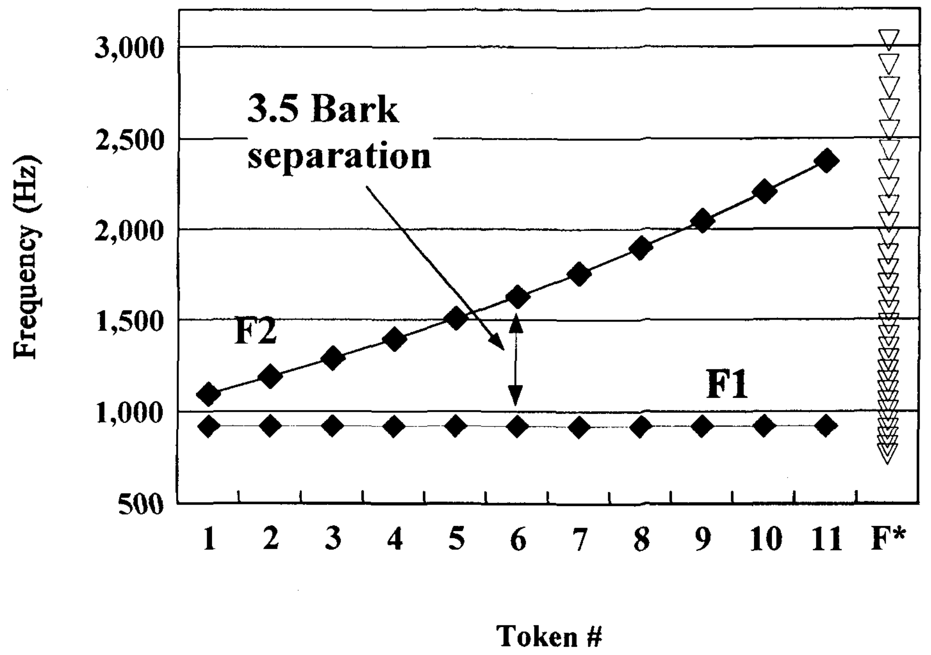


Figure 1. Formant frequencies of the synthetic stimuli used in Experiment 1. The solid symbols show the frequencies of F1 and F2 of the two-formant stimuli, and the open symbols on the right show the frequencies of F* in the one-formant stimuli. As indicated, F1 and F2 are separated by at least 3.5 Bark in two-formant stimuli 6-11.

The two-formant stimuli formed an 11-step continuum (shown in Figure 1) with F1 fixed and F2 varied across the members of the continuum. When the F2 was low, these stimuli sounded like the vowel in "odd", and when the F2 was high, they sounded like the vowel in "add". The distance between F1 and F2 was 3.5 Bark in token 6.

The one-formant stimuli also formed a continuum, labelled F* in Figure 1. Their frequencies ranged from 764Hz to 3027Hz in 0.3 Bark steps. We synthesized 8 stimuli at each frequency step spanning a range of bandwidth values from 50Hz to 400Hz in 50Hz steps. This manipulation was added because we noticed that single formant stimuli (especially those with high F*) sounded more natural with wider bandwidths.

The spectral integration hypothesis leads to a prediction about this matching experiment (illustrated with some hypothetical data in Figure 2). If closely spaced formants are integrated into a single perceptual formant, we expect that, when the formants of the two-formant standard are close to each other, listeners will choose F* values which are between the formants of the two-formant standard, but when they are not close to each other, listeners will tend to adjust F* so that it matches either the F1 or the F2 of the standard. Chistovich et al. (1979) stated that they expected F* to match the F2 of the standard. We can see no principled motivation for this expectation. Note that this pattern of results gives us a definition of "closely spaced" (the critical distance between formants), as well as a test of the spectral integration hypothesis.

As predicted, listeners did tend to choose F* values which were between F1 and F2, early in the continuum (see Figure 3). The boxes in Figure 3 enclose 50% of the responses, the notches enclose 33% of the responses, and the horizontal line marks the median response. As can be seen in the figure, there was greater variability in the listener's responses as F1 and F2 separated. However, we are hard pressed to identify any point along the continuum where there is a sudden shift in the response pattern. Also, contrary to our expectation, listeners chose F* values in between the formants even when F1 and F2 were widely separated.

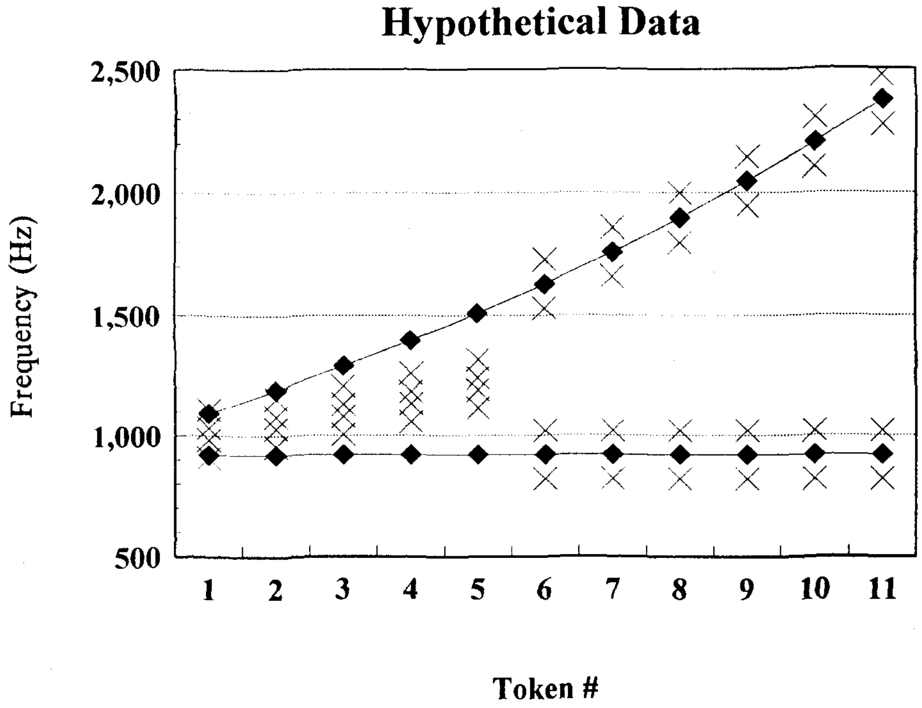


Figure 2. Pattern of matching responses predicted by the spectral integration hypothesis. Frequencies of F1 and F2 of the two-formant stimuli are indicated by the filled diamonds, and hypothetical best matching one-formant F* values are marked with x.

Our decision to allow the listeners to adjust the bandwidth of the one-formant stimuli, may have clouded the results of this experiment. As Figure 4 shows, the listeners tended to choose tokens with wide bandwidths (that is, the stimuli which sounded to us less harsh or nonspeech-like), so, the F* choices may have been corrupted by the less sharply defined spectral envelopes of these wide bandwidth tokens.

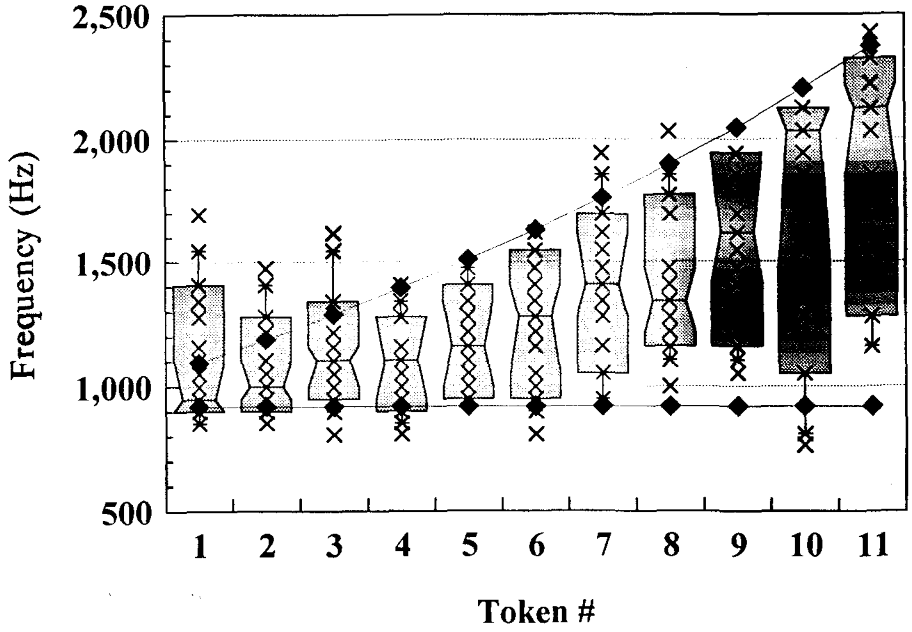


Figure 3. Results of Experiment 1. Frequencies of F1 and F2 are marked by filled diamonds and the observed F* values are marked with x. The overlaid boxes show the median response and the thirds and quartiles of the response distributions. The error bars include 90% of the responses.

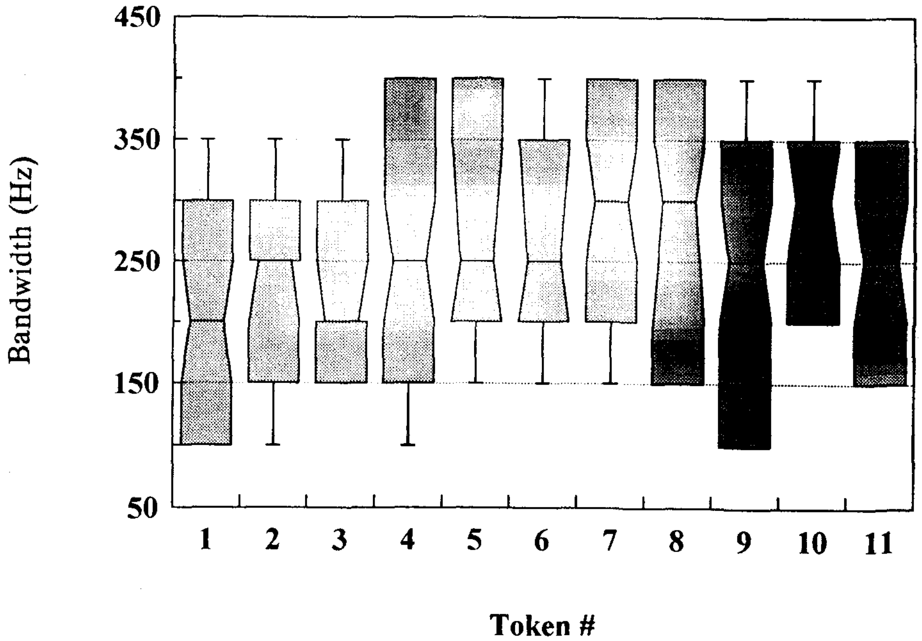


Figure 4. Results of Experiment 1. Bandwidth values selected by the listeners for each two-formant token are plotted by token number. Box and whisker display as in Figure 3.

Experiment 2

Therefore, in Experiment 2 we tested the effect of changing the bandwidth of the F* in a matching experiment. We also manipulated the overall amplitude of the two-formant stimuli, on the assumption that higher amplitude might result in some frequency smearing which would affect the results. Unfortunately, our amplitude manipulation was not very large, and no effects of the manipulation were observed. Therefore this report will focus on the bandwidth manipulation.

One group of 3 listeners performed the matching task as in Experiment 1, with narrow-bandwidth F* stimuli, while another group of 3 listeners performed the task with wide bandwidth stimuli. The bandwidths were 150Hz and 250Hz respectively.

Responses in the wide bandwidth condition (Figure 5) were very scattered. There are a few observable trends, such as the steady increase in the median frequency of F* across the continuum, but on the whole the data in this condition were quite messy. Thus, given the relatively wide bandwidths chosen by listeners in Experiment 1, it seems likely that the data in that experiment included some of the variability we see in Figure 5.

Data from the narrow bandwidth condition (Figure 6) showed much less scatter. However, this box-and-whisker display leads to the false impression that listeners were equally probable to chose any of the stimuli surrounded by the boxes. The distribution of responses is clarified in a spectral plot of the same data (Figure 7). In this figure, responses to the stimuli are plotted by token number and F* value with the number of responses coded as a shade of gray. The darker the shading, the greater the number of responses at that F* frequency. Note how the distribution of responses goes from unimodal for tokens 1 through 5 (that is, the responses fall near one frequency) to bimodal for tokens 6 through 11. This is pooled data, so it is important to note that the distributions were bimodal for each listener. Recall that this was the pattern of data we predicted given the spectral integration hypothesis. Chistovich et al.'s prediction that F* would follow F2 was not confirmed by these data, but the location of the shift from unimodal to bimodal performance corresponds with Chistovich & Lublinskaya's (1979) estimate of the critical distance.

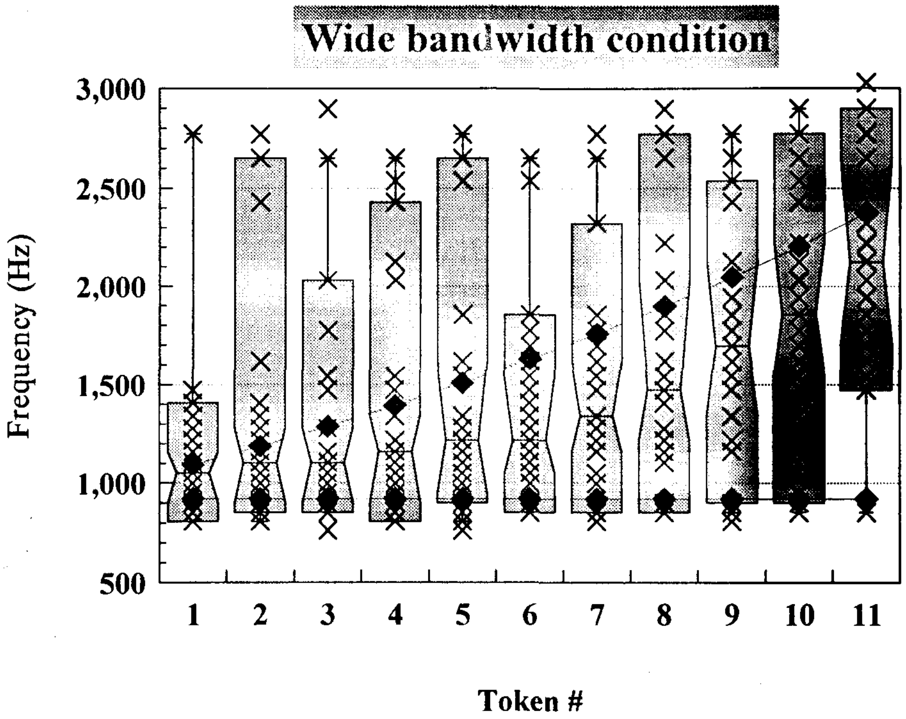


Figure 5. Results of Experiment 2. Data from the wide bandwidth condition.

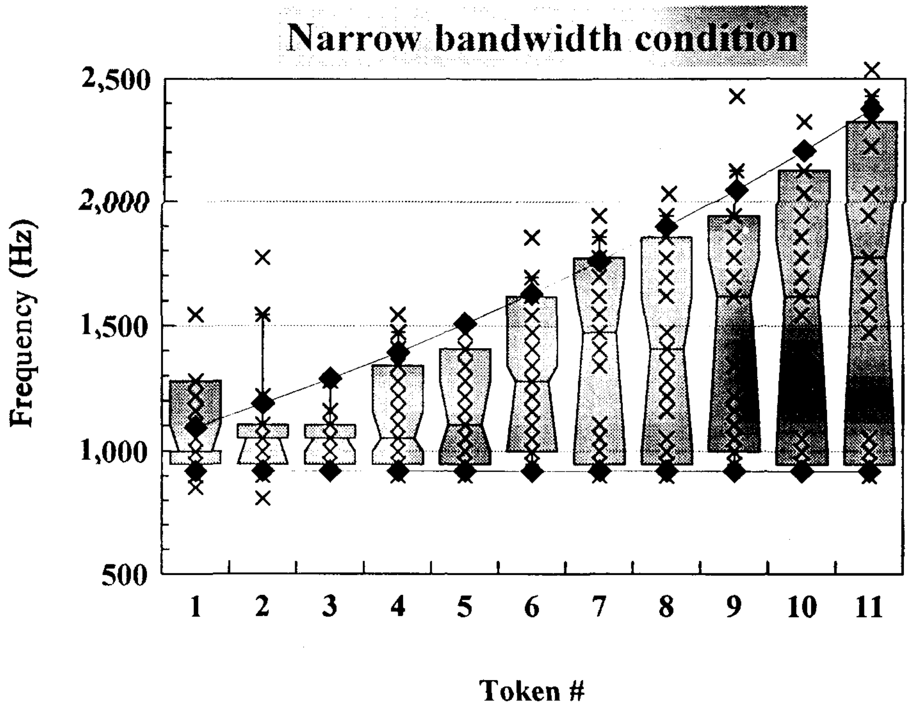


Figure 6. Results of Experiment 2. Data from the narrow bandwidth condition.

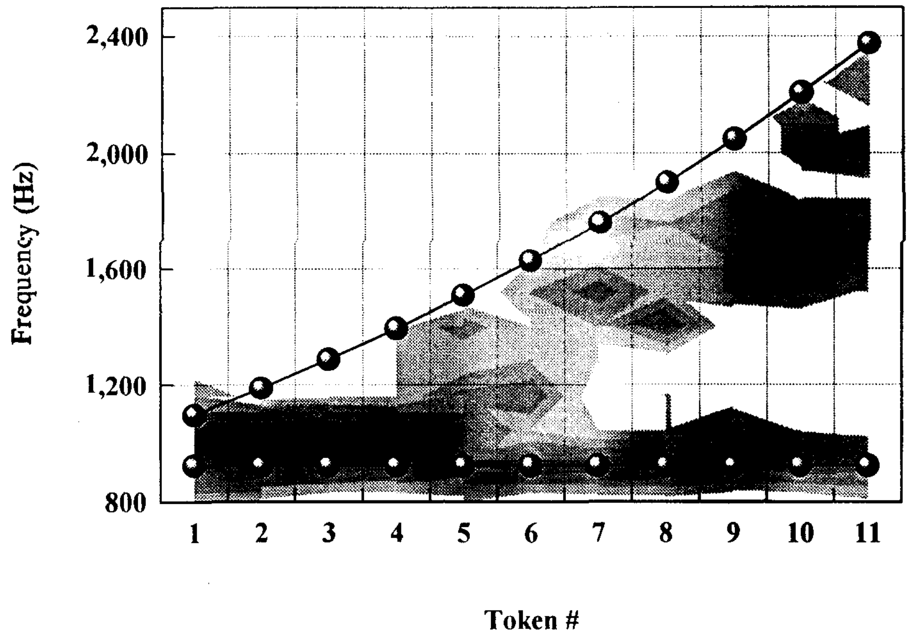


Figure 7. Results of Experiment 2. Data from the narrow bandwidth condition shown in a spectral plot. The number of responses to each two-formant token having a particular F* value is indicated as a shade of gray; the darker the shading, the greater the number of responses.

Experiment 3

Earlier we stated the spectral integration hypothesis this way: "When two formants are within some critical distance they merge into a single perceptual formant". Experiment 2 found evidence in favor of this hypothesis. Experiment 3 was designed to further investigate this hypothetical spectral integration.

The experiment was a very simple discrimination task. We saught to discover whether listeners can detect the difference between two-formant stimuli and their best-matching one-formant counterparts. For stimuli in which the two formants are far apart we can't really talk about the best-matching one-formant stimulus because the distribution of best-matches is bimodal, and even when you compare the most common matching tokens they sound nothing like the two-formant stimuli. However, when the two-formants are close to each other, the one- and two-formant stimuli do sound surprizingly similar (to our ears). Klatt (1985) and Hanson & Javkin(1990) asked listeners to rate the similarity of tokens which had the same calculated spectral center of gravity, but different formant frequencies and formant amplitudes. They found that sounds with the same center of gravity were judged to be less similar than acoustically identical sounds. Hanson & Javkin (1990) attributed the difference between matching results and similarity judgements to differences in the tasks. They described the matching task as "linguistic" because it calls for a linguistic judgement. The distinction that Hanson & Javkin (1990) make between linguistic judgements and other sorts of judgements is important, but not very well developed. Our understanding of this distinction is somewhat different from theirs.

When thinking about the relevance of spectral integration in vowel perception it is important to distinguish between (1) information used to identify sounds, and (2) ways that sounds can be similar. We assume that any aspect of a sound in its auditory representation may be used for identification, while judgements of sound similarity may disregard auditory differences (or ignore some of the available information). Therefore, the question we are posing in Experiment 3 can be stated this way: Does spectral integration affect the information available for vowel identification, or does it affect sound similarity only? Many researchers have assumed that 3-Bark integration is relevant for vowel identification, because they have assumed that the merged 'perceptual formant' is an **auditory** property of sounds with close formants. When two formants are closer than 3.5 Bark they become auditorily one and the listener has no record or representation of their former two-formantness. We will call this the **auditory hypothesis**. An alternative hypothesis, which we will call the **perceptual hypothesis**, is that spectral integration data (such as Experiment 2 above) reflects the existence of a dimension of perceptual similarity. However, this perceptual aspect is built up in addition to the auditory representation, and does not replace the more detailed spectral information in the auditory representation. To quote Chistovich et al. (1979), "information about spectrum shape is not exhausted by the centre of gravity location".

These two hypotheses about 3 Bark spectral integration make different predictions about the discrimination of one- and two-formant vowels. The auditory hypothesis predicts that listeners will not be able to discriminate one- and two-formant vowels when the two formants are within the critical distance, while the perceptual hypothesis predicts that listeners **will** be able to distinguish one- and two-formant vowels.

For purposes of this experiment we defined "best-matching" as the average F* value found in the narrow bandwidth condition in Experiment 2. In a pilot experiment, with the authors as listeners, we used long vowels, in the main experiment we used short vowels. We also manipulated the interstimulus interval in the pilot and the main experiment.

Results of the pilot experiment (Figure 8) showed that regardless of "spectral integration" the listeners could easily detect the differences between the one- and two-formant stimuli. This result is consistent with the perceptual hypothesis rather than the auditory hypothesis. But what happens when we bring this discrimination function down from ceiling? We made the task more difficult by reducing the durations of the stimuli from 225ms to 60ms. Discrimination performance for naive listeners at long ISI's was near chance on stimuli 1-3 (Figure 9). It is interesting that the discrimination function is gradual, there is no sudden shift in performance as might be predicted by the auditory hypothesis. Rather, this gradual shift in discriminability seems to be more indicative of a gradual decrease in perceptual similarity.

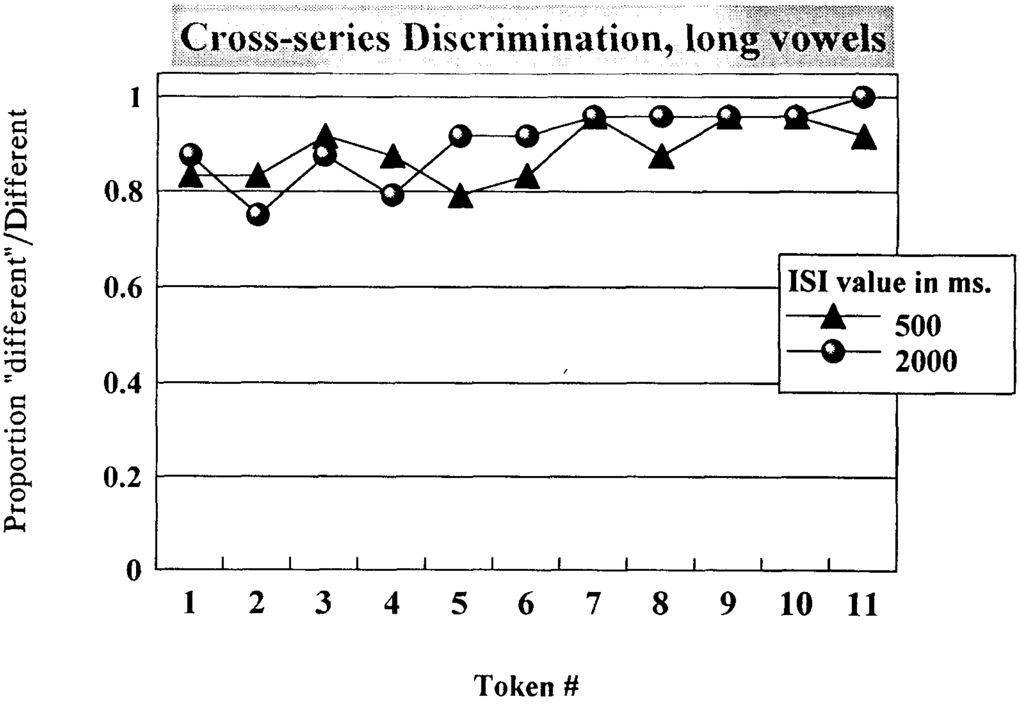


Figure 8. Results of the preliminary cross-series discrimination experiment. The proportion "different" responses for trials in which the stimuli were actually different is plotted by token number for short (triangles) and long (balls) interstimulus intervals.

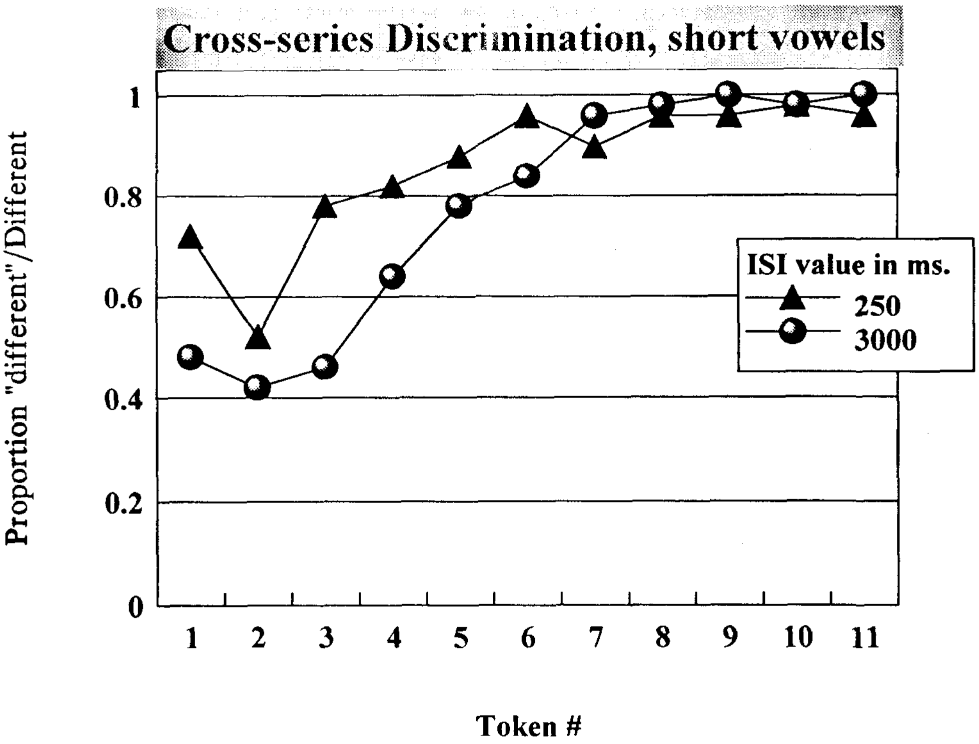


Figure 9. Results of Experiment 3. The proportion "different" responses for trials in which the stimuli were actually different is plotted by token number for short (triangles) and long (balls) interstimulus intervals.

Conclusion

We found evidence in a matching experiment to support the spectral integration hypothesis. The data also confirmed Chistovich et al.'s estimate of the critical distance between formants. Additionally, the results of Experiment 3 suggested that spectral integration affects perceptual similarity, but not auditory representation.

Note

Johnson is now at: Dept of Biocommunication, UAB, Birmingham, AL 35294-0019. The experiments were carried out as undergraduate honors projects (Fernandez - Experiment 1, Henninger - Experiment 2, and Sandstrum - Experiment 3) at UCLA, Spring, 1992. Presented to the 124th meeting of the Acoustical Society of America, November 3rd, 1992.

References

Chistovich, L.A. and Lublinskaja, V.V. (1979) The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research* **1**, 185-195.

Chistovich, L.A., Sheikin, R.L. and Lublinskaja, V.V. (1979) "Centres of Gravity" and spectral peaks as the determinants of vowel quality. In (B. Lindblom & S. Öhman, Eds.) *Frontiers of speech communication research.* London: Academic Press.

Hanson, B.A. & Javkin, H.R. (1990) Evidence for the three bark integration interval. *STL Research Reports* **2**, 2-1 -- 2-7, Santa Barbara, CA: Speech Technology Laboratory.

Klatt, D.H. (1985) A shift in formant frequencies is not the same as a shift in the center of gravity of a multi-formant energy concentration. *J. Acoust. Soc. Am.* **77**, S7.

The hyperspace effect: Phonetic targets are hyperarticulated

Keith Johnson
Edward Flemming
Richard Wright

Abstract

A common but rarely defended assumption is that phonetic reduction processes apply to hyperarticulated phonetic targets. If this assumption is correct, a two stage model of phonetic implementation is indicated; at the first stage distinctive features are mapped to hyperarticulated phonetic targets, and at the second stage these phonetic targets are reduced. The experiments reported in this paper supported this model by showing that phonetic targets are hyperarticulated. Listeners adjusting the first and second formants of synthetic vowels chose values which were only found in hyperarticulated speech.

Introduction

Most approaches to phonetic realization that directly address the issue of casual speech assume canonical phonetic representations, or targets, which are hyperarticulated, with variation being introduced by reduction processes. For example, Lindblom (1990) conceptualizes speech production as a feedback system in which the input is the goal, and the extent to which it is achieved depends on the "gain" of the feedback loop, so the gain is analogous to something like effort. Browman and Goldstein (1990) propose that casual speech variants of canonical lexical gestural representations are produced by increasing gestural overlap and reducing gestural magnitudes.

In fact it appears that this general approach is widely assumed since almost all discussions of the casual speech - clear speech continuum are cast in terms of "undershoot", "reduction" and related concepts. However, this is not the only possible account of these phenomena. We could postulate the existence of both reduction and hyperarticulation processes, in which case the canonical phonetic representation would be of an intermediate level, perhaps akin to citation forms. We can probably reject the third logical possibility, which is the hypothesis that there are only hyperarticulation processes, since the most reduced forms of words can be so indistinct that if they were the starting point it would be difficult to derive the clear distinctions between them that exist in hyperarticulated speech. This is essentially the same type of argument that leads Jakobson & Halle (1956, p. 6), among others, to state that the most clearly articulated speech is most relevant to phonological analysis since it contains the most information.

In this paper, we report the results of several method of adjustment studies in which listeners controlled the first two formants in synthetic vowels. We will show that listeners' responses in the method of adjustment task differ systematically from their own productions and that this systematic deviation from normal speech is consistent with the view that phonetic targets are hyperarticulated.

Phonetic comparison of vowel systems and the method of adjustment

In the method of adjustment (Nooteboom, 1973; Ganong & Zatorre, 1980; Samuel, 1982; Johnson, 1989) a listener is given control over one or more parameters of a speech synthesizer and is asked to adjust them until the machine pronounces a particular speech sound. Where other methods used in speech perception research focus on the boundaries between phonetic categories, the method of adjustment provides information about linguistic/phonetic targets for speech sounds. Repp & Liberman (1987) discuss the need for data on the internal structure of phonetic categories, as opposed to category boundaries, and assume that the method of adjustment provides information about the “prototypes” of phonetic categories (see also Samuel, 1982 concerning this assumption). However, they note that “until recently, no one had used methods designed to identify prototypes” (p. 90) and that “the application of such methods has so far failed to yield entirely satisfactory results”.

The method of adjustment is useful for cross-linguistic comparisons of vowel systems because personal differences in vocal tract anatomy (vocal tract size, oral cavity to pharynx cavity ratio, palate doming, lip shape, etc.) give rise to acoustic differences between speakers (Ladefoged & Broadbent, 1957). Although research since the late 1940's has shown that the speech signal varies quite considerably from speaker to speaker even within the same language and dialect (Joos, 1948; Peterson & Barney, 1952), cross-linguistic acoustic/phonetic comparisons of vowels confound personal and linguistic differences. Thus, any cross-linguistic acoustic comparison of vowel systems includes some unknown amount of personal variation.

One way to compensate for personal variation in making cross-linguistic comparisons is to scale the measured formant values by a factor which is related to vocal tract size. The scale factor may be derived from the range of observed formant values for a particular speaker (Gerstman, 1968), the mean of the observed formant values (Lobanov, 1971), the mean of the log-transforms of the observed formant values (Nearey, 1977), or a function of the speaker's fundamental frequency (F0) (Miller, 1989; Syrdal & Gopal, 1986). Disner (1980) suggested that it is valid to use the formant mean or range "so long as the data are drawn from a single language or dialect, such that the same set of vowel phonemes is shared by all speakers" (p. 257). But, in making cross-linguistic or cross-dialectal comparisons, reliance on mean formant values or formant range is not usually valid. Methods of formant normalization which use the speaker's F0 are also flawed because they rely on the observed *rough* correlation of F0 with vocal tract length, and are also only valid for comparisons of vowels produced in similar prosodic contexts.

Another class of normalization techniques uses cross-linguistic formant averages as normalizing factors. For instance, Disner (1980) used PARAFAC (a three-mode factor analysis technique, Harshman, 1970) to compare the vowels of English, German, Danish, Swedish, and Dutch. The PARAFAC procedure relates measurements from individuals to the overall mean of the data set and finds speaker constants which scale the individual's vowel space to the overall vowel space. In later work, Disner (1986) compared the vowels of various languages using analysis of variance models which included factors for vowel, speaker and language. In both of these approaches the differences between languages are tested as deviations from cross-linguistic means. Two limitations are inherent in this class of normalization technique. First, only comparable vowel qualities can be tested cross-linguistically, which means that it is not possible to compare whole vowel systems when they contain unequal numbers of vowels. Second, the method assumes that the average individual deviation from the overall mean is not correlated with language. To see the problem with this assumption consider an extreme example. If all of the data for language x is taken from recordings of female speakers while all of the data from language y is taken from recordings of male speakers, an analysis of variance would show quite large differences in formant values as a function of language even though speaker is included as a factor in the model. Thus, although the statistical techniques proposed by Disner may provide a better solution to the normalization problem (for cross-linguistic comparisons) than those provided by other approaches, the problem is not solved by using cross-linguistic formant averages as normalizing factors because the results still depend on the comparability of the groups of speakers who represent each language (see Behne, 1989).

The method of adjustment offers a better way of making cross-linguistic phonetic comparisons of vowel systems. By using a single synthetic voice, the method of adjustment makes it possible to ascertain the listener's expectations for vowel sounds in a particular language *for that voice*. Thus, the linguistic and personal aspects of the speech signal are disentangled.

In the studies of the Southern California English vowel space reported here, we found that listeners chose vowel formants which did not match those produced by *any* speaker in normal speech. The discrepancy between listeners' choices in the method of adjustment and speakers' productions is interesting because it is systematic. The perceptual vowel space was expanded relative to the production space; high vowels were higher, low vowels lower, front vowels more front, and back vowels more back. We will present data which indicates that the perceptual vowel space reflects vowels produced in clear or hyperarticulated speech.

Preliminary study

A preliminary study was designed to give a first indication of the feasibilty of the method of adjustment for cross-linguistic phonetic research. We sought answers to two questions: (1) can listeners do the task in a reasonable amount of time and with reasonably low within- and between-listener variability? and (2) will the task provide interpretable data for cross-linguistic comparisons?

Subjects. Ten female and four male university students served as volunteer subjects. They had self-reported normal speech and hearing and had recently completed a one-quarter course in phonetic transcription. This pool of subjects represented fairly diverse linguistic backgrounds: four were monolingual Southern Californians, six were English-dominant Southern Californians, one was a native of Maryland, two were native speakers of Serbo-Croatian, and one was a native speaker of Spanish. We will present data collected from the Southern Californians and the Serbo-Croatians.

Materials. 330 steady-state isolated vowel stimuli were synthesized using a software formant synthesizer (Klatt & Klatt, 1990). The stimuli were generated by varying F1 and F2 independently over a large range of values. There were fifteen possible values of F1 and twenty-two possible values of F2. F1 ranged from 250 Hz to 900 Hz with a step size of 0.37 Bark, and F2 ranged from 800 Hz to 2800 Hz also with a step size of 0.37 Bark. These formant step sizes are slightly larger than the just-noticeable-differences reported by Flanagan (1957). Figure 1 shows the stimuli used in Experiment 1 (described below) and illustrates the type of two dimensional array also used in the preliminary experiment, although with a slightly larger range of possible formant values.

Other parameters of the synthesizer were fixed across the set, or were estimated by rule given the values of F1 and F2. The duration was fixed at 250 ms. and the F0 started at 120 Hz and fell over the last half of the vowel. F3 was estimated by the regression formulas published by Nearey (1989). The fourth formant was constrained to be at least 300 Hz higher than F3 and no lower than 3500 Hz. The bandwidths of the formants were estimated by regression formulas relating the bandwidth values suggested by Klatt (1980) to F1, F2 and F3. The formulas for the

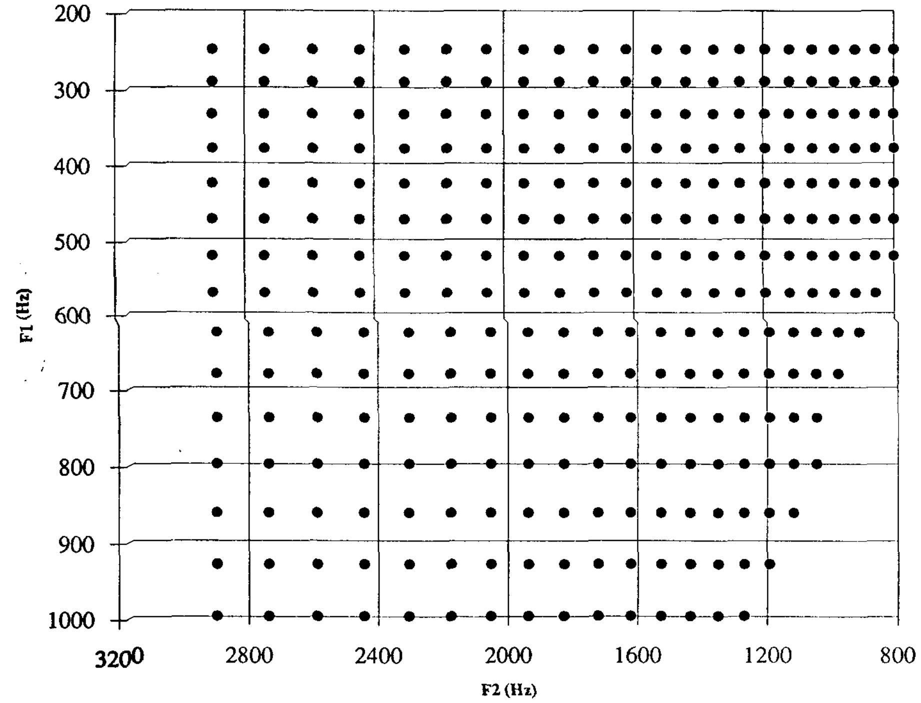


Figure 1 Formant values of stimuli used in the method of adjustment task in Experiment 1. Each filled circle represents a stimulus. The stimuli are equidistant on the Bark scale and therefore are separated by larger Hz intervals as the frequencies of F1 or F2 increased.

bandwidths of F1 through F3 are given in (1) through (3) respectively.

(1) B1 (in Hz) = 29.27 + 0.061*F1 - 0.027*F2 + 0.02*F3, r^2=0.605
(2) B2 (in Hz) = -120.22 - 0.116*F1 + 0.107*F3, r^2=0.497
(3) B3 (in Hz) = -432.1 + 0.053*F1 + 0.142*F2 + 0.151*F3, r^2=0.595

As the r^2 values suggest, formulas (1)-(3) provide only a rough fit to the bandwidth values suggested by Klatt, and extreme formant values resulted in unnatural bandwidths. This contributed to the unnaturalness of stimuli which were already unnatural due to their formant values, while simultaneously contributing to the naturalness of tokens which had humanly possible combinations of formants. The bandwidth of F4 was fixed at 200 Hz. To further increase the naturalness of the stimuli, the "natural" voice-source in the synthesizer was used and the values of amplitude of aspiration, open quotient and glottal tilt varied over time to simulate the changes in glottal vibration seen in naturally produced syllables (see Klatt & Klatt, 1990).

In addition to these synthetic stimuli used in the perception part of the experiment, we compiled a list of common English words illustrating the vowels of English for use in the production part of the experiment. The list was: heed, hid, aid, head, had, HUD, odd, awed, owed, hood, who'd. These words have either a glottal stop or /h/ initially and a final /d/. They were also used as the visual stimuli in the perception experiment.

Procedure. The experiment involved two tasks. First, the subjects were asked to read ten repetitions of the list of English words (in the carrier phrase say ___ again). The order was randomized separately each time through the list. The subjects were seated in a sound booth and recordings were made using high quality equipment (Sennheiser microphone, Symetrix SX202 preamplifier, and Tascam 122 cassette recorder). Formant values from these recorded utterances were measured using CSpeech (Paul Milenkovic) from an LPC spectrum which was calculated at a point early in each vowel as determined in a digital waveform display.

The second task was the method of adjustment task using the same list of words as visual stimuli. This part of the experiment was run online by an IBM PC-AT. Stimuli were stored on disk and were converted to analog waveforms by a Data Translation DT2801A board in the PC. The sampling rate was 10kHz and the signal was low-pass filtered at 4.2 kHz before being amplified (BGW Systems, Model 85) and presented diotically over headphones (Sony MDR-V4). (For more details concerning the setup see Johnson & Teheranizadeh, 1992.) The listener saw a word at the top of a CRT screen, and a two dimensional grid (see Figure 2 for a sample display). Each square in the grid corresponded to one of the vowel sounds in the F1, F2 matrix, and the listener used a mouse to select a particular square and clicked a mouse button to hear the synthetic vowel associated with that square. The task was then to find the location in the grid which produced a synthetic vowel which sounded like the vowel in the word. After choosing the F1 and F2 values for the vowel of a particular word, the listeners were asked to rate their choice on a scale from one to ten. The rating data were used to eliminate mistakes, primarily accidental terminations of trials. This task was repeated 10 times for each of 11 words in the list.

Note one complication in relating the grid shown in Figure 2 to a set of vowel stimuli such as that shown in Figure 1. In the region of the vowel quadrangle where F1 and F2 are close to each other the acoustic vowel space has a corner cut off (F1 was always at least 250Hz below F2), while in the visual display this is not true. This complication was handled by filling in the corner of the visual display with copies of nearby tokens. If the listener were to choose the square which would correspond to an F1 of 1000Hz and an F2 of 1200Hz, for example, a token with the same F1 and the next higher F2 value would be presented. So, the vertical dimension of the display always corresponded to different F1 values but, in one corner of the space, changes in the horizontal dimension of the grid did not result in changes of F2.

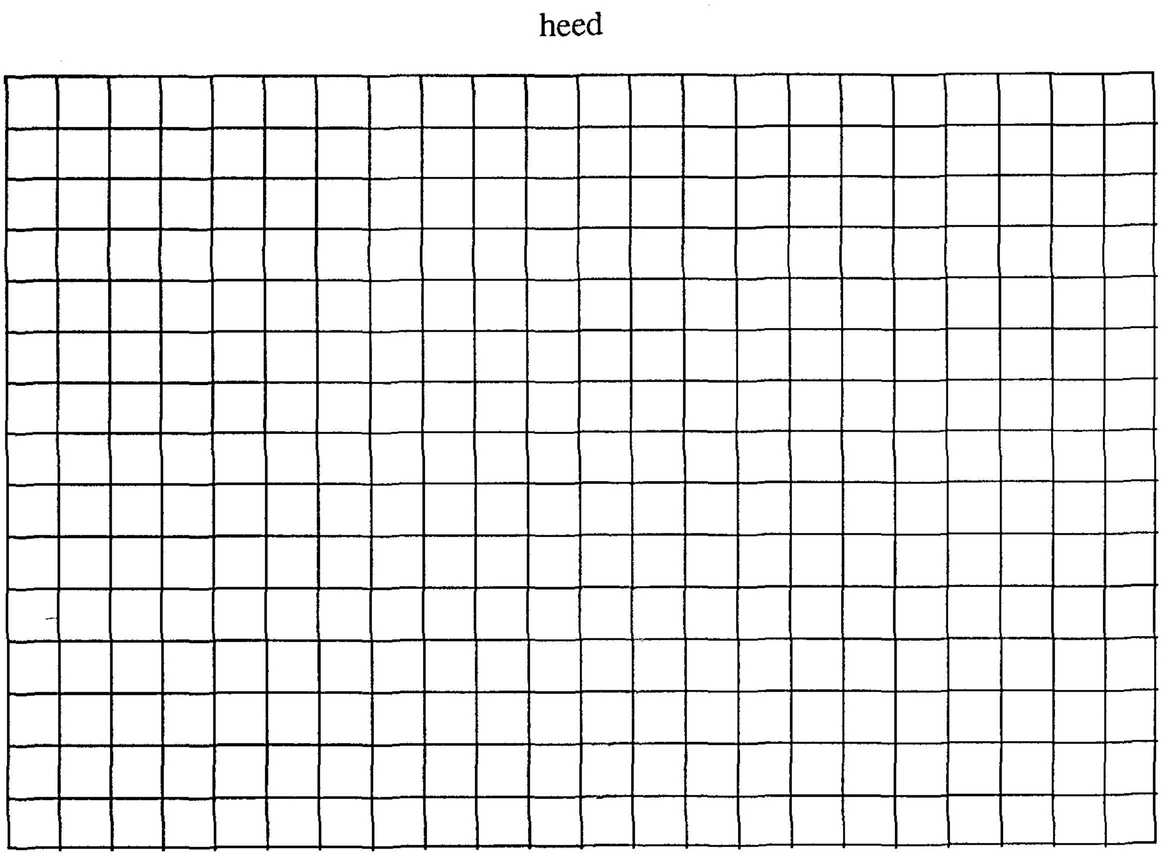


Figure 2 An example of the visual display presented to listeners in the method of adjustment task. The word at the top of the screen changed from trial to trial, and each square (except in the region of vowels with low F2 and high F1) in the display corresponded to a different F1, F2 combination (see text for further details).

Because we wanted to collect several adjustment trials for each of several vowels, and we did not want the listeners to simply rely on visual cues in making their judgements, we changed the orientation of the acoustic vowel space on the screen randomly from trial to trial. So, on 50% of the trials the high F1 stimuli were at the top of the screen and on the other 50% of the trials they were at the bottom of the screen. Similarly, the relationship of F2 to the horizontal dimension of the grid also changed from trial to trial.

Results and Discussion. Table 1 shows the standard deviations in the method of adjustment task for the native Californians. The first two columns show the between-listener standard deviations (calculated across all responses) of F1 and F2 while the next two columns show the average within-listener standard deviations of F1 and F2 (calculated for each listener separately and then averaged). The average ratio of within- to between-listener standard deviation for F1 is 0.83. The ratio of within- to between-listener standard deviation for F2 is 0.73. These ratios suggest that most of the variability in the method of adjustment task occurred within the responses of individual listeners rather than appearing as between-listener variability in the formant values chosen. Note also that standard deviations tend to be higher for higher formant values (SDF2 is higher than SDF1). This reflects the fact that the equal Bark increments in the stimulus set (Figure 1) resulted in larger acoustic differences between the stimuli as the frequency increased. One surprizing observation to be noted in Table 1 is that /i/ showed more between-listener variation in F2 than did the other vowels and /u/ showed more between-listener variation in F1. Listeners were internally consistent in their choices for these vowels but showed relatively more discrepancy

Table 1. Comparison of variability between and within listeners in the preliminary experiment. The first two columns show the overall standard deviations (in Hz) of F1 and F2 in the method of adjustment trials. The middle two columns show the average within-subject standard deviations (in Hz) for the same data. The last two columns show the ratio of within to between listener variability.

	Between Subj.		Within Subj		Ratios	
word	SDF1	SDF2	$\overline{\text{SDF1}}$	$\overline{\text{SDF2}}$	F1	F2
heed	19.5	164.6	16.2	99.9	.83	.61
hid	29.7	158.4	26.3	124.9	.89	.79
aid	55.9	229.6	37.1	151.2	.66	.66
head	49.7	173.0	41.7	114.6	.84	.66
had	61.4	175.5	56.4	126.3	.92	.72
odd	56.5	56.6	50.0	52.2	.89	.92
awed	55.7	42.4	48.8	35.8	.88	.84
HUD	48.9	98.5	41.1	70.6	.84	.72
owed	35.5	57.6	31.5	46.0	.89	.80
hood	36.2	96.9	35.7	85.8	.99	.89
who'd	67.9	112.4	41.9	86.2	.61	.77
average	47.0	124.1	38.8	90.3	.83	.73

between listeners than with most of the other vowels. Also note that the choices for F1 and F2 of aid were more variable between listeners than they were within listeners. This is probably due to the fact that the synthetic stimuli were steady-state vowels while the target vowel is diphthongal. The vowel in owed is less diphthongal in this dialect than it is in other dialects of English. Although there are some interesting patterns in the variability found in this preliminary study, the most important finding is that the variability is low; averaged standard deviations of about 50Hz for F1 and 100Hz for F2.

Measurements of the acoustic vowel space produced by 8 Southern Californian males in the citation readings of Experiment 1 (shown in Figure 3a) suggested that the vowels in odd and awed are merged. (These production data from Experiment 1 are shown here because there were too few male subjects in the preliminary experiment and because it is necessary to compare these method of adjustment results with male formant values because the synthetic stimuli had a male voice.) Note also that /eɪ/ and /ɪ/ had very similar formant values.

The merger of the vowels in odd and awed was also found in the perception results (Figure 3b) from the native Southern California subjects in the preliminary experiment. In addition to the merger of the low back vowels, the data indicate that the listeners kept the vowels of aid and hid more spectrally separated in the method of adjustment than they did in production. This tendency was noted in an earlier method of adjustment study of vowels (Johnson, 1989). As suggested in that earlier report, it may be that when potential cues such as intrinsic vowel duration, pitch, and formant movement are not available, listeners will exaggerate an existing small spectral difference in the method of adjustment in order to maintain a linguistic distinction. There is also the possibility that the production data in Figure 3a don't accurately represent the spectral properties of /eɪ/ because of our choice of measurement location. This possible explanation of the discrepancy between the production and perception result seems rather unlikely because we made the acoustic measurements early in the vowel. Therefore, we would expect if anything to see an even lower F1 and higher F2 for /eɪ/ if we were to measure later in the vowel. Thus, if we were to make the formant measurements later in the vowel we would expect to see even more discrepancy rather than less.

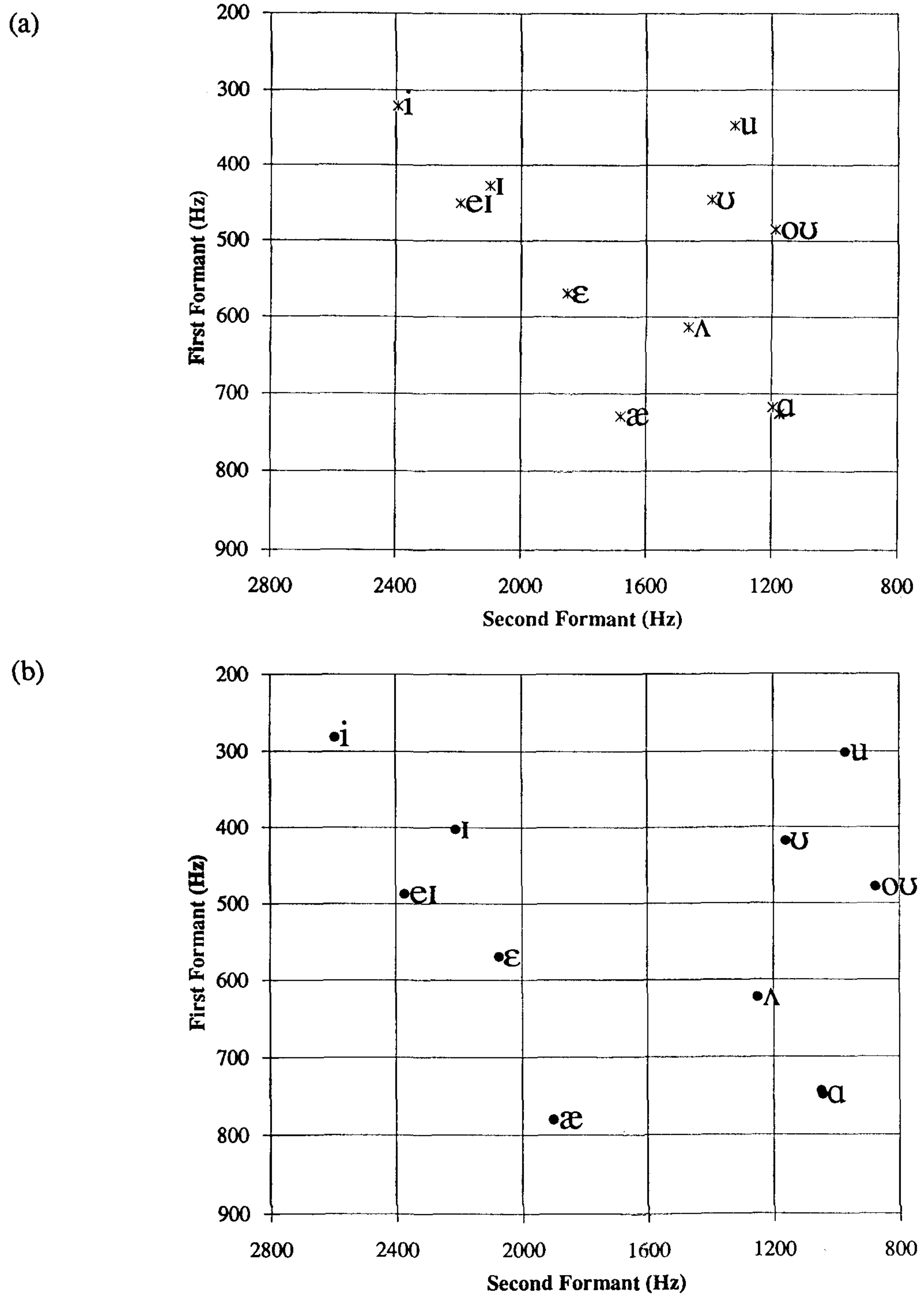


Figure 3 (a) Average measured formant values of vowels produced by the eight male native Southern California English speakers from Experiment 1. These vowels were produced in the "citation" reading condition. (b) Average method of adjustment results for the native Southern California English speakers in the preliminary experiment.

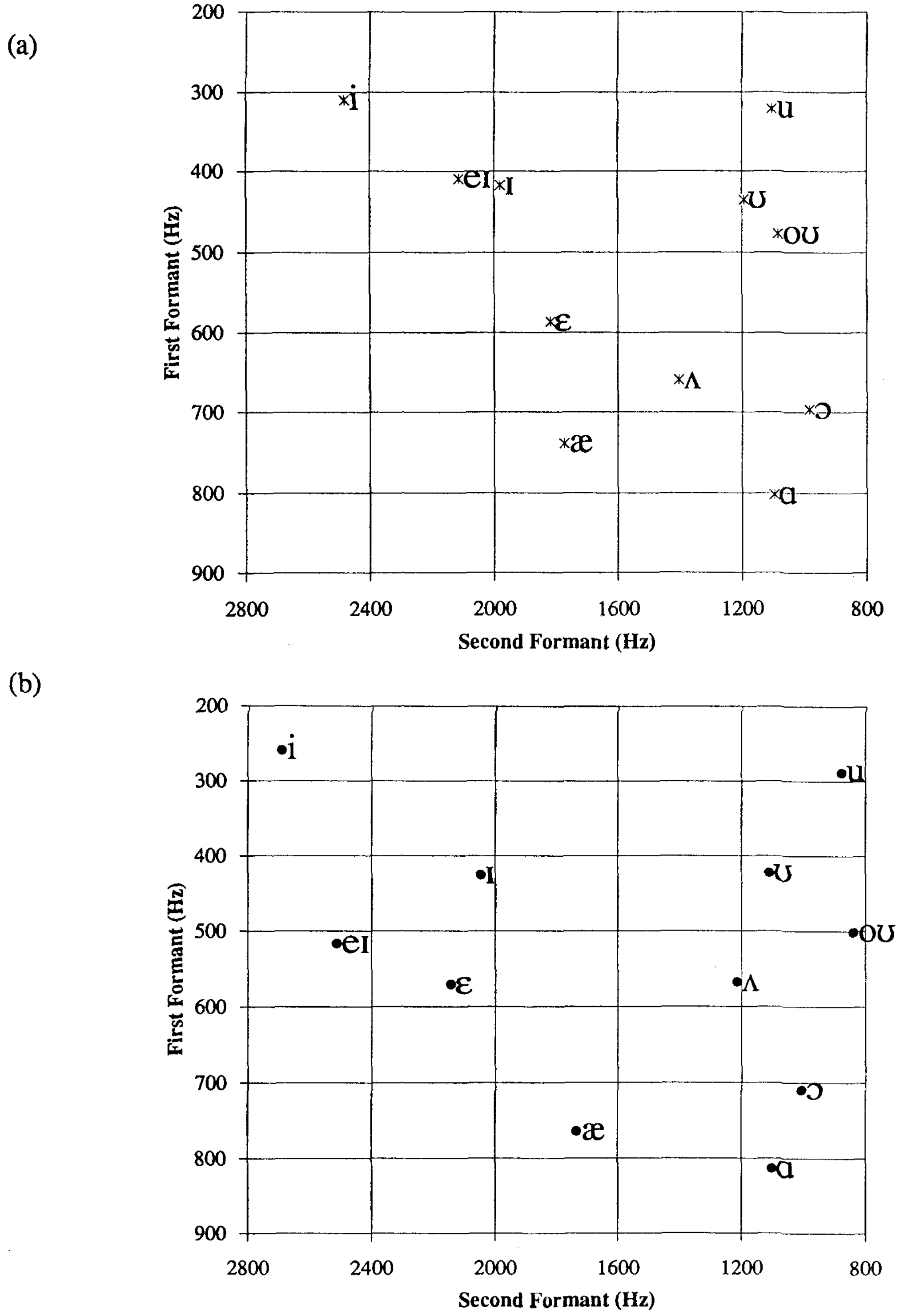


Figure 4 (a) Average measured formant values of vowels produced by one male speaker in the preliminary experiment. This speaker maintained a distinction between the vowels in awed and odd. (b) Average method of adjustment results obtained from the same speaker.

One of the listeners in the preliminary experiment was older than the other listeners and was born in New York City. Figure 4a shows that he maintained the distinction between odd and awed in production. In other respects his acoustic vowel space is similar to the average vowel space shown in Figure 3a. Figure 4b shows that this subject also selected different formant values for the vowels in odd and awed in the perception experiment (the difference between /eɪ/ and /ɪ/ was also quite expanded in the perception space). Keep in mind that although the comparison between production vowel spaces (Figures 3a & 4a) gives us about the same picture of the difference between this speaker and the others, the perception vowel spaces (Figures 3b & 4b) allow for a better comparison because the listeners are telling us their expectations for the vowel space of a single synthetic speaker, thus speaker and dialect information are not confounded.

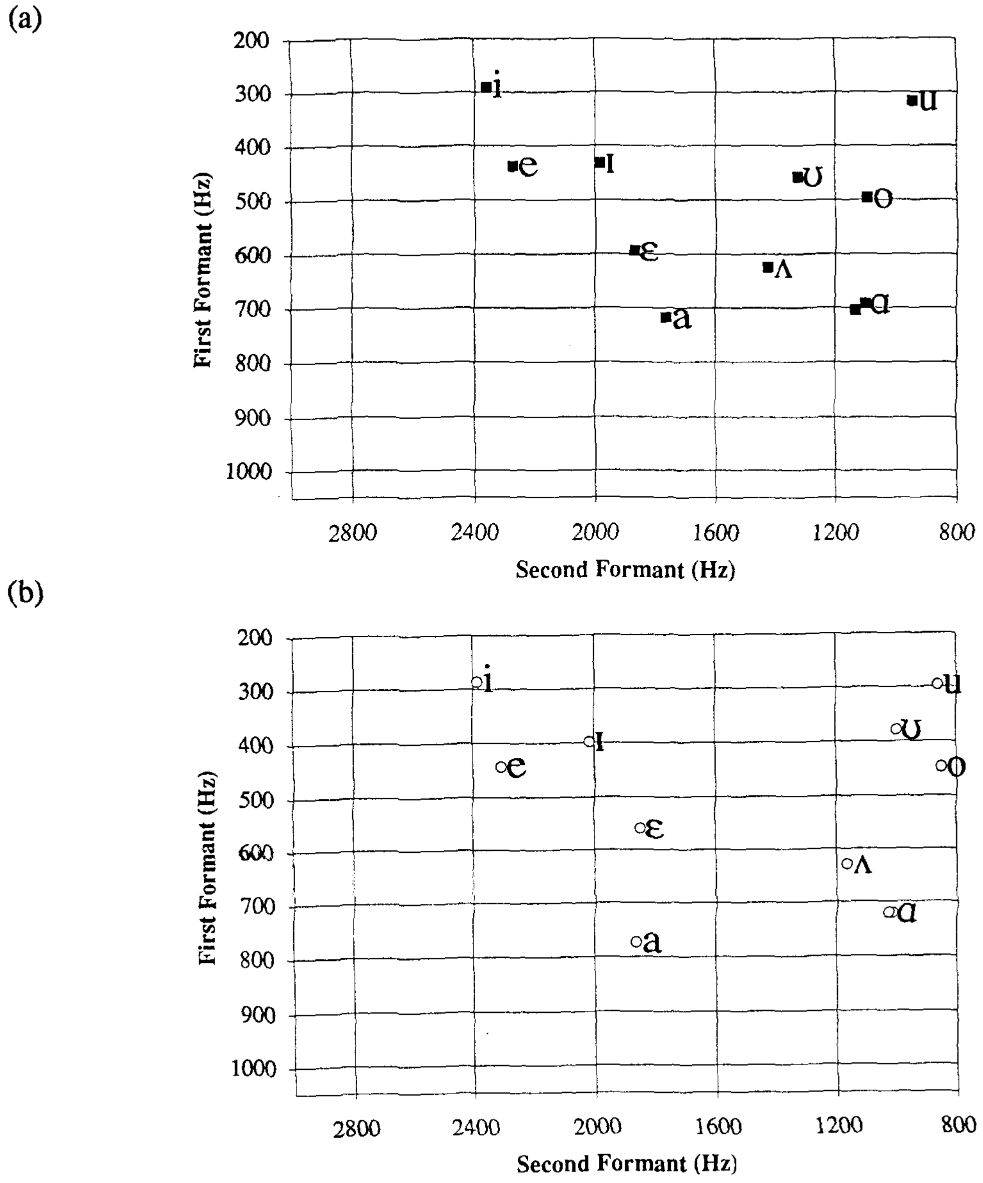


Figure 5 (a) Average measured formant values of English vowels produced by the male Serbo-Croatian speaker. (b) Average method of adjustment results for English vowels obtained from the male Serbo-Croatian speaker.

Two of the listeners in the preliminary study were native speakers of Serbo-Croatian who moved to Los Angeles at different ages. The male speaker was 7 years old when he moved to LA, and both his production and perception data for English were generally comparable to the native English speakers' data (Figure 5). One interesting difference is that his back vowels in both production and *perception have lower F2 values than the native speakers. This corresponds to the* lower F2 found for Serbo-Croatian /u/ and /o/.

The female Serbo-Croatian speaker (who had moved to Los Angeles at the age of 19) showed a much more striking pattern of deviation from the native Southern Californians (Figure 6). Her production vowel space showed some interesting deviations from the native speakers' space. In particular, the vowel in had was more central and lower and the vowel in owed was also lower (compare Figure 6a with Figure 3a). Also, her back vowels had relatively lower values of

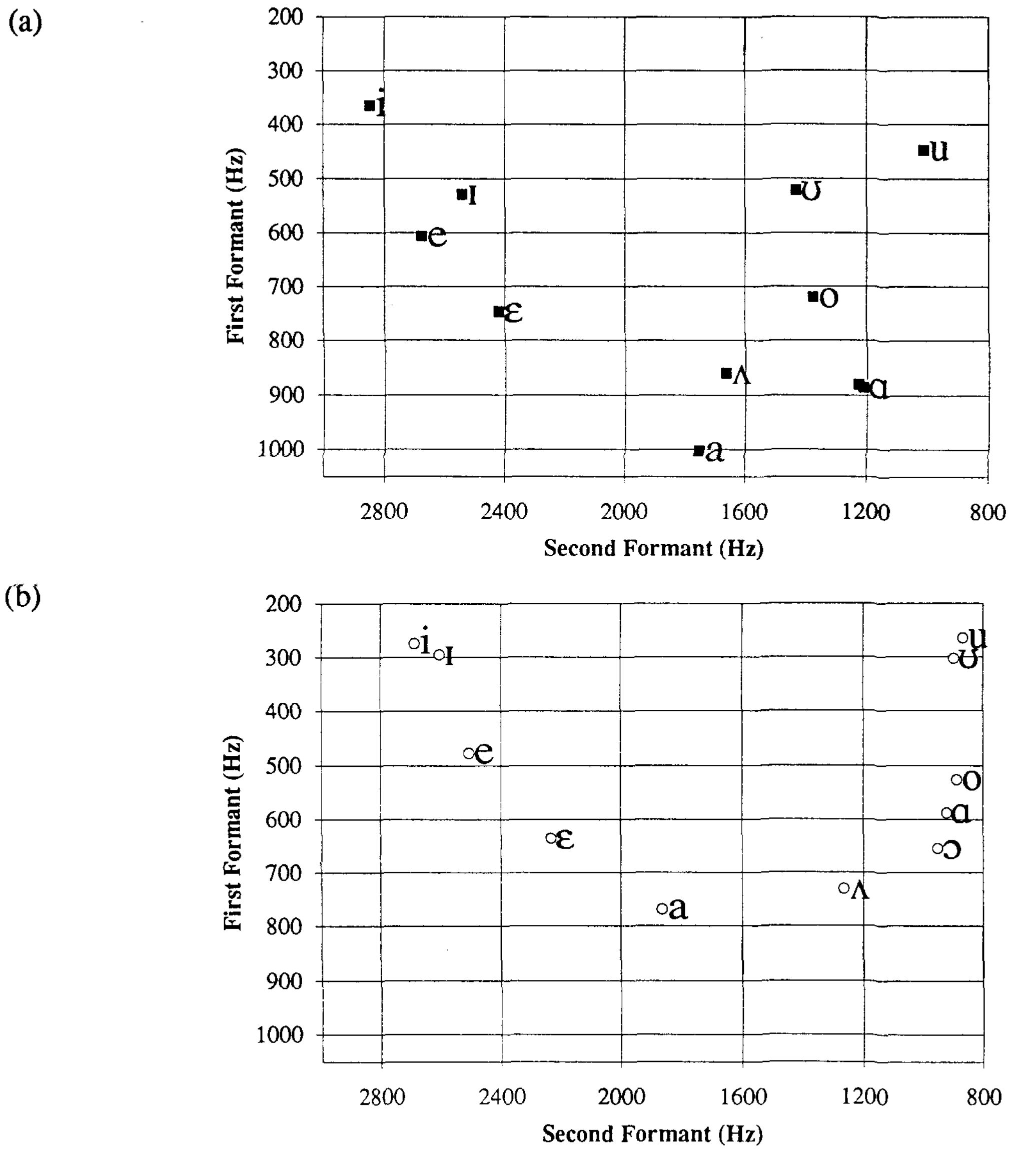


Figure 6 (a) Average measured formant values of English vowels produced by the female Serbo-Croatian speaker. (b) Average method of adjustment results for English (solid circles) and Serbo-Croatian (open circles) vowels obtained from the female Serbo-Croatian speaker.

F2, similar to the pattern seen in the male Serbo-Croatian speaker's production and perception results. However, the most surprizing result from this speaker has to do with her perception vowel space (Figure 6b). She chose peripheral formant values for all of the English vowels and merged the high tense/lax pairs. Figure 6b also shows this speaker's average responses when asked to select vowel sounds for words illustrating the five vowels of Serbo-Croatian. Comparison of her English and Serbo-Croatian perception data indicates that the high vowels of English were associated with the high vowels of Serbo-Croatian but that the rest of the vowels of English (with the possible exception of /o/) were assigned unique values of F1 and F2. Interestingly, however, the entire English vowel space was restricted to the periphery of the acoustic vowel space. This pattern suggests that the listener's perceptual expectations for second language may be influenced globally as well as locally by the first language, and this despite the second language speaker's relative success in producing the second language.

Experiment 1: Instruction set

The preliminary data suggest that the method of adjustment may be a useful tool in the study of dialect differences, cross-linguistic differences and second-language acquisition. In the remainder of this paper we will focus on a methodological puzzle and its significance both for the use of the method of adjustment in studying vowel spaces and for theories of phonetic representation.

The puzzle is illustrated by a comparison of the average perception vowel space from the preliminary experiment (Figure 3b) with the average citation production space from the 8 male speakers in Experiment 1 (Figure 3a). This comparison shows that the vowel space chosen in the method of adjustment was expanded relative to the production vowel space. In other words, listeners' expectations for vowels produced by a male synthetic voice were quite different from the vowels as actually produced by male speakers in citation speech. This is a conundrum if we assume that listeners' perceptual expectations are based on experience.

There were a couple of aspects of the preliminary experiment which made us doubt the validity of this discrepancy between production and perception vowel spaces. The listeners were not phonetically naive; they had completed an undergraduate course in phonetics and thus knew the cardinal vowel system. So, they may have been inclined to select extreme cardinal vowel qualities in the method of adjustment task while more naive speakers might choose formant values more similar to those found in production. Additionally, we suspected that the instructions given to the listeners may have biased them toward extreme vowel qualities. We asked the listeners to find the "best" vowel sound for each word. After the fact we realized that this instruction could have been interpreted to mean, "find the most distinct example of the vowel".

Experiment 1 was designed to investigate these issues by using (1) naive listeners and (2) a careful manipulation of instruction set. One group of listeners was instructed to find the best example of the vowel in each word (the *best* condition). While another group of listeners was instructed to find the vowel sound which most closely matched their own pronunciation of the vowel in each word (the *as you say it* condition).

Subjects. Ten females and eight male university students were recruited through the university newspaper and paid a small sum for their participation. They were monolingual English speakers who reported normal speech and hearing ability and had attended high school in Southern California. The subjects were divided into two groups as described below, with five females and four males in each group.

Materials. As in the preliminary experiment, 330 steady-state isolated vowel stimuli with fifteen possible values of F1 and twenty-two possible values of F2 were synthesized using a software formant synthesizer (Klatt & Klatt, 1990). The formant ranges in Experiment 1 (shown in Figure 1) were larger than those used in the preliminary experiment because the listeners in the preliminary experiment chose formant values which were more extreme than we had anticipated. F1 ranged from 250 Hz to 1000 Hz in increments of 0.42 Bark, while F2 ranged from 800 Hz to 2900 Hz in 0.39 Bark increments.

Procedure. Experimental sessions in Experiment 1 were very much like sessions in the preliminary experiment. However, after the subjects had completed the perception part of the experiment they were asked to read the word list a second time. In this second reading of the words we elicited hyperarticulated or clear-speech versions of the words by saying "what?" or "huh?" after each sentence, prompting the speaker to read each sentence again more clearly. This procedure was explained to the speakers prior to starting the tape recorder. We will call the first reading the *citation* reading and the second the *hyperarticulated* reading. One other procedural difference in Experiment 1 concerned the instructions given to the listeners in the method of adjustment task. We asked one group of listeners (5 female, 4 male) to find the best examples of the vowels and another group of listeners (5 female, 4 male) to find the vowel sound which most closely matched their own pronunciation of the vowel in each word. We will call the first instruction set the *best* condition and the second set the *as you say it* condition.

Finally, the recordings were analysed using CSL (Kay Elemetrics) rather than CSpeech. In analysing these productions we chose measurement points from spectrographic displays (where we had only used waveform displays in the preliminary experiment) and were therefore able to identify an early steady-state portion of the vowel as the point representing the acoustic vowel "target".

Results. The perception results from Experiment 1 are shown in Figure 7a. In separate repeated-measures analyses of variance there were no reliable effects of instruction set on the choices made by the listeners on F1 or F2 for any of the vowels. The most robust difference as a function of instruction set was for the F1 of awed which tended to be greater in the *as you say it* group than in the *best* group (806 Hz versus 758 Hz respectively) but this difference was only marginally reliable (F[1,16]=3.19, p=0.093), no other differences between groups proved to be statistically reliable.

Although the vowel spaces chosen were not affected by instruction set, the ratings given to the synthetic stimuli (shown in Table 2) were. Listeners in the *best* condition were consistently more critical of the stimuli than were the listeners in the *as you say it* condition. The statistical comparison of these conditions was complicated by ceiling effects in the rating data, but the trend is clear. The result is that the instruction set manipulation had an effect on ratings, but it did not have an effect on the formant values chosen in the method of adjustment.

In addition, a comparison (shown in Figure 7b) of the perceptual vowel spaces of naive listeners from Experiment 1 and phonetically trained listeners from the preliminary experiment

Table 2. Rating values given to synthetic vowels in the perception part of Experiment 1. Rating values (on a scale from 1 to 10) were averaged without including vowels rated 1 because the listeners were asked to use 1 to indicate that they had accidently terminated a trial early. Data are presented by vowel and by listening condition.

word	*best*	*as you say it*
heed	8.9	9.7
hid	**8.6**	9.0
aid	**8.7**	9.5
head	**8.9**	9.4
had	**8.6**	9.4
odd	**9.0**	9.7
awed	**8.8**	9.9*
HUD	**8.8**	9.5
owed	**8.1**	9.8**
hood	**8.2**	9.4**
who'd	**8.6**	9.2
average	8.7	9.5

*p < 0.1 **p < 0.05

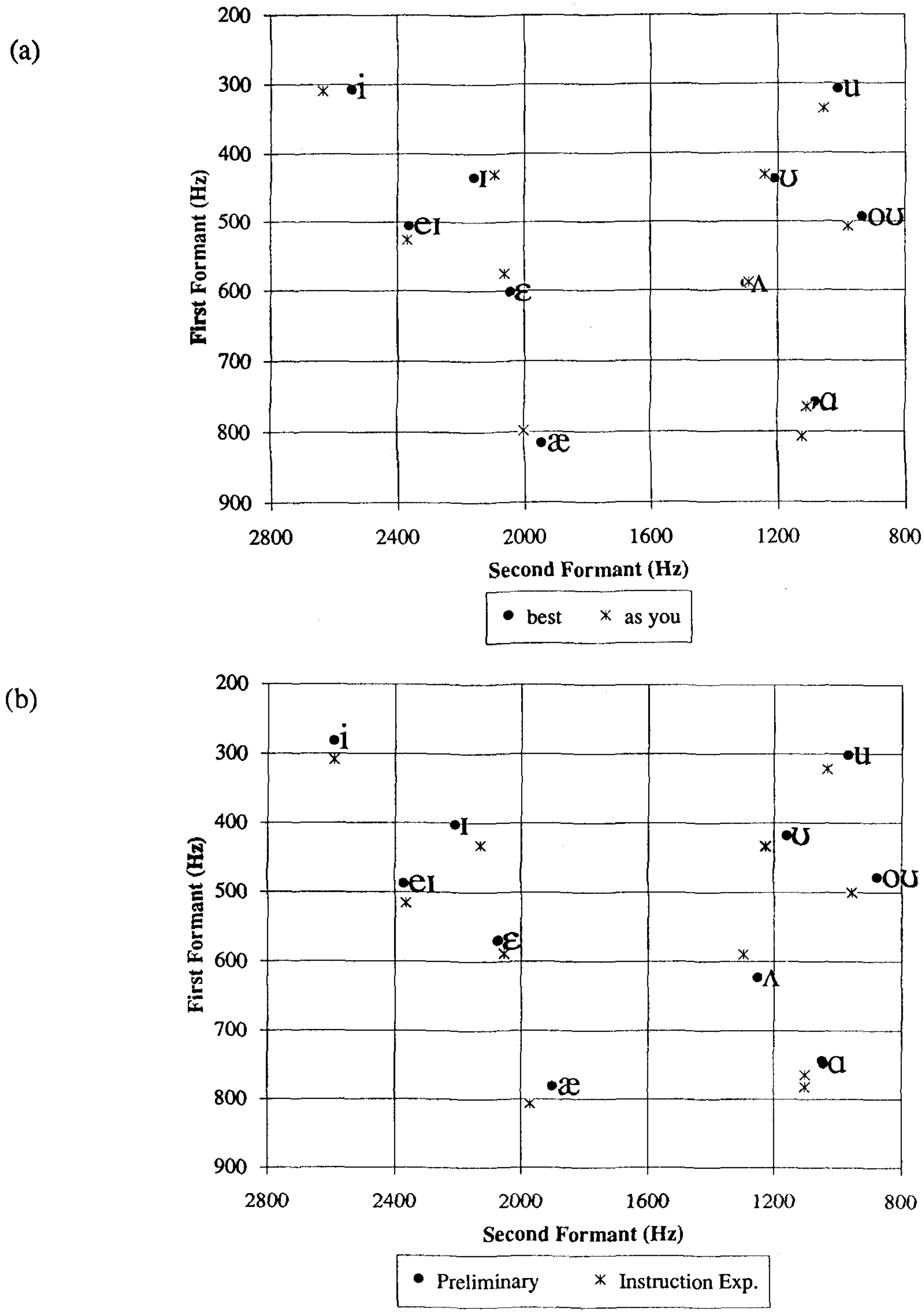


Figure 7 (a) Method of adjustment results of the instruction set experiment. Filled circles are the average values chosen by the listeners in the *best* group, and stars are the average values chosen by the listeners in the *as you say it* group. (b) Comparison of the acoustic vowel spaces chosen by listeners in the preliminary experiment and in Experiment 1. Filled circles are the average responses of the listeners in the preliminary experiment, and stars are the responses (averaged across instruction condition) of the listeners in Experiment 1.

averaged over instruction conditions suggests that phonetic training (at least from the first author) had no effect on the results of the method of adjustment task. It is not valid to attempt a statistical comparison of the data shown in Figure 7b because there were several small changes in the method (particularly the range of possible F1/F2 combinations was expanded in Experiment 1). Still the differences appear to be of the same magnitude as the nonsignificant differences found as a function of instruction set (Figure 7a) and are certainly nothing like the differences between measured values from citation forms and the method of adjustment results (Figure 3a versus Figure 3b).

So, the perceptual vowel space resulting from the method of adjustment task is robust. Speakers of the same dialect give the same answers regardless of instruction set or their previous training in phonetics. This robust pattern of performance raises an interesting question. Namely, what in the experience of the listener underlies the method of adjustment vowel space which is so different from the acoustic vowel space found in normal productions of the same words?

One hypothesis is that the perceptual vowel space which is found in the method of adjustment task reflects hyperarticulated versions of the vowels, rather than the vowel qualities found less carefully produced speech. We will call this the hyperspace hypothesis. With this hypothesis in mind, we asked the speakers to read the word list in a hyperarticulated style. As has been found before (Picheny, Durlach, & Braida, 1986; Moon & Lindblom, 1989), the hyperarticulated versions of the vowels had generally more extreme vowel formants than did the less carefully produced vowels (see Figure 8a). This is just the sort of vowel space expansion that we saw in comparing the perception results with citation readings of the words.

A comparison of the average vowel formants in hyperarticulated productions and the method of adjustment results from Experiment 1 (shown in Figure 8b) suggests that the hyperspace hypothesis is on the right track. Further, when we looked at the hyperarticulated vowel spaces for individual speakers we found that all of the formant values chosen in the perception task were represented in the productions of at least one speaker. Thus, it seems that the listeners' responses in the method of adjustment experiment are not only robust, but are also based upon their experience of very clearly articulated versions of the vowels.

Experiment 2: Intrinsic F0 and duration

One factor which may have had an effect on the method of adjustment results in both the preliminary experiment and in Experiment 1 is that the stimuli were impoverished relative to natural speech. While in English vowels differ in intrinsic pitch, duration and formant trajectories (Peterson & Barney, 1952, Peterson & Lehiste, 1960, Lehiste & Peterson, 1961), the stimuli which we used in the method of adjustment task did not vary along these dimensions. The stimuli were impoverished in this way because we were interested in designing a tool for the cross-linguistic comparison of vowel spaces, and therefore we avoided maipulating pitch, duration and formant trajectories because the redundancies oberserved in English vowels are not cross-linguistic universals of vowel systems (Lehiste, 1970; Keating, 1985). This decision complicates any interpretation of the method of adjustment results because listeners may have attempted to compensate for a loss in the overall distinctiveness of the different vowel qualities (resulting from the absence of redundant cues) by increasing distinctiveness in the spectral domain. Therefore, the hyperspace effect may have been an artifact of the experimental design. We tested this possibility in a second experiment.

In Experiment 2, American English intrinsic vowel F0 and duration were modelled in the synthetic stimuli used in a method of adjustment study. F0 and duration were made to vary as a function of F1 and F2 in a way which is similar to their observed variation in English. Thus, some portion of the redundant information which was missing from the stimuli used in the preliminary experiment and in Experiment 1 was present in these stimuli. If the hyperspace effect occurred in these earlier experiments because of the lack of redundant information in the stimuli, we should find a reduction (but probably not a total elimination) of the effect in Experiment 2.

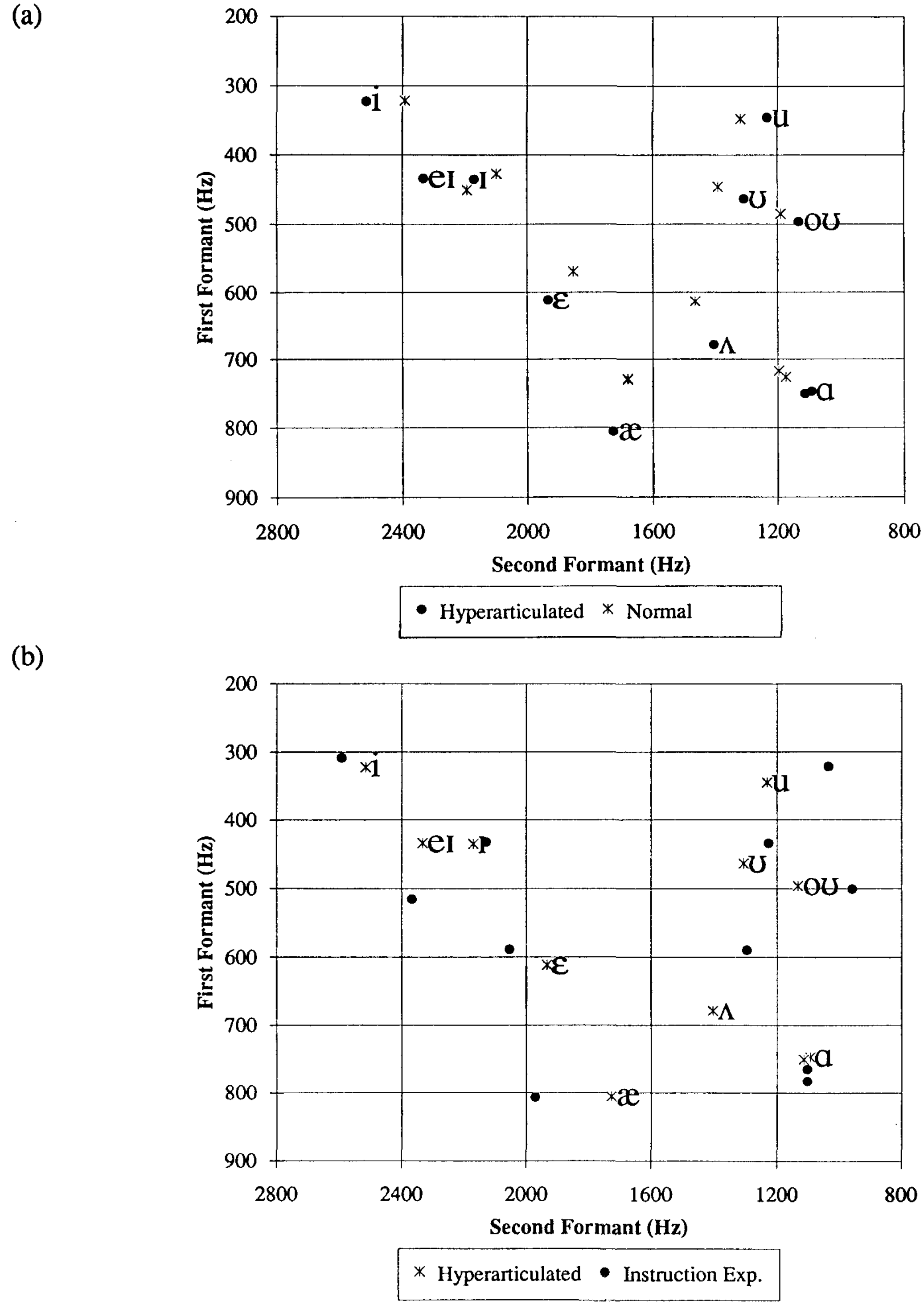


Figure 8 (a) Average measured formant values of the vowels produced by the eight male speakers in Experiment 1. Filled circles are the average values in the hyperarticulated condition and stars are the average values of the vowels in the citation-form condition. (b) Comparison of hyperarticulated productions (8 male speakers from Experiment 1) with method of adjustment results. Stars are the average measured formant values from hyperarticulated versions of the vowels and filled circles are the method of adjustment results (averaged across listener and instruction condition) from Experiment 1.

Subjects. Two male and one female native speakers of Southern Californian English (by the criteria used in Experiment 1) volunteered for the experiment. The listeners reported normal speech and hearing abilities and had completed two introductory phonetics courses. Because we found no differences in performance in the task as a function of phonetic training between the preliminary experiment and Experiment 1, these listeners were taken to be representative of Southern California English.

Materials. As in the earlier experiments, 330 isolated steady-state vowels were synthesized. The formants and bandwidths were the same as those in the stimuli synthesized for Experiment 1, however F0 and duration varied from stimulus to stimulus rather than being fixed as they were in the earlier experiments.

The method used to derive F0 and duration values for the stimuli was analogous to that used in the earlier sets to derive bandwidth values by rule (formulas 1-3 above). Average F0 and formant values for male speakers from Peterson & Barney's (1952) study of American English vowels were entered into a regression analysis in which F0 was predicted by F1 and F2. The resulting regression formula, shown in (4), indicates that F0 is negatively correlated with both F1 and F2. As a result of using this formula to calculate F0 values for the synthetic stimuli, F0 ranged from 110Hz to 142Hz. As in the earlier experiments, F0 was steady over the first half of the vowel and then fell gradually to about 85% of its original value over the last half.

Similarly, average duration measurements from Peterson & Lehiste (1960) and formant measurements from Peterson & Barney (1952) for tense vowels in English were analysed and a regression formula (5) was calculated for duration as a function of F1 and F2. The resulting durations ranged from 210ms to 305ms for the range of F1, F2 combinations in the vowel array. The duration equation is problematic for English because lax vowels have much shorter durations than their tense counterparts, even though their formant values are comparable. Thus, the duration formula only captures vowel variation which is correlated with F1 and F2 variation and the duration differences between tense and lax vowels are not captured by the manipulation. All stimuli had 50ms on- and off-ramps for the amplitude of voicing.

(4) F0 (in Hz) = $153.44 - 0.035*F1 - 0.00275*F2$, $r^2 = 0.939$
(5) Dur (in ms) = $191.754 + 0.121*F1 - 0.00347*F2$, $r^2 = 0.792$

Procedure. Unlike the earlier experiments, no production data were collected in Experiment 2. The method of adjustment task was conducted using the same equipment and software as in the earlier experiments. The listeners were instructed to find vowels which sounded like the ones they produced in the words (the *as you say it* condition of Experiment 1). Each of the eleven English words used in the earlier experiments was presented in random order 7 times.

Results and Discussion. Figure 9 shows the results of Experiment 2 compared with the average results of Experiment 1. This figure indicates that there were only very minor differences between the vowel formants chosen when F0 and duration varied as a function of F1 and F2 and the vowel formants chosen when these two redundant parameters were held constant across the vowel array. These results suggest that the expanded vowel space which was found in the preliminary experiment and in Experiment 1 was not an artifact of the synthetic stimuli. If the absence of redundant information such as intrinsic F0 differences or duration differences between vowels had caused an expansion of the vowel space we would have expected some contraction of the vowel space in this experiment. This result did not occur; the hyperspace effect persisted.

Conclusion

Why do listeners choose a hyperspace in the method of adjustment task? One answer is that the experimental situation biases listeners in this direction. Although we have tried to test some ways in which this might have happened (instruction set, phonetic training, and lack of redundant cues), there may still be some aspect of the stimuli or of task itself which biases listeners toward a hyperspace. For example, if we were to present synthesized versions of whole words

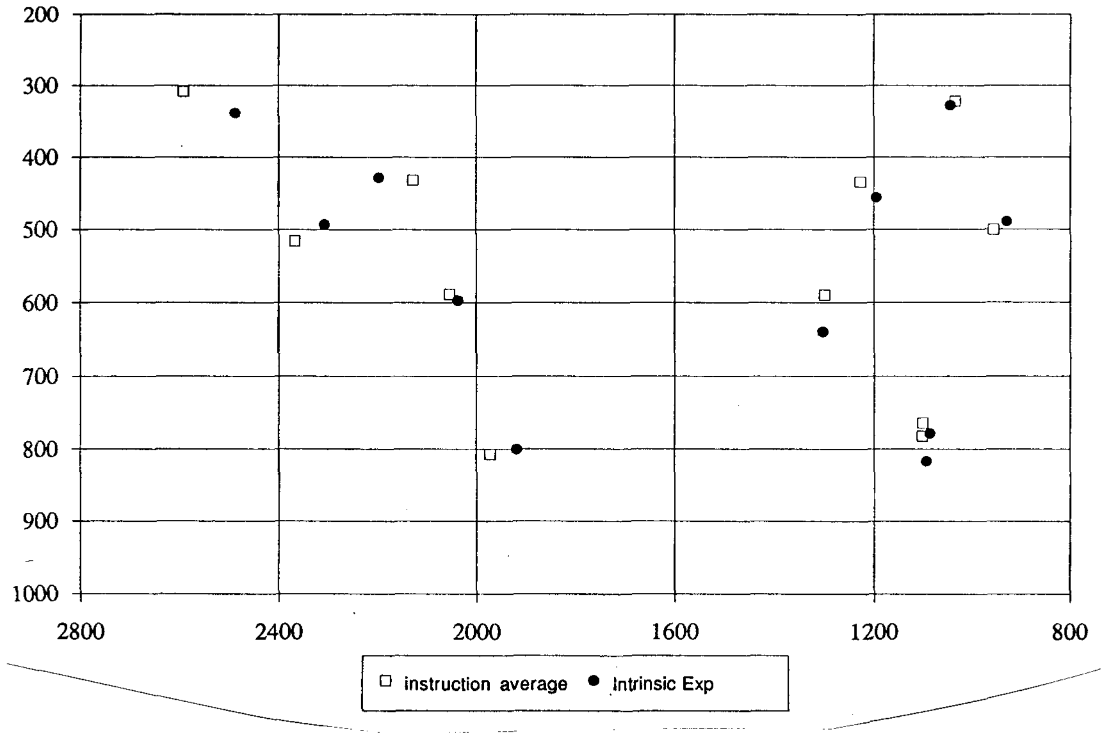


Figure 9 Average method of adjustment results from Experiment 2 (filled circles) and Experiment 1 (open squares).

rather than isolated vowels, listeners might be inclined to choose less extreme vowel qualities (Lindblom & Studdert-Kennedy, 1967). This hypothesis, although reasonable, is probably not right because measured formant values from /hVd/ contexts are very similar to those found in isolated vowel productions (Peterson & Barney, 1952).

One other explanation of the effect is that the formality of the test situation biased the listeners toward a hyperspace in the method of adjustment task. While it is difficult to elicit casual speech in experimental situations, it should be noted that it is also quite difficult to elicit hyperarticulated speech. We found that speakers, when simply instructed to speak clearly, would initially produce quite hyperarticulated speech, but after only a few utterances would revert back to the same style of speaking that they used in the citation reading. This is why we adopted a special procedure to elicit hyperarticulated speech in Experiment 1. So, it is not clear why hyperarticulation would be the *listener's* response to the formality of an experimental situation and not the *speaker's*.

The results reported here suggest that hyperarticulated versions of speech sounds are more basic than less clearly articulated versions of those same sounds (or as one of our speakers put it, the hyperarticulated versions are the *real* sounds). Jakobson & Halle expressed this view when they said, "The slurred fashion of pronunciation is but an abbreviated derivative from the explicit clear-speech form which carries the highest amount of information. ... When analyzing the pattern of phonemes and distinctive features composing them, one must resort to the fullest, optimal code at the command of the given speakers" (1956, p. 6). The hyperspace result can be interpreted as empirical support for this generally assumed, although not explicitly defended, point of view since the notion that speakers utilize hyperarticulated phonetic representations or targets provides a natural rationale for the peripheral vowels selected by the listeners in the method of adjustment.

An alternative account is that phonetic implementation rules are context sensitive. For instance, the feature [+high] might be realized as particular F1 targets which differ depending on prosodic context or the degree of effor the speaker is willing to expend. In this model, the parametric output is determined as a function of both the distinctive feature and various parameters of the performance context. A conceptual difficulty with this type of implementation model is that the different contextually determined realizations of a feature all have equal status as phonetic realizations of that feature (this is also a problem for Keating's, 1988 window model of coarticulation). Thus, a reduced schwa-like version of /i/ is just as good an example of a high vowel as is a hyperarticulated, maximally distinct /i/. As shown in the method of adjustment task, this runs counter to the intuitions of naive listeners.

The theory of phonetic realization must account for the wide range of realizations of the same utterance that a single speaker produces in differing situations. Some of the variation may be introduced by optional categorical phonological rules, but much of the variation is continuous and very low-level in nature, and must surely be the result of phonetic implementation. The details of a phonetic implementation model consistent with our experimental results have not been fully worked out, but an outline is clear. This type of model includes (1) a mapping from categorical representations to parametric representations corresponding to hyperarticulated speech, and (2) a second mapping from maximally distinct parametric representations to reduced forms. The first mapping is what we normally think of as phonetic implementation. It maps distinctive features to phonetic parameters like vocal tract shapes or formant values. By including a second mapping in the model, as suggested by the experimental data, some of the complications that arise in devising schemes of phonetic implementation may be alleviated. In particular, the wide range of realizations of the same utterance that a single speaker can produce does not have to be accounted for by a single mapping from phonological features to phonetic parameters. One noteworthy description of a mapping from hyperarticulated parametric representations to reduced parametric representations is Browman & Goldstein's (1986, 1989) articulatory phonology in which reduction phenomena have been described as gestural overlap, hiding, and blending. There are conceptual difficulties in interpreting a gestural model as the second mapping in a two stage model of phonetic implementation (clearly this is not what Browman and Goldstein have in mind) but there are some compatibilities between a gestural model and the requirements of the mapping because the gestures in articulatory phonology specify articulatorily extreme targets which become reduced in normal speech.

Acknowledgments

Special thanks to Doug Whalen who suggested Experiment 2, Bob Port who offered encouragement early on, and Peter Ladefoged whose interest and support inspired us.

References

Behne, D.M. (1989) A comparison of the first and second formants of vowels common to English and French. *Research on Speech Perception Progress Report no. 15*, 269-282. Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Browman, C.P. & Goldstein, L. (1986) Towards an articulatory phonology. *Phonology*, **3**, 219-252.

Browman, C.P. & Goldstein, L. (1989) Articulatory gestures as phonological units. *Phonology*, **6**, 201-231.

Browman, C.P. & Goldstein, L. (1990) Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, **18**, 299-320.

Disner, S.F. (1980) Evaluation of vowel normalization procedures. *J. Acoust. Soc. Am*,67, 253-261.

Disner, S.F. (1986) On describing vowel quality. In (Eds.) *Experimental phonology*, ed. by J.J. Ohala & J.J. Jaeger, 69-79. Orlando: Academic Press.

Flanagan, J. (1957) Estimates of the maximum precision necessary in quantizing certain 'dimensions' of vowel sounds. *J. Acoust. Soc. Am*,**29**, 533-534.

Ganong, W.F. & Zatorre, R.J. (1980) Measuring phoneme boundaries four ways. *J. Acoust. Soc. Am*, **68**, 431-439.

Gerstman, L.H. (1968) Classification of self-normalized vowels. *IEEE Trans. Audio*

*Electroacoust.***AU-16**, 78-80.

Harshman, R. (1970) Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis. *UCLA Working Papers in Phonetics*, **16**.

Jakobson, R. & Halle, M. (1956) *Fundamentals of Language*. 'S-Gravenhage: Mouton & Co.

Joos, M. (1948) *Acoustic Phonetics*. Linguistic Society of America Language Monograph No. 23 (Baltimore: Waverly Press).

Johnson, K. (1989) On the perceptual representation of vowel categories. *Research on Speech Perception Progress Report no. 15,* 343-58. Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Johnson, K. & Teheranizadeh, H. (1992) Facilities for speech perception research at the UCLA phonetics lab. *UCLA Working Papers in Phonetics,* **81**.

Keating, P.A. (1985) Universal phonetics and the organization of grammars. *Phonetic linguistics: essays in honor of Peter Ladefoged*, ed. by V. Fromkin, 115-32. Orlando: Academic Press.

Keating, P.A. (1988) The window model of coarticulation: articulatory evidence. *UCLA Working Papers in Phonetics,* **69**, 3-29.

Klatt, D. (1980) Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.* **67**, 971-995.

Klatt, D. & Klatt, L. (1990) Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am,* **87**, 820-857.

Ladefoged, P. & Broadbent, D. (1957) Information conveyed by vowels. *J. Acoust. Soc. Am,* **29**, 98-104.

Lehiste, I. (1970) *Suprasegmentals.* Cambridge, MIT Press.

Lehiste, I. & Peterson, G.E. (1961) Transitions, glides and diphthongs. *J. Acoust. Soc. Am,* **33**, 268-277.

Lindblom, B. (1990) Explaining phonetic variation: A sketch of the H&H theory. *Speech production and speech modelling*, ed. by W.J. Hardcastle & A. Marchal, 403-439. Dordrecht: Kluwer Academic.

Lindblom, B. & Studdert-Kennedy, M. (1967) On the role of formant transitions in vowel recognition. *J. Acoust. Soc. Am,* **42**, 830-843.

Lobanov, B.M. (1971) Classification of Russian vowels spoken by different speakers. *J. Acoust. Soc. Am,* **49**, 606-608.

Miller, J.D. (1989) Auditory-perceptual interpretation of the vowel. *J. Acoust. Soc. Am,* **85**, 2114-2134.

Moon, S.J. & Lindblom, B. (1989) Formant undershoot in clear and citation-form speech: A second progress report. *STL-QPSR*, **1**, 121-123.

Nearey, T.M. (1977) Phonetic feature systems for vowels. PhD Dissertation, University of Connecticut, Storrs, CT.

Nearey, T. (1989) Static, dynamic and relational properties in vowel perception. *J. Acoust. Soc. Am,* **85**, 2088-2113.

Nooteboom, S.G. (1973) The perceptual reality of some prosodic durations. *J. Phon,* **1**, 25-46.

Peterson, G.E. & Barney, H.L. (1952) Control methods used in a study of vowels. *J. Acoust. Soc. Am,* **24**, 175-184.

Peterson, G.E. & Lehiste, I. (1960) Duration of syllable nuclei in English. *J. Acoust. Soc. Am,* **32**, 693-703.

Picheny, M.A., Durlach, N.I. & Braida, L.D. (1986) Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *J. Speech & Hearing Res,* **29**, 434-446.

Repp, B.H. & Liberman, A.M. (1987) Phonetic category boundaries are flexible. *Categorical Perception,* ed by S.N. Harnad. New York: Cambridge University Press.

Samuel, A. (1982) Phonetic prototypes. *Perc. & Psychophys,* **31**, 307-314.

Syrdal, A. & Gopal, H. (1986) A perceptual model of vowel recognition based on the auditory respresentations of American English vowels. *J. Acoust. Soc. Am,* **79**, 1086-1100.

Trading relations and informational models
Richard Wright

1.0 Introduction

For the past several decades there has been a lively debate about the nature of the perceptual and cognitive processes involved in translating the acoustic speech signal into discrete linguistic units such as phonemes. The central theme of this debate has been summed up in a single question: "Is speech special?". On one side of the debate there is the hypothesis, proposed by proponents of the motor theory (e.g. Liberman & Mattingly, 1985, 1989), that speech perception involves special mechanisms that interpret the acoustic signal using articulatory information. On the other side of the debate there is the hypothesis that, at very low levels at least, speech is processed in the way that all other auditory input is, but that there may be bias effects on perception (e.g. Pastore, 1981). An alternative response to the question is seen in the direct realist approach (Fowler, 1986; Fowler & Rosenblum, 1991) which, like the motor theory, assumes an event based approach, but, like the auditory theories, asserts that this approach is used in all forms of perception not just in perceiving speech. At the same time, informational theories, such as the Fuzzy Logic Model of Perception (FLMP) (Derr & Massaro, 1980; Massaro, 1987), have seen a recent resurgence in popularity.

The aim of this thesis is to review some of the phenomena that proponents of the motor theory have cited as supporting evidence, to review alternative interpretations of these phenomena, and to describe a set of experiments designed to test the hypothesis that the auditory signal is interpreted gesturally by a phonetic perception module. Specifically, the prediction that only features that are gesturally significant in the signal should affect the listener's perception of speech is tested using the trading relations phenomenon.

1.1 The motor theory

Over the years the motor theory has undergone many revisions. The version investigated here is the one proposed by Liberman and Mattingly (1985, 1989) and by Whalen and Liberman (1987) in which there is a "biologically distinct module" which produces percepts that are immediately phonetic, bypassing the standard auditory pathway used for non-speech percepts. The phonetically relevant parts of the signal are preempted by the "phonetic" module, and the portion of the signal that is not phonetically relevant is left for general auditory processing. The module intercepts the speech part of the signal by picking out the parts of the signal that correspond to a limited class of sounds that is defined by the acoustic consequences of phonetically relevant vocal tract configurations, or gestures. An important aspect of the present version of the motor theory is that intended articulatory gestures are the basis of speech perception. An intended gesture is abstract and invariant and has features that specify it uniquely, whereas its physical realization can vary greatly, although systematically, due to factors such as speaking rate, and coarticulation. A second aspect of the current version of the theory is that the innate "phonetic" module perceives the intended gestures directly (as per Gibson, 1966) without the intervention of higher level auditory processing.

1.2 Other theories of speech perception

When the motor theory was initially proposed by Liberman, Cooper, Shankweiler & Studdert-Kennedy (1967) it provided an attractive alternative to other theories because it offered a way to account for aspects of the speech signal that were seen as problematic at the time. There was a general effort to find invariant and linear correlates of segments which was thwarted by variation and overlap within the speech signal. The overlapping nature of the acoustic signal is due to the continuous and interactive nature of the vocal tract configurations in speech production. An often cited example is the variation of formant transitions of a given consonant due to the vowel context in which it is found (Delattre, Liberman & Cooper, 1955; Liberman, Delattre & Cooper, 1954). For example, while a falling F2 transition can be used as a cue to the alveolar place of articulation of a stop before a low back vowel, a rising F2 transition occurs when an alveolar consonant precedes a high front vowel. An alternative approach that has been proposed in various informational models in various forms from Klatt's (1979) LAFFS model to Massaro and Oden's (1980) FLMP (see also Derr & Massaro, 1980; Oden & Massaro, 1978; Massaro, 1972, 1987,1989; Massaro & Cohen, 1976, 1977), is that of matching elements of the acoustic signal to a stored representation using some sort of best fit algorithm.

Rather than seeing variation within the signal as something to overcome, informational models allow variation to be broken into two groups: information and noise. Information comes from lawful variation such as that caused by coarticulation (e.g. Delattre, Liberman & Cooper, 1954; Liberman, Delattre & Cooper, 1955) and speaker gender or index (Ladefoged and Broadbent, 1957). Noise is nonrelevant or unpredictable sources of variance within the signal such as environmental masking or a talker's mouthfull of food. Noise may cause distortion of some dimension of the speech signal thereby reducing the sum total of information available to the perceiver, or it may be filtered out of the signal as irrelevant. The speech signal is replete with redundant cues (information) because it is the richness of the signal that allows it to be transmitted and perceived in less than optimal conditions. The richness is derived not only from the varied and related acoustic cues, but from visual, phonological, prosodic, syntactic, semantic and discourse information.

The FLMP as defined by Massaro (1989) uses fuzzy logic, "a continuously valued logic that represents the truth of propositions in terms of truth values that range between zero (false) to one (true)" [p.724]. Goodness of match to a subjectively derived prototype description in memory, arrived at through experience with the language of the listener, is the criterion for perception of a particular set of features as a particular perceptual unit such as the phoneme. The input is the acoustic signal which undergoes an acoustic analysis by the peripheral auditory system. Continuous values are assigned to the various acoustic[1] features of the signal based on the perceptual system's certainty that the feature appears in the signal; then the valued features are integrated and matched against the possible candidate prototypes. Because best fit algorithms are used, an absolute match is not needed for the process to achieve results.

Connectionist models, such as the TRACE model (Elman & McClelland 1984, 1986; McClelland & Ellman 1986), are similar to the FLMP in that they also rely on continuous rather than discrete representations and that by using activation levels they also allow for varying degrees of support for perceptual hypotheses. Connectionist models also allow for the evaluation and integration of multiple sources of input and rely on a pattern matching scheme with a best fit algorithm. Where the connectionist models and the FLMP differ is in the degree to which top down influences can affect low level processes. As Massaro (1989) points out, connectionist models that have two way interactions are too powerful; "they are capable of predicting not only observed results, but results that do not occur" [p. 755]. As a result, this study will use informational models rather than connectionist models . However, where a phenomenon is analyzed using the FLMP, a connectionist model could just as easily have been used since the degree to which top-down influences affect auditory perception are not being considered in the present study.

[1] The model also allows visual cues but since the scope of this study is auditory the discussion is limited to acoustic features.

2.0 Evidence in favor of the motor theory

Many of the findings that were initially interpreted as evidence in favor of the motor theory (e.g. Wood, 1975) have proven to be less conclusive than was originally thought. An example of such a finding is the early interpretation of categorical perception in speech, dubbed "phoneme boundary effect" (Studdert-Kennedy et al, 1970), as evidence for a "phonetic mode" of perception. It was found that on a continuum between two stops, such as a voiced-voiceless contrast, a listeners sensitivity to small changes in the acoustic signal was best at the crossover point between the two categories and was poorest within either category. Moreover, sensitivity in discrimination tasks was predictable from identification tasks. However, it was subsequently shown that categorical perception also occurred when a variety of non-speech continua were used (e.g. Pisoni, 1977; Cutting and Rosner, 1974). Additionally, it was found that categorical discrimination occurred when chinchillas were exposed to synthesized speech stimuli (Kuhl and Miller, 1978). In the end, categorical perception, as a perceptual mechanism rather than the output of cognitive processes, has been abandoned by both sides of the debate (see Massaro, 1976, 1989; Studdert-Kennedy, 1989, for a discussion) because of findings that within category discrimination occurs (e.g. Barklay, 1972; Pisoni, 1973). Nevertheless, there is a continuing debate over the status of categorical versus continuous perception in speech perception in general (e.g.. Macmillan *et al*, 1987; Massaro, 1989; Schouten & van Hessen, 1992).

2.1 Current evidence

Of the remaining phenomena that are interpreted as support for a phonetic module, the two that are cited most often are duplex perception and trading relations. The phenomenon of duplex perception (Rand, 1974) has frequently been cited as strong evidence for a phonetic module which is separate from and preemptive of more general auditory perception (Liberman, 1982; Liberman & Mattingly, 1985, 1989; Repp, 1982; Studdert-Kennedy, 1982; Whalen & Liberman, 1987). In a typical duplex perception experiment, the listener is presented a portion of a single formant transition in one ear, F2 or F3, while the remainder of the syllable is simultaneously presented in the other ear. The result is the percept of a nonspeech "chirp", the isolated transition, in one ear while a syllable is perceived in the other ear. In the view of Whalen and Liberman (1987), the phonetic module preempts the general auditory processes, uses the phonetically relevant parts of the signal, then passes the non-speech portion, the single formant transition, on to the general auditory processing system.

Another explanation of this phenomenon is put forward by Bregman (1978, 1987, 1990) in his theory of auditory stream segregation; presenting the formant transition in one ear while the rest of the syllable is presented in the other spatially segregates the formant transition from the rest of the signal. At the same time, however, the formant transition onset and the onset of the rest of the syllable are temporally aligned suggesting that the two are the same perceptual object. As a result, the transition and the syllable "base" are integrated at the onset, but segregate as the unitary percept breaks down.

Another alternative explanation comes from the direct realist approach as proposed by Fowler (1986) and Fowler and Rosenblum (1990). This theory of perception, which extends Gibson's (1966) theory of visual perception to the auditory domain, claims that perception is event based and that the medium of transmission is only the carrier and is not involved in the process of perception itself. Unlike the motor theory the direct realist approach does not see speech as special; it is perceived directly without the intervention of a module just as all other auditory stimuli are perceived. Accordingly, duplex perception should be found for non-speech events as well. To test this hypothesis, Fowler and Rosenblum (1990) performed an experiment in which the sound of a metal door slamming was recorded and then low pass and a high pass filtered versions were made. The high frequency noise was presented to one ear and the low frequency noise, which sounded like a wooden door, was presented to the other. When the high frequency noise was played at a slightly higher amplitude than the low frequency noise, listeners heard a metal door plus the high frequency noise. Thus, the strength of duplex perception as evidence in favor of a low level phonetic module has been brought into question, but the claim that speech is perceived directly still remains intact.

However, the trading relations phenomenon has remained relatively unchallenged. It is well known that any phonetic contrast can be signaled by a number of acoustic cues. It has been found that a change in one cue can be offset by change in another (e.g. Hoffman, 1958; Summerfield & Haggard, 1977; Fitch, Hawles, Liberman, Erickson & Liberman, 1980; Best, Morrongiello & Robson, 1981). For example, Fitch et al (1980) conducted an experiment using a fricative-stop-vowel continuum (sa-sta) in which changes in a spectral cue, the F1 transition, were found to be compensated for by changes in a temporal cue, the duration of the prevocalic silence. Listeners heard /sta/ with either a longer silence and a higher F1 onset, or a shorter silence and a lower F1 onset. The perceptual offset of one cue in the speech signal by another has come to be referred to as a trading relation. The cues that have been shown to trade off in this manner are directly relevant to articulatory gestures; this is referred to as phonetic relevance. In the Fitch et al example, silence is a cue to stop occlusion and the F1 transition is a cue to stop release. What is more, in an experiment that used a sine wave analog of speech, Best et al (1981) demonstrated that when a stimulus continuum is perceived as speech, trading relations are found in both the identification and discrimination functions, but when the identical stimulus is perceived as non-speech, trading relations disappear (but see Johnson & Ralston, 1990). It has frequently been argued (Liberman & Studdert-Kennedy, 1977; Fitch, Hawles, Erickson, & Liberman, 1980; Best, Morrongiello, & Robson, 1981; Repp, 1982, 1983a; Liberman & Mattingly, 1985) that since the cues that participate in trading relations are those that are phonetically relevant, hence direct correlates of gestures, trading relations are evidence for a special phonetic module that interprets the signal in terms of intended gestures. Liberman and Mattingly (1985) state:

"As for the perceptual equivalence among diverse cues that is shown by the trading relations, explaining that on auditory grounds requires ad hoc assumptions. But if, as the motor theory would have it, the gesture is the distal object of perception, we should not wonder that the several sources of information about it are perceptually equivalent, for they are products of the same linguistically significant gesture." [p. 12]

While such experiments such as those by Best et al (1981) demonstrate that the listener reacts differently to a stimulus that is perceived as speech than when it is perceived as nonspeech, they fail to demonstrate that the difference is due to the intervention of a phonetic module rather than to some other more general perceptual process.

It should be pointed out that in such experiments it is the practice to investigate only cues that are phonetically relevant. Thus it is no surprise that the results show that gesturally relevant cues participate in trading relations. The results would be more conclusive if it could be established that trading relations occur only between features of the signal that are gesturally relevant, and do not occur between a feature that is gesturally irrelevant and one that is relevant. Such an experiment is described below.

2.2 Alternative interpretations

Two alternative interpretations of the trading relation phenomenon have been suggested: psychoacoustic and informational. In psychoacoustic interpretations such as the one put forward by Pastore (1981), trading relations can be seen as resulting from the interaction of several complex auditory phenomena. One is the effect of forward masking, for example from the 's' frication in the 'sa-sta' continuum, that masks cues at the onset of voicing such as formant transitions or release bursts. Equally the backward and upward masking of the low level energy in the onset of voicing could effectively mask transitional cues in the 's' frication. The silent period between the end of the frication and the onset of voicing allows for maximum persistence of the perceptual effects of the preceding stimulus, and segregates the preceding and following stimuli. The silent period can act as a temporal marker that helps to integrate the various aspects of the signal into a unitary percept. Delgutte (1984) has shown that inserting a 100 ms period of silence between an 's'-like fricative noise and a vowel created a dramatic increase in the response discharge rate in the auditory nerve of the cat at vowel onset. It has also been shown that there is a high discharge rate followed by a gradual decay in reaction associated with the sudden onset of vowel like noise. Thus one explanation of trading relations is that an increase in the temporal cue, silence, allows the spectral cues in the voicing onset to become salient because masking no longer distorts them. A long period of silence also contributes to the percept of a stop because it allows for maximal recovery, thus higher excitation levels, of the auditory nerve. The sudden onset of voicing causes an upward spread of masking creating the percept of a broad frequency noise typical of some release bursts. Thus when the temporal cue is short, the formant transitions are masked by the spectral features of the preceding consonant. When the duration of the silence is long enough, the formant transitions escape masking and become perceptually salient. A formant transition that is a cue to the presence of a particular stop will result in a stop percept, but a transition that is not appropriate for the presence of a particular stop may still result in the percept that there is no stop. When the silent period is long enough for the auditory nerve to recover from the excitation of the 's' noise, there may be the percept of a release burst, which may lead to the percept that there is a stop regardless of the lack of appropriate formant transitions.

As Pastore has pointed out, however, while psychoacoustic studies have shed light on some acoustic correlates of segments, these explanations alone cannot account for trading relations but require integration with cognitive processes. Early psychoacoustic explanations such as that of Pastore (1981) were based on categorical perception. In fact psychoacoustic studies of category effects (e.g. Pisoni & Lazarus, 1974; Samuel, 1977, Massaro & Oden, 1980) indicate that the category like effects that are seen in speech perception are not due to general processes of audition, but rather to processing which occurs after the peripheral auditory stage of perception. Because of such findings, attempts to define speech categories in terms of a raw physical response to a stimulus have become less popular in recent years.

There are two other reasons that a purely psychoacoustic explanation of trading relations is undesirable. First of all, while average crossover points in continua appear fairly uniform for both identification and discrimination, and while there tends to be within subject consistency, there is dramatic between subject variability as to where the crossover point occurs (e.g. Liberman et al, 1980; Best et al, 1981). Given that the signal was identical for all subjects, one would expect that the response would be more uniform if a trading relations are to be explained on purely psychophysical grounds. Secondly, a strong argument against a purely psychophysical explanation are the results of experiments, such as the one conducted by Best et al (1981), that have demonstrated that trading relations occur when a stimulus continuum is perceived as speech, but fail to occur when the same continuum is not perceived as speech. The failure of psychophysical accounts is the primary reason that the trading relations phenomenon remains an argument for the motor theory.

A second alternative account of trading relations is the informational approach. In informational models there are typically several stages of processing. This hierarchical approach to processing was first proposed by Studdert-Kennedy (1974) who had four stages of processing: auditory, phonetic, phonological, syntactic-semantic. The FLMP, introduced on p. 3 above, is an attractive model because it allows for perception under conditions of degradation of the signal and for multiple sources of input, ie visual. In the FLMP the first stage is general auditory processing in which the various acoustic features of the auditory signal are evaluated using continuous values ranging from 0 to 1 depending on the degree of certainty of the listener that a given feature is present in the signal. At this stage the processing of the signal is purely auditory or psychoacoustic but coupled with some sort of feature detection in order to evaluate the features. The various evaluated features are then integrated. It is the integrational stage that allows the model to account for the interaction of multiple acoustic cues to a particular segment as well as the integration of multiple cues and features from other peripheral sources of input such as vision (McGurk & MacDonald, 1976). The integrated features are then matched against a subjectively derived prototype using a best fit paradigm to arrive at a classification.

In many experiments on trading relations there is one cue that is described as a spectral cue and one that is temporal. The spectral cue is usually present or absent, for example the presence or absence of a formant transition that is appropriate for a particular stop, and the temporal cue is usually present as varying silence duration. In the informational approach, the longer the duration of the silent cue, the higher the value that is assigned to it. In this approach, the higher value for the longer silent duration can offset the absence of other cues to the presence of a particular stop because it raises the best fit score for the signal. When the spectral cue is present, less silence is needed to achieve an equivalently high best fit score. It is the continuous nature of the values assigned to the features that allows them to interact equivalently. The obvious argument against this approach is that it is too unconstrained. It predicts that aspects of the signal that are not phonetically relevant, such as amplitude differences, should be able to raise the best fit score and trigger trading relations. This prediction is at odds with that of the motor theory. While such nonlinguistic factors have been shown to interact with other nonlinguistic factors, they have never been demonstrated to participate in speech trading relations (Repp, 1983a).

2.3 Previous experiments

While experiments on trading relations have been cited as evidence for a phonetic module that uses articulatory information to interpret the signal, experimenters have consistently used stimuli that are based on cues that are seen as significant to the gesture as a unit. This is, of course, sensible practice on one level, but when the results are cited as evidence that the signal is perceived in terms of gestural intent, the experimenter has failed to rigorously test the model. For example, in an experiment conducted by Fitch et al (1980) that looked at a fricative-vowel to fricative-stop-vowel (sa-sta) continuum, the F1 frequency at the vowel onset was manipulated so that in one set of stimuli, it had the steeper increase in frequency that is associated with the presence of a voiceless stop, and in the other set of stimuli, it had the shallower increase in frequency associated with the presence of a voiceless fricative. When the F1 transition was appropriate for the presence of a stop, a shorter silent duration was needed in order for the stimulus to be perceived as having a stop than when the F1 transition was appropriate for the fricative.

By choosing the stimuli so that they represent acoustic correlates of either one or the other gestural category being examined, experimenters have skirted one of the most important questions raised by the phenomenon of trading relations: can an acoustic stimulus that is not significant to a particular gesture be found to interact with one that is? If this sort of trading relation can be established, then using the results of trading relation experiments as evidence of a phonetic module that interprets the speech signal using gestures will be brought into question. If the phonetic module interprets the acoustic signal in terms of intended gestures, then changes that are acoustically significant but gesturally insignificant should not participate in trading relations. If, on the other hand, such changes do participate in trading relations, then there is reason to doubt the efficacy of a phonetic module, especially in light of models such as the FLMP which would allow such trade offs. With this goal in mind a pair of experiments was designed.

In an experiment that was novel for its use of unambiguous, real speech rather than synthesized speech, Repp (1983a) demonstrated a trading relation between the presence *vs* the absence of a stop burst and closure duration. For this experiment he used a 'slit-split' continuum in which the prevocalic period of silence, representing the stop closure, and the presence *vs* absence of the 'p' release burst were the 'cues' being studied. The presence of the burst was the spectral cue to the labial release gesture, and the absence of the burst indicated the absence of the labial gesture.

3.0 New experiments

The present experiments differ from that of Repp (1983) described above in one key way: the spectral cue used was high *vs* low amplitude of the burst.[2] This cue cannot be considered significant to the distinction of the gestures involved in production of the stimuli; while the burst amplitude is a gestural cue to the voicing of a stop, it should not cue the absence vs presence of a stop. Rather, in terms of distal events, as long as it is audible, the presence of the burst should be a cue to a labial release gesture, regardless of whether it is more or less audible. The amplitude difference should not affect the length of closure silence needed to achieve the percept of a 'p'. Thus, the motor theory predicts that no trading relation should occur. However, given a model that assigns continuous values to features based on their degree of presence in the signal, it is expected that the decreased saliency of a cue within the signal would lead to a poorer fit to the stored representation which could be compensated for along another dimension. Increased amplitude should increase the value assigned to the burst cue because listeners will have a higher degree of certainty of its presence. The higher relative value will result in a better fit which could itself offset a shorter (lower value) silent cue. Thus, the FLMP predicts that a trading relation will occur.

3.1 Experiment 1: Variable AX Discrimination Task

The first experiment used a variable AX paradigm to establish that the sensitivity to the stimuli is appropriate for that normally found for speech stimuli and to gather reaction time data on the responses. This is of issue here because it has been claimed (Repp 1983a, Best 1981) that variations in the signal that produce non-phonetic variation have sensitivity scores that are monotonic across the continuum rather than having a peak at the crossover point between the two categories. A two interval AX task was used because it is a more sensitive measure for deriving auditory discriminability than other paradigms such as the ABX or oddity discrimination tasks which put a heavier demand on auditory short term memory (Pisoni, 1973). A variable rather than a standard AX design was used to avoid anchoring effects at the edges of the stimulus continuum such as those seen in Repp (1983a). The discrimination experiment preceded the identification experiment to avoid biasing the subjects. The second experiment, introduced in 3.2, used a forced choice identification task to determine whether or not a trading relation could be seen between the amplitude of the 'p' burst and the duration of the pre-burst silence. A trading relation typically displays a shifting of the crossover point in identification functions due to the manipulation of the cues.

[2] An experiment using burst amplitude as a spectral cue in a voiced-voiceless stop continuum was conducted by Repp (1983); however, his findings indicate that while a trading relation could be established, it occurred between nonlinguistic aspects of the signal. This conclusion was based on the results of the discrimination tasks which did not show the peak in sensitivity peak at the crossover typical of speech.

3.1.1 Methods

Subjects:

Seven paid volunteers participated in the experiment. Of the seven, three were female and four were male. All were UCLA students.

Stimuli:

The original stimulus consisted of the word "split" recorded by a female speaker[3]. The recording was made in a single walled sound attenuated chamber at the UCLA Phonetics Laboratory using a standard analog cassette recorder. The recording was then digitized at 10 kHz using the Kay Computerized Speech Laboratory. After the digitizing process all of the consonants were clear.

Initial duration measurements of the 'split' recording were taken from the waveform and a time aligned spectrogram: the pre-closure 's' frication was 101.5 ms, the 'p' closure was 80.9 ms, the 'p' release burst was 13.3 ms, the voiced 'li' section, as measured from the first glottal pulse, was 188.3 ms, the 't' closure was 78.7 ms, and the release burst and aspiration was 66.9 ms.

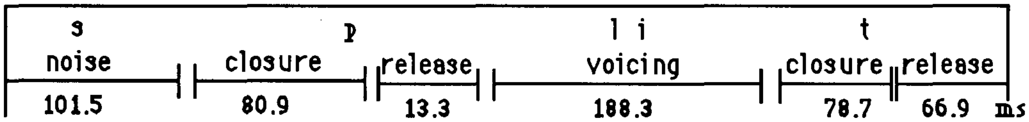


Figure 1

Using digital editing, the 's' frication was separated from the rest of the waveform and the final 13 ms of the 's' frication was removed in order to dilute the spectral cue to labial closure within the 's'. Then, two 'blit' stimuli were made by 1) doubling the gain of the original 'p' burst and 2) halving the gain of the original burst. The amplitude manipulations were accomplished by marking off the burst section of the waveform, then multiplying the marked section by a factor of 2 and .5 respectively. Thus, there was a 6 dB difference between the RMS energy of the two bursts. The RMS energy of the quiet burst was 10 dB higher than the average RMS energy of the 's' frication, and was 13 dB lower than the average RMS energy of the first 15 ms of voicing. A comparison of a token in which the burst was completely removed with the quiet and loud burst tokens established that both the quiet and the loud bursts produce 'blit' percepts that were acceptable to listeners; the form without the burst sounded like 'lit' while the other two forms sounded like two versions of 'blit'[4]. The 's' frication was then recombined with the strong and weak bursts so that there was a silent gap with a duration of 10-90 ms between the frication and the onset of the burst, resulting in a total of 18 base stimuli.

[3] The speaker's voice fell within the average range for female speakers of American English, as determined by comparing the speaker's long term spectral average with data from Cox & Moore (1988), and the voice exhibited no overt pathologies.

[4] As Repp (1983) noted, to native speakers of English, an unaspirated voiceless stop is perceived as a voiced stop. This is due to the fact that, for most speakers of English, the VOT value of 'p' after 's' is the same as that of 'b'.

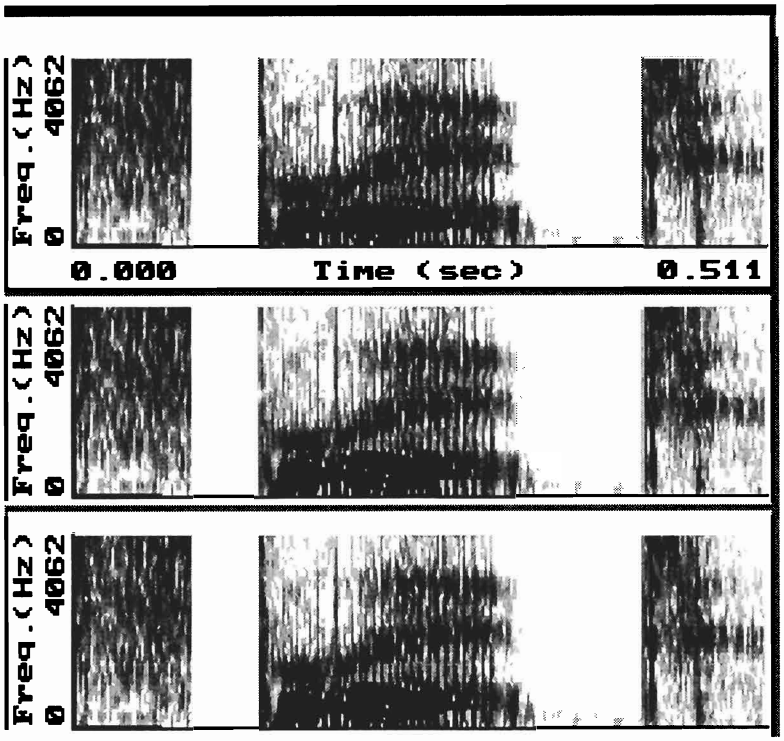


Figure 2 Spectrograms of 'split' with normal, halved and doubled 'p' bursts

The two interval variable AX discrimination test included the equivalent of previous researchers' comparison types, a "cooperating cues" and a "conflicting cues" comparison type[5]. In the "cooperating cues" comparison a token with a quiet burst (Q) was paired with a token that had a loud (L) burst and a 20 ms longer silent duration: [10-70 ms]Q-[30-90 ms]L. The term "cooperating" here refers to the fact that subjects were asked to discriminate between a stimulus that had a loud burst and a longer silent gap and one that had a quiet burst and a shorter silent gap potentially making the two stimuli more discriminable. Discrimination could be made on either dimension, and taking both into account maximized the distinctness of the stimuli. In the "conflicting cues" comparisons a loud burst token was paired with a quiet burst token that had a 20 ms longer silent duration: [10-70 ms]L-[30-90 ms]Q. The term "conflicting" is used here because the discrimination is between a stimulus that had a loud burst and a short silent gap and one that had a quiet burst and a longer silent gap potentially making the two stimuli less discriminable. Discrimination could be made on either dimension, but taking both into account minimizes the distinctiveness of the stimuli. There were 7 possible pairings and 2 possible pair internal orderings resulting in 14 stimuli pairs per comparison type. The stimulus pairs were matched with an equal number of same-same pairs.

[5] Following Best et al (1981) and Fitch et al (1980) I have chosen to refer to these categories as "cooperating" when the two cues point to the same choice, and conflicting cues when the two cues point to opposite choices. However, since the burst amplitude does not point away from a particular gesture it could be argued that the term "conflicting" is a misnomer. This point only serves to highlight the bias towards a gestural interpretation in the previous testing paradigms.

The stimuli were randomized and played out on line using the speech perception facility at the UCLA Phonetics Laboratory (see Johnson & Teheranizadeh, 1992). There was a 300 ms interstimulus interval and a 1000 ms inter-block interval. The subjects were seated in a single walled sound booth and listened to the stimuli binaurally using headphones at a comfortable listening level. Each subject completed ten full repetitions of each stimulus set. Responses were collected by computer using labeled button boxes. Subjects were instructed to pay close attention to the stimuli as the differences would be very subtle, and they were also instructed to ignore as much as possible whether or not the two stimuli sounded like two tokens of the same word or like two different words. These instructions are comparable to those of previous studies.

3.1.2 Results

Subjects' responses to the discrimination test are shown as percent different in figure 3. Figure 4 shows the average A-prime (A') values for all stimulus pairs for the subjects. Average discrimination values are shown here in keeping with previous studies. A' is a nonparametric analog of d' that ranges from 0 to 1. It is used instead of d' here because it better represents subjects' sensitivity when there is a small number of subjects (Grier, 1971). A', like d', is a more accurate measure of a subject's perceptual sensitivity than raw data because it takes into account both the hit and the miss rate (Kaplan, MacMillan & Creelman, 1978; Grier, 1971). An analysis of variance was performed on the A' scores to compare the effects of two cooperating and two conflicting cues on discrimination. A two-way repeated measures design was used; the factors were Cue combination (cooperating vs conflicting) and Mean gap (20, 30, 40, 50 ms). Data from the 60-80 ms conditions were not used because there was no variance beyond 60 ms. There was an overall higher discriminability in the cooperating cues pairs [F(1, 6) = 60.76, p<.001] relative to the conflicting pairs. There were also significant differences in discriminability between the conditions of gap duration [F(3,18) = 9.74, p<.001]; discrimination was greatest at the 40 ms mean duration. The 40 ms condition also corresponds to the greatest difference in discrimination functions between the cooperating cues pairs and the conflicting cues pairs [F(3,18) = 4.42, p<.02] as shown by the significant interaction. The results indicate that the two cues interact to determine discriminability. The results show the same distinction between cooperating and conflicting cues as in previous studies.

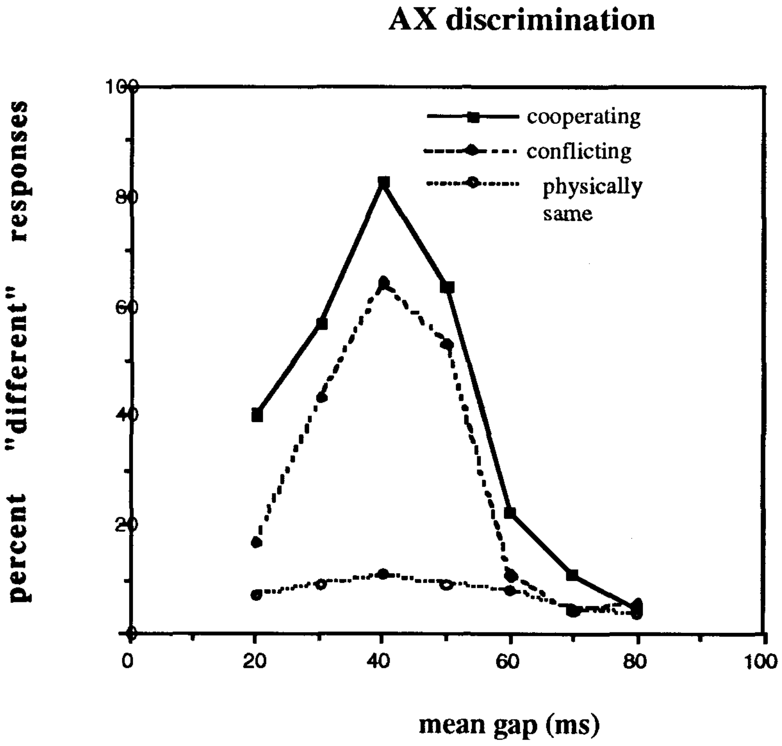


Figure 3 AX discrimination: Percent different

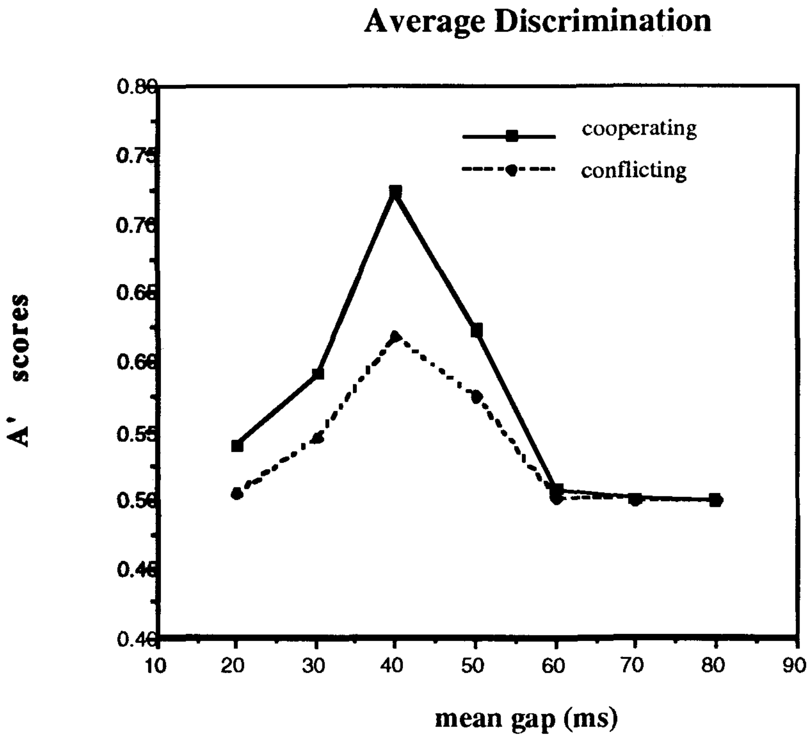


Figure 4 AX Discrimination: A' scores

3.2.0 Experiment II: Forced Choice Identification

3.2.1 Methods

The subjects from the first experiment participated in the identification experiment. The same set of 18 stimulus tokens was used. There was a 1000 ms inter-stimulus interval. The tokens were randomized and played out one at a time binaurally at comfortable levels using ear phones. As in the first experiment, subjects were seated in a single walled sound booth. The subjects were told that they would hear words and that after each one they should respond by pressing either the button labeled "slit" or the button labeled "split" depending on which word they heard. The stimulus files were played out and the responses and response times recorded using the computer-based speech perception setup in the UCLA Phonetics Laboratory.

3.2.2 Results

The results for the forced choice identification test are shown in figures 5. Crossover points (50 % 'split' responses) were determined by fitting a third order polynomial curve to the data points and taking a graphic measure. The average crossover point for the loud burst fell at 38.2 ms mean silence duration and at 50.2 ms for the quiet burst.

The two crossover points for each subject were entered into a two tailed paired t-test. The crossover difference was significant (t=6.348, p < .001). To be perceived as 'split' the quiet burst stimulus needed 12 ms more silence than did the loud burst stimulus. Again there was considerable between subject variability as to where the crossover point fell along the stimulus continuum (24.5-48 ms for the loud burst and 37.5-64 ms for the quiet burst). The between subject variation is in agreement with previous experiments on trading relations. For example the range of crossover points in Best et al (1981) showed a range of 11.4-52.0 ms and 40.0-94.0 ms in two conditions, and Fitch et al showed a 13 ms variability. Figures 6-12 show the individual results for the identification test.

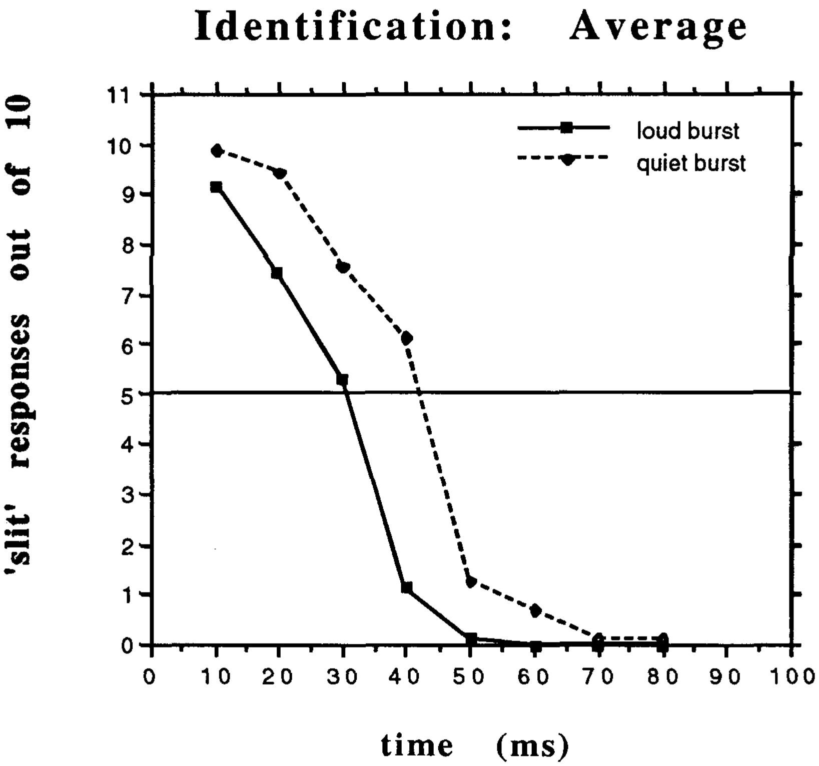


Figure 5 Average identification results

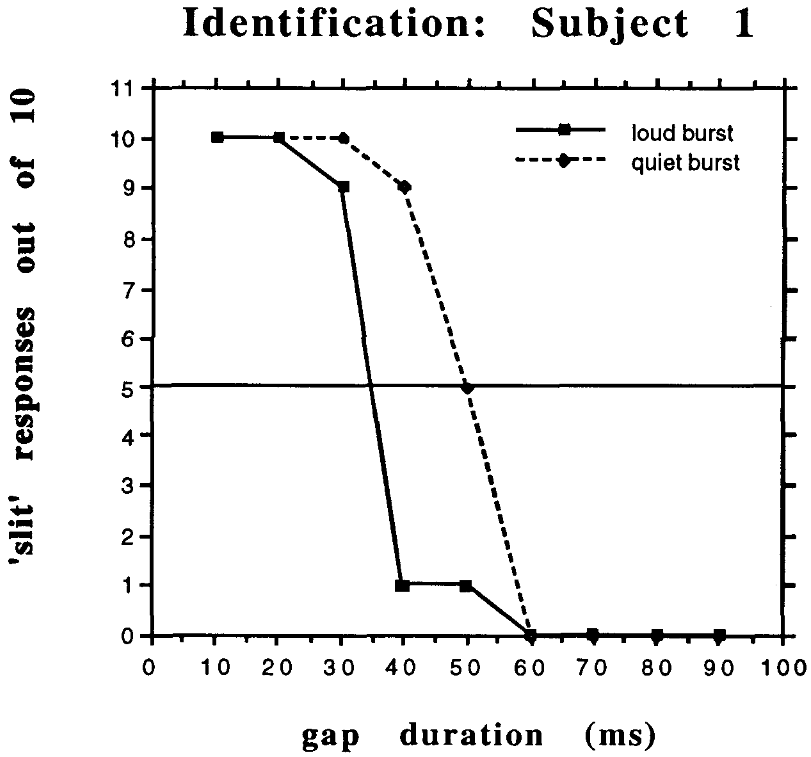


Figure 6 Identification results: subject 1

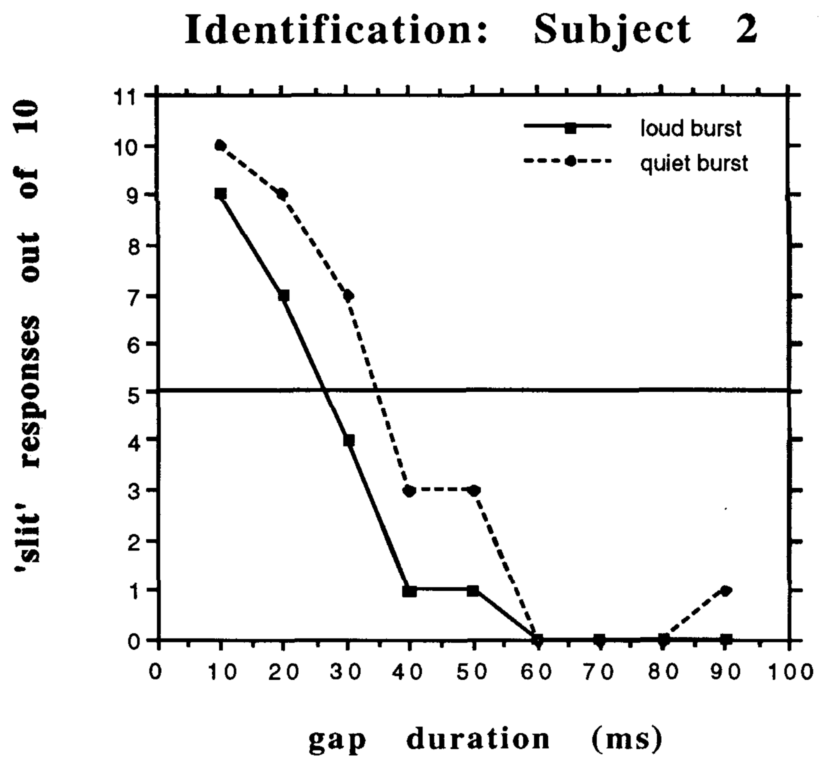


Figure 7 Identification results: subject 2

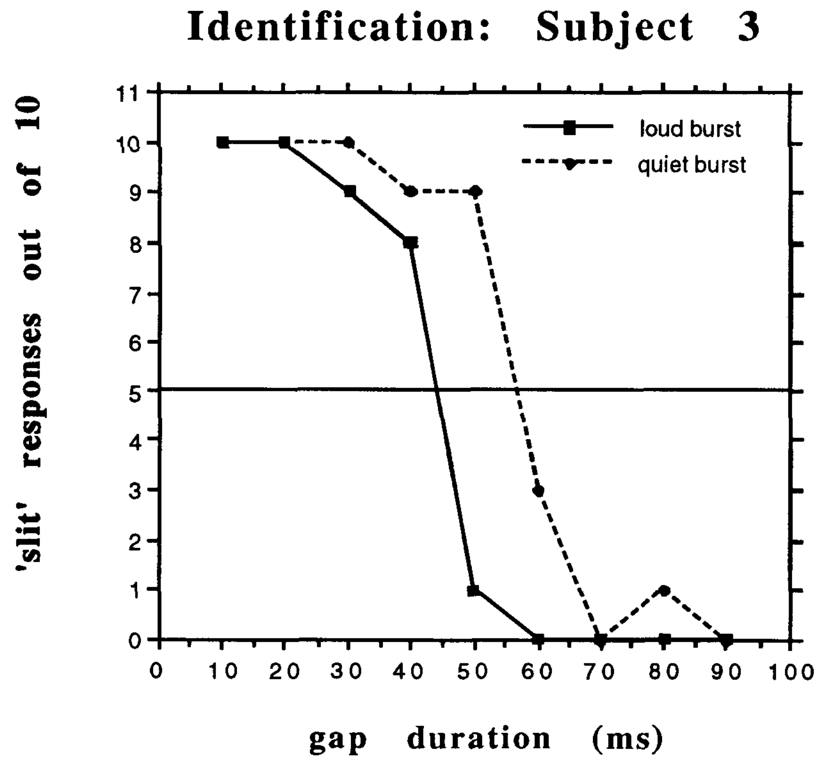


Figure 8 Identification results: subject 3

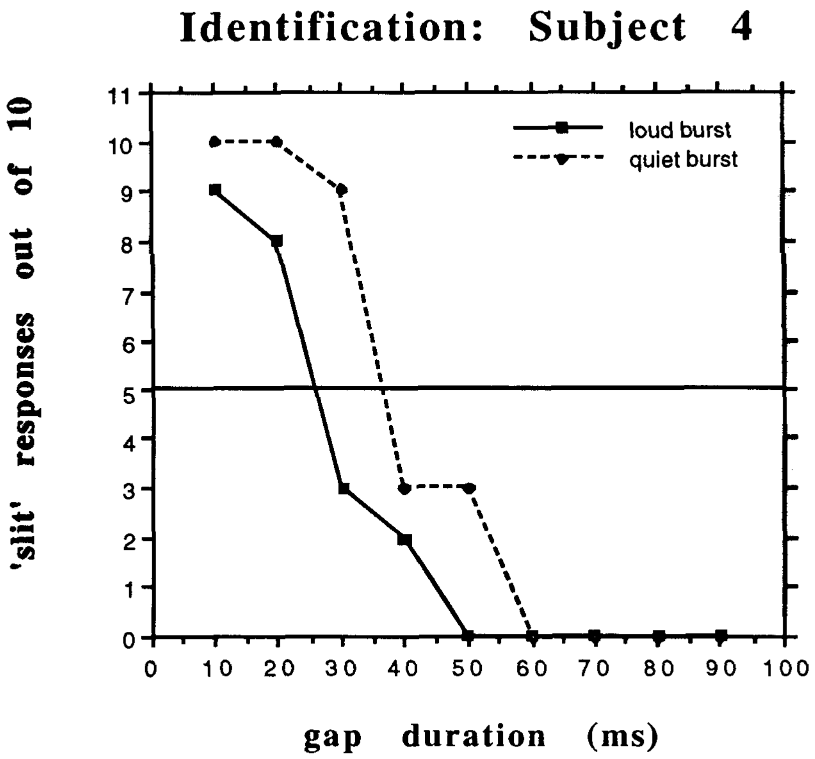


Figure 9 Identification results: subject 4

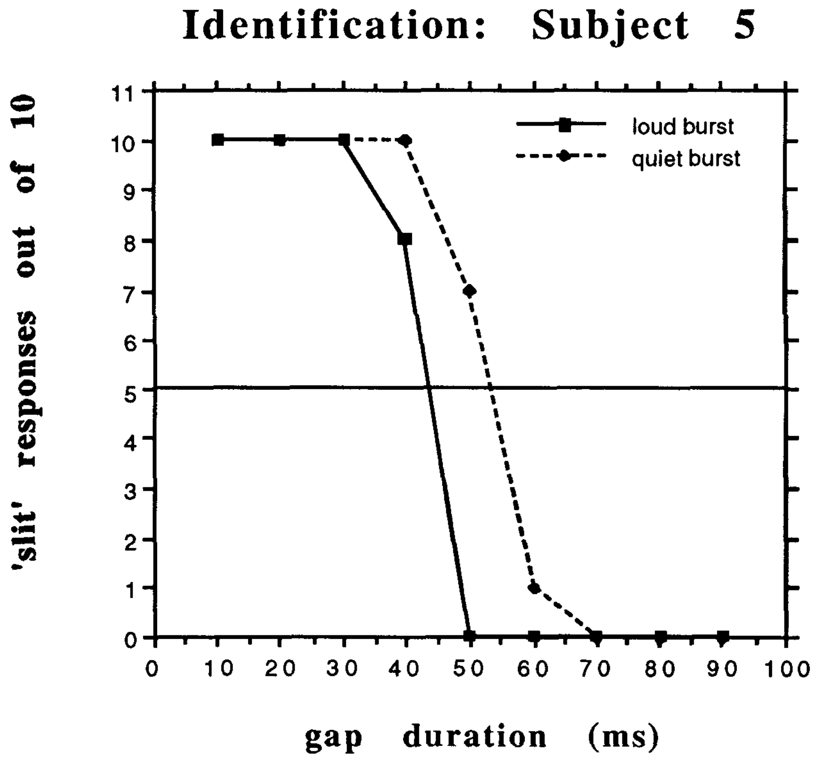


Figure 10 Identification results: subject 5

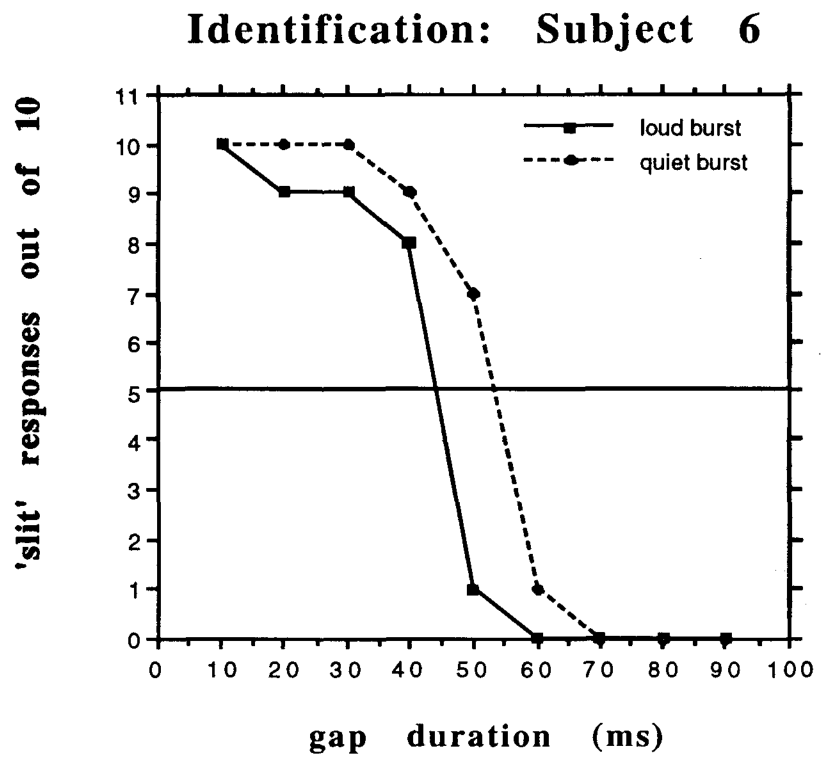


Figure 11 Identification results: subject 6

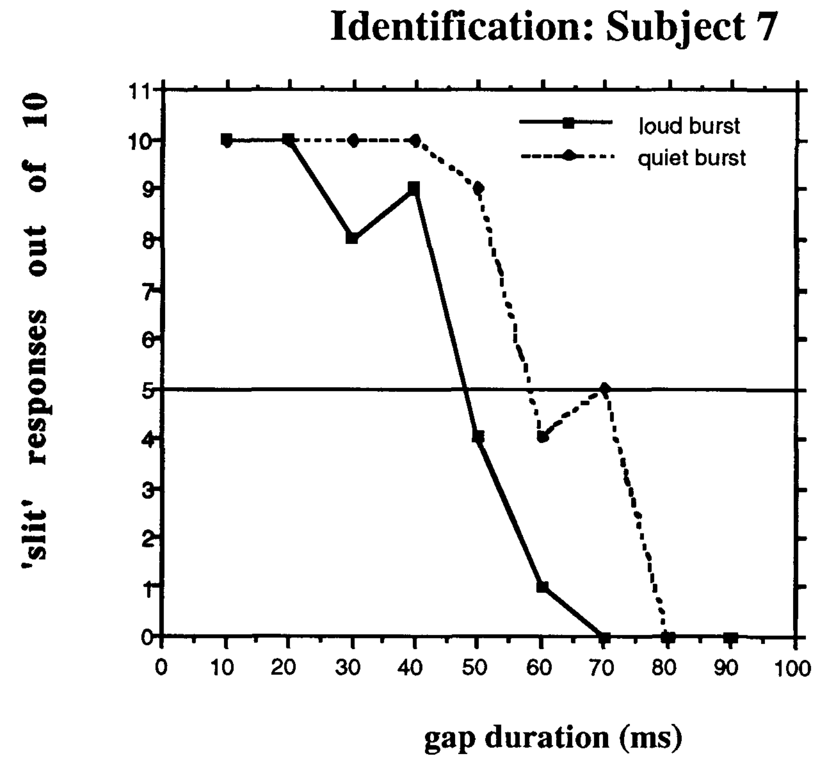


Figure 12 Identification results: subject 7

3.2.3 General discussion

All of the subjects demonstrate a clear trading relation between the amplitude of theburst and the duration of the silence (referred to as the gap in the graphs). The fact that the amplitude of a burst, a cue that is not phonetically relevant for the presence *vs* absence of a stop, participates in a trading relation in a way that is similar to cues that are relevant to the distinctions between gestural categories poses a problem for a theory that relies on direct perception of a distal event. In an informational model, however such a result is expected. In the AX discrimination test the increase in sensitivity at the crossover point is due to subject uncertainty at the point where the signal produces an equally good fit for either mental representation.

The phonetic module account, on the other hand, predicts that subjects would not show an overall higher sensitivity for the cooperating cues stimulus pairs than for the conflicting cues stimulus pairs. That is, fluctuations in amplitude of the burst should not bear on the output of the phonetic module because a stimulus is perceived as the distal event and any sufficient burst should point to the "split" end of the continuum. Any presence of a burst that is appropriate for a particular stop should contribute to the percept of the stop's presence. Subjects' sensitivity to amplitude differences should be equally high across the spectrum since the phonetic module would pass on the amplitude differences to the general auditory processing mechanism. With a short enough silence duration the presence of the burst could be overridden of course. The only cue present in the signal that should affect the crossover from one category to the other is the duration of the silence since it is the only gesturally relevant cue. The results clearly show that this isn't the case; thus, the phonetic module account must be wrong.

As in the results of the discrimination experiment, the identification results support an information integration approach such as the FLMP. There was a clear trading relation established between the "phonetically significant" silent cue and the phonetically insignificant spectral cue. This trading relation is predicted by a model that assigns a constant value to a particular cue that is dependent on the subject's certainty of the cue being present in the signal. What is more, the general lack of agreement among listeners as to where the exact crossover point is can also be seen as evidence against a model such as the motor theory which is based on discrete units such as gestures. All of the stimuli were made from a single speaker's utterance of a single word. It seems that innately specified units would be more uniform in their boundaries. Of course it can be argued that while the crude boundaries are innately specified, the actual boundaries are arrived at subjectively through experience with the language. In fact if enough modification is made to the motor theory it can even incorporate a FLMP style of pattern matching algorithm to explain the workings of the phonetic module. However, the fact that a cue that was not gesturally correlated traded off with a gesturally correlated cue would be difficult to explain given a phonetic module that interprets the signal directly in terms of the speaker's intended gestures. Thus these results can also be seen as evidence against the direct realist approach to auditory perception.

4.0 Conclusion

Proponents of the motor theory have claimed that models of auditory perception are unable to account for either the many to one relation of cues to segments or the perceptual equivalence among cues. However, informational models such as the FMLP are designed to use multiple and continuously valued cues in auditory perception. This makes the nature of the cues that are involved in trading relations one of the key arguments that will decide which of the two models is more accurate in its predictions. It has been claimed by proponents of the motor theory that the fact that cues that are "phonetically" significant are involved in trading relations is evidence for a phonetic module. Informational models, on the other hand, would predict that variations in the signal that are not phonetically relevant could nevertheless be involved in trading relations if they heightened the saliency of a particular feature. A pair of experiment was designed to test the predictions of the two theories using a cue that is phonetically significant (the silence duration), and one that is not phonetically significant (the amplitude of the release burst). The presence of the burst itself is phonetically significant as it is a cue to labial release; what is not phonetically significant is the difference in audibility of the high amplitude burst and the low amplitude burst.

The results of the experiment clearly show a trading relation occurring between the audibility of the burst and the duration of silence preceding the burst. The results can be interpreted as evidence in favor of a model that does not involve a low level phonetic module that perceives the gestures directly as a distal event since an audible release burst is the same distal event regardless of its amplitude. The FLMP provides a convenient model that predicts phenomena such as trading relations as part of the general auditory perceptual process. What is particularly attractive in accounting for trading relations using this model is the continuous nature of the values that are assigned to the features in the signal.

A continuous value between 0 and 1 is assigned to a feature depending on the perceptual system's certainty of the feature being present in the signal. The greater the certainty, the higher the value. In the perception of the stop consonant /p/ in this experiment, there were two key features that were being manipulated: the release burst of the /p/ which was assigned one of two values (since there were only two levels of loudness): i and j; and the silent cue that was assigned a value that had a variable value assigned to it depending on its duration: n. In the FLMP, values are multiplicative. That is, when two features are present in the signal their values are multiplied by each other so that the combined result is greater than the sum of the features. In addition, degradation of input only lowers the values of the distorted or masked features, but a negative value for a feature cannot be achieved. When a stimulus that has a quiet burst also has a preceding duration of silence that is adequately long it can be an equally good fit to a stored representation of the phoneme /p/ as a stimulus with a loud burst but a shorter preceding silent duration. Thus the manipulation of the burst can be compensated for by equivalent manipulation of the silence duration. This sort of relationship between the values of cues accounts for the perceptual equivalence of different stimuli in trading relations without encumbering the perceptual process with unwieldy mechanisms that are specific to speech.

REFERENCES

Abramson, A. S. & Lisker, L. , 1965. Voice onset time in stop consonants: Acoustic analysis and synthesis. In D. E. Commins (Ed.), *Proceedings of the 5th Congress of International Acoustics*. Liège, Belgium: Thone, 1965.

Barclay, J.R., 1972. Noncategorical perception of a voiced stop: A replication. *Perception and Psychophysics*, **11**, 269-273.

Best, C.T, Morrongiello, B., & Robson, R., 1981. Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception and Psychophysics*, **31**, 65-85.

Blumstein, S. E., & Stevens K. N., 1979. Acoustic invariance in speech production: Evidence from measurements of stop consonants. *Journal of the Acoustical Society of America*, **66**, 1001-1017.

Bregman, A. 1978. The formation of auditory streams. In R. Requin (ed.), *Attention and Performance VII*. Hillsdale: N.J.: Earlbaum.

Bregman, A., 1987. The meaning of duplex perception, sounds as transparent objects. In Schouten, M.E.H. (Ed.), *The Psychophysics of Speech Perception*. Dortrecht: Martinus Nijhoff Publishers.

Bregman, A., 1990. *Auditory Scene Analysis*. Cambridge: MIT Press.

Cox, R.N., & Moore, J.N., 1988. Composite speech spectrum for hearing aid gain prescriptions. *Journal of Speech and Hearing Research*, **31** (1), 102-107.

Cutting and Rosner, 1974. Categories and boundaries in speech and music. Perception and Psychophysics, **16**, 564-570.

Delattre, P.C., Liberman, A.M., & Cooper, F.S., 1955. Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, **27**, 769-773.

Derr, M.A., & Massaro, D.W., 1980. The contribution of vowel duration, F0 contour, and frication duration as cues to the /juz/ - /jus/ distinction. *Perception and Psychophysics*, **27**,51-59.

Elman, J.L., & McClelland, J.L., 1984. The interactive activation model of speech perception. In N. Lass (ed.), *Language and Speech*. New York: Academic.

Fitch, H.L., Hawles, T., Erickson, D.H., & Liberman, A.M., 1980. Perceptual equivalence of two acoustic cues for stop manner. *Perception and Psychophysics*, **27** (4), 343-350.

Fowler, C.A., 1986. An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, **14**, 3-28.

Fowler, C.A., & Rosenblum, L.D., 1990. Duplex perception: A comparison of monosyllables and slamming doors. *Journal of Experimental Psychology: Human Perception and Performance*, **16**, 742-754.

Gibson, J.J., 1966. *The Senses Considered as Perceptual Systems*. Boston, MA: Houghton Mifflin.

Grier, J.B., 1971. Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin*, **75** (6), 424-429.

Hessen, A.J. van, & Shouten, M.E.H., 1992. Modeling Phoneme perception. II: A model of stop consonant discrimination. *Journal of the Acoustical Society of America*, **92** (4), 1856-1868.

Hoffman, H.S., 1958. Study of some cues in the perception of voiced stop consonants. *Journal of the Acoustical Society of America*, **30**, 1035-1041.

Johnson, K.A., & Ralston, J.V., 1990. Automaticity in speech perception: Some speech/nonspeech comparisons. *Research on Speech Perception: Progress Report, Indiana University*, **16**, 65-98.

Johnson, K.A., & Teheranizadeh, H., 1992. Facilities for speech perception research at the UCLA phonetics lab. *UCLA Working Papers in Phonetics*, **81**, 123-139.

Kaplan, H.L., MacMillan, N.A., & Creelman, C.D., 1978. Methods and Designs: Tables of d' for variable-standard discrimination paradigms. *Behavior Research Methods and Instrumentation*, **10** (6), 796-813.

Klatt, D.H., 1979. Speech perception: A model of acoustic-phonetic analysis and lexical access. In R.A. Cole (ed.), *Perception and Production of Fluent Speech*. Hillsdale, NJ: Erlbaum.

Klatt, D.H., 1989. Review of selected models of speech perception. In W. Marslen-Wilson (ed.), *Lexical Representation and Process.* Cambridge: MIT Press.

Kuhl, P., & Miller, J.D., 1978. Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America,* **63**, 905-917.

Ladefoged, P., 1980. What are linguistic sounds made of? *Language,* **56**, 485-502.

Ladefoged, P., and Broadbent, D.E., 1957. Information conveyed by vowels. *Journal of the Acoustical Society of America,* **29**, 119-131.

Liberman, A.M., 1982. On finding that speech is special. *American Psychologist,* **37** (2), 148-167.

Liberman, A.M, Delattre, P.C., & Cooper, 1954. The role of consonant-vowel transitions in the perception of stop and nasal consonants. *Psychological Monographs,* **68**, 1-13.

Liberman, A.M., Cooper, F.S, Shankweiler, D.S., & Studdert-Kennedy, M., 1967. Perception of the speech code. *Cognition,* **21**, 1-36.

Liberman, A.M. & Studdert-Kennedy, 1978. Phonetic Perception. In R. Held, H.W. Libowitz, & H.L. Teuber (eds.), *Handbook of Sensory Physiology Vol VIII: Perception.* New York: Springer-Verlag.

Liberman, A. M., & I. G. Mattingly. 1985. The motor theory of speech perception revised. *Cognition,* **21**, 1-36.

Liberman, A. M., & Mattingly, I.G., 1989. A specialization for speech perception. *Science,* **243**, 489-494.

Macmillan, N.A., Braida, L.D., & Goldberg, R.F., 1987. Central and peripheral processes in the perception of speech. In M.E.H. Schouten (ed.), *The Psychophysics of Speech Perception.* The Hague: Martinus Nijhoff.

Massaro, D.W, 1972. Perceptual images, processing time, and perceptual units in auditory perception. *Psychological Review,* **79**, 124-145.

Massaro, D.W., 1987. *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Erlbaum.

Massaro, D.W., 1989. Multiple book review of *Speech perception by ear and eye: A paradigm for psychological inquiry. Behavioral and Brain Sciences,* **12**, 741-794.

Massaro, D.W., & Cohen, M.M., 1976. The contribution of fundamental frequency and voice onset time to the /zi/ - /si/ distinction. *Journal of the Acoustical Society of America,* **60**, 704-717.

Massaro, D.W., & Cohen, M.M., 1977. The contribution of fundamental frequency and voice onset time as cues to the /zi/ - /si/ distinction. *Perception and Psychophysics,* **22**, 373-382.

Massaro, D. W., and G. C. Oden. 1980. Evaluation and integration of acoustic features in speech perception. *Journal of the Acoustical Society of America,* **67**: 996-1013.

McClelland, J.L., & Elman, J.L., 1986. The TRACE model of speech perception. Cognitive Psychology, **18**, 1-86.

McGurk, H., & MacDonald, J., 1976. Hearing lips and seeing voices. *Nature*, **264**, 746-748.

Miller, J.D., 1982. Auditory-perceptual approaches to phonetic perception. *Journal of the Acoustical Society of America*, **71**, S112 (A).

Miller, J. D. & A. Jongman. 1987. Auditory-perceptual approaches to stop consonants. *Journal of the Acoustical Society of America*, **82**: S82 (A).

Oden, C.G., & Massaro, D.W., 1978. Integration of featural information in speech perception. *Psychological Review*, **85**, 573-594.

Pastore, R.E., 1981. Possible psychoacoustic factors in speech perception. In Eimas, P.D., & Miller, J.L. (eds.), *Perspectives on the study of speech*. Hillsdale, NJ: Erlbaum.

Pastore, R. E., Ahroon, W. A., Baffuto, K. J., 1976. A comparative evaluation of the AX and two ABX procedures. *Journal of the Acoustical Society of America*, **60** (Suppl.), S120. (Abstract)

Pisoni, D.B., 1973. Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, **13**, 253-260.

Pisoni, D. B., 1975. Auditory short term memory and vowel perception. *Memory and Cognition* 1975, **3**, 7-18.

Pisoni, D.B., 1977. Identification and discrimination of the relative onset of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America*, **61**, 1352-1361.

Pisoni, D. B., & Tash, J., 1974. Reaction times to comparisons within and across phonetic boundaries. *Perception and Psychophysics*, 1974, **15**, 285-290.

Pisoni, D.B., & Lazarus, J.H., 1974. Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, **55**, 328-333.

Rand, T.C., 1974. Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, **55**, 678-680.

Repp, B. H., 1979. Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. *Language and Speech*, 1979, **22**, 173-189.

Repp, B. H. 1982. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, **92**: 81-110.

Repp, B.H., 1983a. Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization. *Speech Communication*, **2**, 341-361.

Repp, B.H., 1983b. Categorical perception: Issues methods and findings. In Lass, N.J. (ed.), *Speech and Language: Advances in Basic Research and Practice, Volume 10*. New York: Academic.

Repp, B. H., & Mann, V. A., 1980. Perceptual assessment of fricative-stop coarticulation. *Journal of the Acoustical Society of America*, 1980, **67** (Suppl.), S100 (Abstract QQ8).

Repp, B. H., Liberman, A. M., Eccardt, T. & Pesetsky, D., 1978. Perceptual integration of temporal cues for stop, fricative, and affricate manner. Journal of Experimental Psychology: *Human Perception and Performance*, 1978, **4**, 621-637.

Samuel, A.G., 1977. The effect of discrimination training on speech perception: Noncategorical perception. *Perception and Psychophysics*, **22**, 321-330.

Schouten, M.E.H., & Hessen, A.J. van, 1992. Modeling phoneme perception. I: Categorical perception. *Journal of the Acoustical Society of America*, **92** (4), 1841-1855.

Smith, R. L. 1979. Adaptation, saturation and physiological masking in single auditory nerve fibers. *Journal of the Acoustical Society of America*, **65**: 166-178.

Stevens, K. N., 1971. The role of rapid spectrum changes in the production and perception of speech. In L. L. Hammerich & R. Jacobson (Eds.), *Form and substance: Festschrift for Eli Fischer-Jørgensen*. Copenhagen: Akademisk Forlag, 1971.

Stevens, K. N., 1974. The quantal nature of speech: Evidence from articulatory-acoustic data. In L. Pinson & D. Denes (Eds.), *Human communication: A unified view*. Cambridge, Mass: M.I.T. Press, 1974.

Studdert-Kennedy, M. 1980. Speech Perception. *Language and Speech*, **23**: 45-66.

Studdert-Kennedy, M. 1985. Development of the Speech Perceptumotor System. Status Report on Speech Research SR-82/83, Haskins Laboratories, New Haven.

Studdert-Kennedy, M., 1974. The perception of speech. In T.A. Sebeok (ed.), *Current Trends in Linguistics, Volume XII*. The Hague: Mouton, 2349-2385.

Studdert-Kennedy, M., 1982. On the dissociation of auditory and phonetic perception. In R. Carlson & B. Granstrom (eds.), *The Representation of Speech in the Peripheral Auditory System*. Amsterdam: Elsevier, 3-10.

Studdert-Kennedy, M., 1990. Paradigm lost. *Behavioral and Brain Sciences*, **12** (4),774-775.

Studdert-Kennedy, M., Liberman, A.M., Harris, K.S., & Cooper, F.S., 1970. Motor theory of speech perception: A reply to Lane's critical review. *Psychological Review*, **77**, 234-249.

Summerfield, Q., & Haggard, M, 1977. On the dissociation of spectral cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America*, **62**, 436-448.

Whalen, D.H., & Liberman, A.M., 1987. Speech perception takes precedence over nonspeech perception. *Science*, **273**, 169-171.

Wood, C.C., 1975. Auditory and phonetic levels of processing in speech perception: Neurophysiological and information-processing analyses. *Journal of Experimental Psychology: Human Perception and Performance*, **104**, 3-20.

54,000 American Stops

Dani Byrd

Abstract

An analysis of oral and nasal stops, affricates, oral and nasal flaps, and glottal stops was conducted using data from the TIMIT database. The results offer descriptions of the frequency of these segments in the large TIMIT corpus, mean segmental durations, voice onset times, and certain effects of voicing, place, word position, and speaker sex. Because of the quantity and diversity of speech data included, this study provides a characterization of American English stops which represents an overview of American speakers' production of these consonants in read materials. The results obtained are considered with respect to standard descriptions in the field which have generally used datasets which are smaller both with respect to context variability and speaker pool.

Introduction

All the world's languages have stop consonants (Maddieson, 1984). Stops can be considered to be composed of three phases: onset, closure, and offset. They can occur at many places of articulation, with many variations in glottal state and airstream mechanism (see Henton, Ladefoged, and Maddieson, 1992). English utilizes only a small subset of these possibilities. Accounts of the nature of stop consonants in English include Keating (1984), Fox and Terbeek, (1977), Crystal and House (1988a,b,c), Lisker and Abramson (1964), Zue and Laferriere (1979) and many others. This paper will augment these efforts with a report on some characteristics of a large and geographically comprehensive sample of American English stops. The TIMIT database of read American English offers an enormous quantity of data produced by many different speakers.

The TIMIT database was designed jointly by the Massachusetts Institute of Technology, Texas Instruments, and SRI International under sponsorship from the Defense Advanced Research Projects Agency-Information Science and Technology Office (DARPA-ISTO) for the development and evaluation of automatic speech recognition systems (Lamel, Kassel, and Seneff, 1986). It is described by Zue, Seneff and Glass (1990) and Pallett (1990). It was intended that TIMIT incorporate sufficient variability to examine the acoustic realization of phonetic segments as affected by canonical characteristics of the phoneme, contextual dependencies, syntactic effects, and speaker-specific factors of age, dialect, sex and education (Lamel et al., 1986). TIMIT includes 2342 different sentences read by 630 speakers (ten sentences per speaker). There were also two "dialect calibration" sentences read by all 630 speakers. All of the sentences are segmented and labeled as outlined by Seneff and Zue (1988). The validity of the results reported here depends on the correctness and consistency of the phonetic transcriptions. See Keating, Blankenship, Byrd, Flemming, and Todaka (1992) and Byrd (1992a) for a description of UCLA's database implementation of TIMIT and its uses in linguistic phonetic research.

The segments, or "phones," included in this study are the six phonemic oral stop consonants, three consonantal and three syllabic nasal stops, the two affricates, the glottal stop, and the oral and nasal alveolar flaps. Included in this sample are 54,384 stops, affricates, and flaps. The breakdown of the quantity of the data is shown in Table 1.

stops	**24,414**	oral stop closures
	18,101	nasal stop closures
affricates	**2,055**	
flaps	**3,649**	oral flaps
	1,331	nasal flaps
glottal stops	**4,834**	

Table 1--data included in this study of the TIMIT database

This then is by far the most comprehensive study in terms of quantity, and probably diversity, of data on American stops offered to date. Labeling in the TIMIT corpus includes two phases for oral stop consonants: closure and offset (release); and similarly for the affricate. However, nasals, flaps, and glottal stops are segmented as closures only. Below distributional frequency and the durational characteristics of these consonants in TIMIT are described. Some discussion of stop release frequency is also offered. Details of the behavior of these stops in assimilatory processes or their acoustic nature are not addressed. Rather TIMIT is exploited for the "big picture" it offers for the general description of American stops.

I. Oral Stops

Randolph (1989) in his dissertation used three databases, one of which was TIMIT. He noted that transcription errors were infrequent and of a limited number of types. He does note that weak stop releases may go undetected and that /t/'s might be more likely than /p/'s and /k/'s to be transcribed as glottal stop in the presences of pitch irregularities.

A. Closures

A search of the TIMIT database was made for oral stop closures and releases. The dialect calibration sentences were excluded so that the frequency counts would not be biased by the repetition of these sentences by all speakers. In the TIMIT transcription, 24,414 oral stop closures and 21,847 oral stop releases were found. This suggests a ratio of releases to closures of about .895. However, note that in sequences of stops in which the first stop is unreleased, the closure is transcribed as having the place of articulation of the first stop and the release of the second stop. Thus, the ratio of release frequency given for each stop should be considered only rough due to frequent stop assimilation in English. The following table records the number of stop closures and releases of each type produced in a search of the database.

stop	closures (n)	releases (n)	rel (n)/ clo (n)
p	3599	3545	.985
b	2651	3067	1.16
t	6736	5326	.791
d	4179	3275	.784
k	5472	4998	.913
g	1777	1643	.925

Table 2--number of oral stop closures and releases

Note that the /t/ and /d/ categories here and their corresponding durations below exclude flaps, which are labeled separately in TIMIT.

The durations of stop closures are given in the table below and shown graphically in Figure 1.

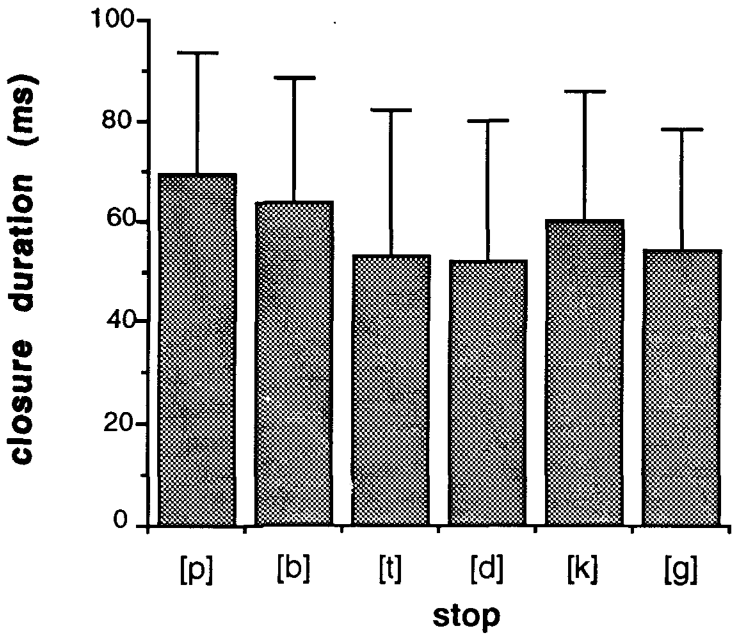


Figure 1: oral stop closure duration by stop identity; [p] 69 (s.d. 24), [b] 64 (s.d. 25), [t] 53 (s.d. 29), [d] 52 (s.d. 28), [k] 60 (s.d. 26), [g] 54 (s.d. 24); all means and standard deviations here and below are reported to the nearest millisecond

Analysis of variance determined there to be a significant effect of stop identity on the duration of closure (F(5,24408)=253.211, p=.0001). A post-hoc Scheffé's S-test showed all pairwise comparisons of closure durations to be significantly ($p<.05$) different except [d] from [t] and [g], and [t] from [g]. Keating (1984) has found fronter closures to have longer durations. This data supports this for the comparison of the labials and velars; however, the intermediate alveolars have the shortest durations. Zue's (1976) finding, not supported in Crystal and House (1988a), of longer closure duration for [p] than for [t] and [k] is supported here.

A further examination of closure durations was conducted with respect to the factors of voicing and place. The voiced closures had a mean duration of 56ms (s.d.=27), and the voiceless closures had a mean duration of 59ms (s.d.=28ms). Figure 2 shows closure duration by place.

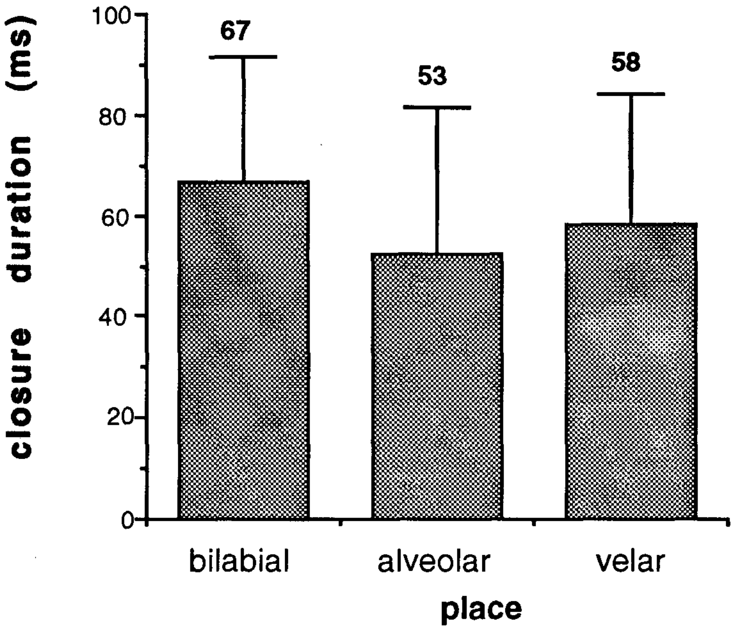


Figure 2: oral stop closure duration by place of articulation; means are shown above the error bars in ms, standard deviations range from 25 - 29ms.

ANOVA determines there to be a significant effect of both voicing (F(1,24405)=122.697), p=.0001) and place (F(2,24405)=525.019, p=.0001). There was also a significant interaction of voicing and place (F(2,24405)=18.365, p=.0001). A post-hoc Scheffé's S-test on place shows all pairwise comparisons to be significantly different (p=.0001). The interaction of place and voicing was significant at the p=.0001 level. The pattern of duration decreasing from bilabial to velar to alveolar is maintained in both voiced and voiceless stop closures. While the mean voiced alveolar stop closure was shorter than the voiceless in accordance with the overall pattern, this difference was on the order of 1ms, whereas the difference at the bilabial and velar places was approximately 6ms. Recall from the above analysis of the effect of stop identity, that the [d] closure duration was not significantly different from that of [t] as shown in the post-hoc test.

These results are in agreement with Crystal and House (1988a,c and 1982) in the general pattern displayed for closure duration as a function of place, with the alveolar closures shorter than those at other places. However, the differences of place found by Crystal and House were slight (1988a). Also in contrast with Crystal and House (1988a,c), these results support Luce and Charles-Luce's (1985) finding that closure duration gets progressively shorter from bilabials to velars to alveolars. Crystal and House do not even find the [p] to be longer than [t] and [k], as found by Zue (1976) and others. The difference in closure duration between the bilabials and other places found here is somewhat smaller than that reported by Fisher-Jørgensen (1964). However, the difference between [b] and [d] is close to that found by Smith (1978). Subtelney, Worth and Sukuda (1966) found that dental stop closures were longer than the labials in their data set, a finding clearly not supported here. Finding labial closures to be longer than those at other places has been suggested to be context free (MacNeilage, 1972). The alveolar closures are shorter presumably because the lesser mass of the tongue tip permits more rapid movement (cf. Kuehn and Moll, 1976). Similar observations concerning the differences between bilabial and velar articulations have to take into account the fact that it seems that the aerodynamic conditions favor a longer closure for bilabials in the case of the stops, perhaps because oral pressure build less rapidly, although Ohala (1983) suggests that this difference is negligible. He suggests that it can be increased through passive and active oral cavity enlargement.

Crystal and House (1988c) found no effect of voicing on closure duration, in contrast to our results at the velar and bilabial places. Note that the studies of Chen (1970) using citation forms and Luce and Charles-Luce (1985) using words in a frame found that voiceless closure duration were longer than voiced. The statistical results reported above for the TIMIT read sentences also generally support this finding. Crystal and House (1988c) remark that it is "sobering to note" that they did not find voiceless stops to have longer closures than voiced stops in their connected speech data as this was found by Meyer (cited in Madebrink, 1955), Medebrink (1955), and Suen and Beddoes (1974) in early work. See also Lisker (1957), Lisker (1972), Umeda (1977) (connected speech), Malécot (1968), Subtelney, Worth and Sukuda (1966), Stathopoulos and Weismer (1983), Slis (1970, 1971), and Prosek and House (1975). Many of these investigators found this difference to be limited to particular word positions such as word-initial, pre-stressed. However, the results above for the large quantity of connected speech data in TIMIT do find voiceless closures to be longer than voiced, although the difference for [t] and [d] did not reach significance in the post-hoc test. Note that Subtelney et al.(1966) found a large difference in the dentals, larger than the bilabials. The difference between the voiceless and voiced bilabials and velars is slightly smaller than that found by Umeda (1977) and Stathopoulos and Weismer (1983) in connected speech; it is substantially smaller than the differences found in the non-connected speech experiments noted above.

B. Release Duration

When followed by a vowel or sonorant, the duration of the labeled stop release phone in TIMIT is equivalent to the phonetic measure of voice onset time. "This period begins with the location of the stop burst and ends at the first sign of periodicity in the waveform" (Randolph, 1989). If a released stop is followed by an obstruent, its release duration will be marked as ending when the characteristic acoustic effect of the next phone begins, i.e. frication for a fricative (Randolph, 1989). Not surprisingly, ANOVA (including both voiced and voiceless stops) finds a significant effect of stop on release duration or VOT (F(5,21841)=1941.821, p=.0001). The release durations are given in Table 3 below. Note the relatively small standard deviations for the voiced stops.

stop	release (ms)	s. d.
p	44	22
b	18	7
t	49	24
d	24	14
k	52	24
g	27	11

Table 3-- stop release durations and standard deviations in milliseconds

A post-hoc Scheffé's S-test shows all pairwise comparisons to be significantly different at the p=.0001 level. In comparison with Lisker's (1964) VOT values for four speakers producing stops in sentences, the TIMIT values are somewhat longer, although they are not as long as Lisker and Abramson's (1964) data in isolated words. Lisker (1957) also notes several cases of prevoicing in his sentence data (see also Lisker and Abramson, 1964; Keating, et al., 1983). The version of TIMIT which is distributed was recorded with a close-talking microphone which does not generally preserves this low amplitude prevoicing voicing, and prevoicing is not segmented/labeled in TIMIT. The TIMIT release duration values for [p, b, d, g] are generally in close agreement with Klatt's (1975) VOT values for word initial single stop consonants spoken in monosyllables by three speakers in a frame sentence, although we might expect the values in that condition to be higher than those for the mixture of read sentences here. However, the VOT values found here for [t] and [k] are much shorter than in Klatt's results.

A two-factor ANOVA testing the effect of place and voicing, and their interaction, on release duration shows there to be significant effects of both place (F(2,21841)=300.759, p=.0001) and, of course, voicing (F(1,21841)=7695.412, p=.0001). There is no significant interaction. The mean release duration for voiced stops is 22ms and for voiceless stops is 49ms.

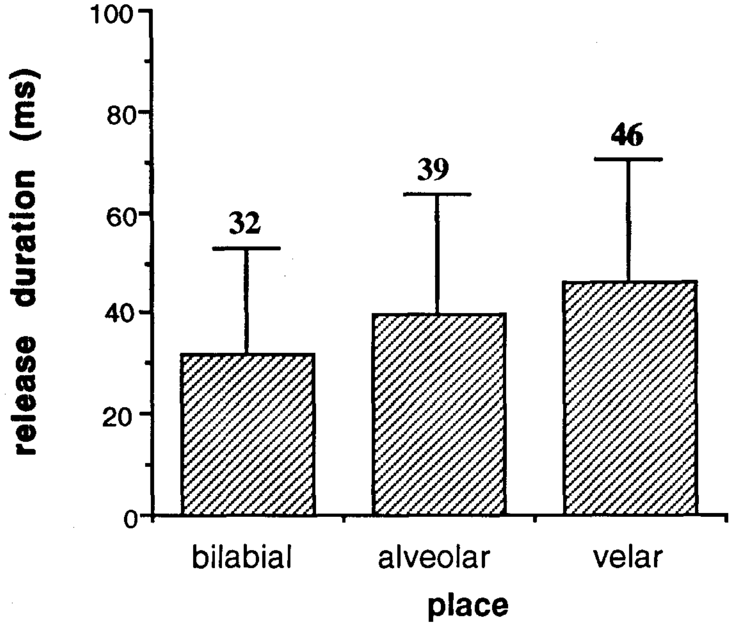


Figure 3: voice onset time by place of articulation

A post-hoc Scheffé's S-test on place show all pairwise comparisons to be significantly different (p=.0001). Henton, Ladefoged, and Maddieson (1992), Crystal and House (1988a), Zue (1976), and others have noted that VOT increases on average as the place of articulation moves from bilabial to alveolar to velar. The TIMIT data is in accordance with this finding. Randolph, using three databases of read sentences, found velars to have the longest release duration at 31ms and labials the shortest at 18ms (Randolph, 1989). This follows the same pattern shown above, but reports generally shorter times.

The differences between the mean voiced and voiceless release duration here are smaller than those reported by Carlson and Granström (1986) but slightly greater than those reported by Crystal and House (1988c). The TIMIT mean release duration for both voiced and voiceless stops is greater than Lisker's (1967) values for 3 speakers in a sentence condition. The TIMIT voiced VOT mean is in close agreement with Klatt's (1975) value for voiced stops before sonorants which he found to be slightly longer than the value before vowels. However, the TIMIT voiceless release duration is much shorter than that found by Klatt (1975) with three speakers recording 25 monosyllabic words beginning with one to three consonants in frame sentences.

Randolph's 1989 dissertation is a major source of segment data from read speech databases. reports voiceless release durations as being twice as long as voiced (49ms vs. 24ms.) in his three database study. Randolph's reported release durations and standard deviations are almost in exact agreement with those found here for the TIMIT database alone. Randolph goes on to note that voiceless stops in syllable onset position have approximately 1.5 times as long a release duration as stops in other syllable positions. He comments that voiceless stops in non-(syllable) initial positions have longer release durations when they precede nasals, glides, and vowels than when preceding affricates, other stops, and fricatives. He notes that stops in a syllable onset have shorter release durations when following obstruents (48ms) then when following nasals, glides, and vowels (Randolph, 1989). Of the onset stops preceded by obstruents, those in clusters with /s/ have the shortest release duration (32ms) (Randolph, 1989). Randolph finds that stops in a non-falling stress environment have the longest release duration of non-onset stops. He comments that this is perhaps an indication of these stops being resyllabified as onset stops.

The two-way interaction of place and voicing is statistically significant at the p=.0040 level. The alveolar and velar stops both have approximately 25ms difference in voiced and voiceless release duration. The bilabial place has approximately a 27ms difference. However, when mean total (closure+release) duration is considered, bilabial and velar releases both have a 31ms difference between voiced and voiceless, and alveolars a 26ms difference between voiced and voiceless. The ratio of voiced total mean duration to voiceless total mean duration is approximately the same for all three places (from .723 for velars to .745 for alveolars.

The relationship between mean closure and release duration for each stop can be seen in Figure 4.

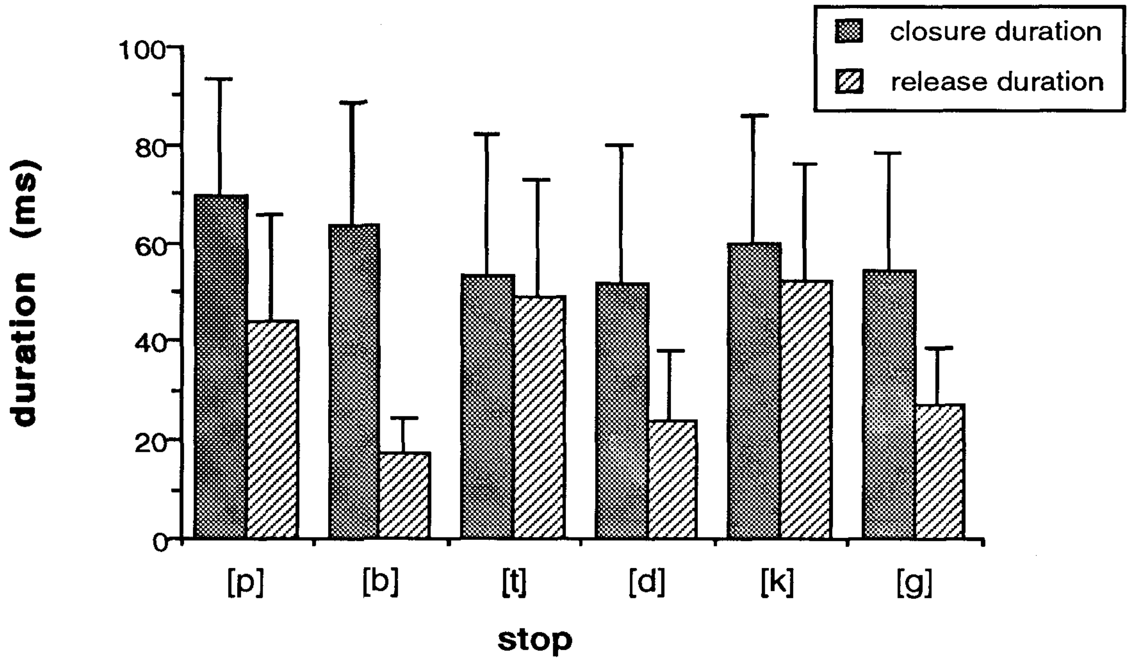


Figure 4: closure versus release duration by stop consonant identity.

Crystal and House (1988a) describe total stop (closure plus release) duration to be similar for alveolars and labials (about 80ms) and longer for velars (100ms). If we add the mean values for closure and release durations calculated from all the stops in TIMIT, regardless of their context, we find that the alveolar place at 92ms is still shorter than the bilabial and velar places. The labial place is the next longest with a value of 99ms, while the velar is longest at 104ms. While the value for the velar place is close to Crystal and House's (1988a) report, the bilabial and alveolar places are somewhat longer and do not follow the pattern suggested by Crystal and House of yielding approximately the same total duration.

C. Release Frequency in Sentence-Final Stops

In order to provide a standard environment in which the release of oral stops could be examined systematically using database searches, the transcription given for the sentence-final position of all (non-calibration) sentences was examined for the occurrence of released and unreleased oral stops.

1130 sentence-final stops were located in the search of the database. 9.1% were bilabial, 77.8% were alveolar, and 13.1% were velar. 37.8% of the stops were voiced, 62.2% were voiceless. The distribution of all six stops is as follows: [k], 11.5%; [t], 43%; [p], 7.7%; [g], 1.2%; [d], 34.8%; [b], 1.8%. A released stop occurred in 59.7% of the cases and an unreleased in 40.3% of the cases.

The place of articulation had a significant effect on whether a release occurred or not as determined by a contingency table analysis (χ^2=40.829, p=.0001). Bilabial stops were released 49.5% of the time, alveolar stops 57% of the time, and velar stops 83.11% of the time. Values for release frequency are presented graphically in Figure 5.

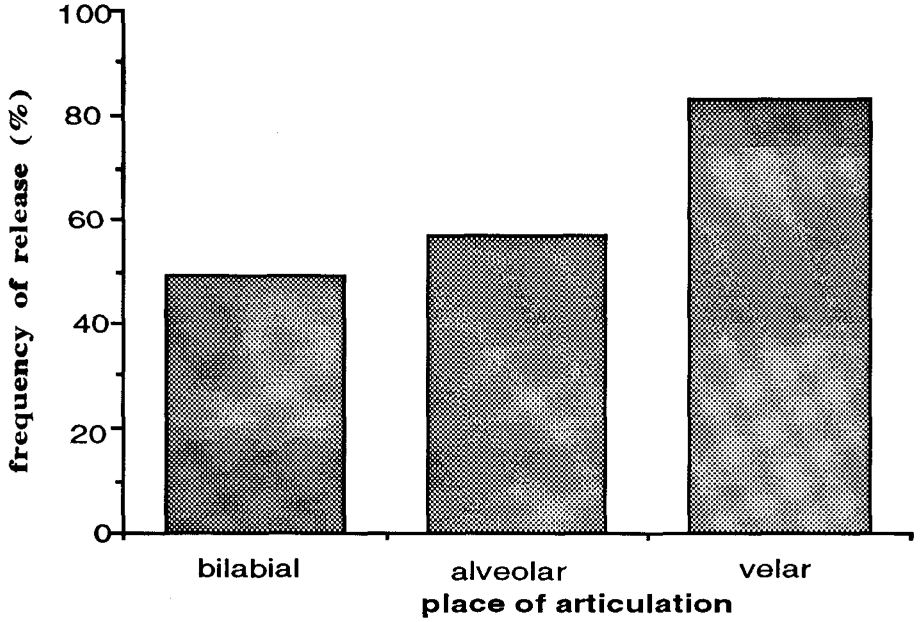


Figure 5: percent frequency of release by place of articulation for sentence-final stops; bilabial 49.5% of the time, alveolar 57% of the time, and velar stops 83.11% of the time.

There are presumably aerodynamic causes for these differences as the velar stops with a small volume of air behind occlusion might produce a more audible release than the alveolar and labial which have increasingly large volumes of air to absorb pressure produced by the pulmonic airstream. Recall however that bilabials also have a longer closure than velars during which pressure may build up, although cavity expansion might mitigate this.

Voicing, however, did not have an overall effect on whether a release occurred, despite the fact that voiced stop bursts are reported as having a lower amplitude than voiceless stop bursts (Halle, Hughes, and Radley, 1957). Voiced stops were released 59.5% of the time and voiceless stops 59.9% of the time. This is in close accordance with Crystal and House's (1988c) report of an overall release frequency of 59% for stop consonant across *all* sentence positions. However, in word-final position, Crystal and House (1988c) found only a 33% release frequency, but they note a problem with this figure due to inadequate diversity in their sample. Crystal and House (1988a) report a tendency for voiceless stops to include a release more often than voiced stops. As an overall effect, this was not evident in the data considered here. Crystal and House (1988c) also find a difference across all contexts in the release frequency between voiced and voiceless stops (65% vs. 33%). This difference is not supported in our data for sentence-final position. Crystal and House (1988a:1557) find "a tendency for voiceless stops to be completed a higher percentage of the time than voiced stops, particularly in word final position." When each place of articulation is tested independently in a contingency table analysis, the bilabial and alveolar place show no effect of voicing on whether a released or unreleased stop occurred. For the velar stops, however, voicing did have a significant effect on whether a released or unreleased stop occurred (χ=3.902, p=.0482). The voiced velars were released in 64.3% of the cases while the voiceless velars were released 85.1% of the time. This fact supports the aerodynamic explanation offered above for the place effects as in American English the voiceless velars are the most likely to have the oral pressure build-up favoring audible stop releases.

Randolph (1989) provides information about stop release frequency as a function of syllable position, information not present in the commercially available TIMIT database. Randolph finds that "stops in the outer onset position are practically always released (97% of the time), whereas in the coda position, they are mostly unreleased (43% vs. 31% released). One also sees that large percentages of stops are released when they are followed by vowels and glides (a necessary but not sufficient condition for the stop belonging to the onset)." (Randolph, 1989, p. 115). Randolph's sub-study of stop realization across three databases included 12,161 tokens which were realized as follows: 7855 released, 2303 unreleased, 1052 flapped, 702 deleted, and 259 glottalized (Randolph, 1989).

Crystal and House (1988a) report a tendency in their data set for the velar consonants to be released more often than bilabial or alveolar consonants and state that this tendency is attributable to the behavior of the voiceless velar consonant. The results reported here for sentence-final stops are in accordance with that finding. A contingency table analysis of the effect of place on the occurrence of a release in which the voiceless velars are excluded yields no significant effect, although the direction of the trend remains unchanged from that shown in Figure 1. As only 18 voiced velar tokens remain in this analysis, a larger sample might raise the trend to significance even without voiceless velars. Another Crystal and House result reported in 1988(b), that "labials, particularly in unstressed syllables, [tend] to be completed [ie. released] more frequently than alveolars and velars" (p. 1580), was not supported by the TIMIT data considered here.

The sex of the talker had a significant effect on release frequency, with women releasing their sentence-final stops more often than men (χ^2=49.146, p=.0001). For a fuller description of these results, see Byrd (1992b). One of the calibration sentences which every speaker read ends in the common stop-final word "that." When the sentence-final position of this sentence is examined, it is found that a released [t] occurs 24% of the time, an unreleased [t] 67% of the time, and a glottal stop 9% of the time. As for sentence-final stops in the rest of the database, when a stop occurred, women released the stop significantly more often than men (χ^2=5.57, p=.0183).[1]

II. Affricates

The database (excluding calibration sentences) was searched for sequences of an alveolar stop followed by a label for an affricate release (which is transcribed differently than the phonetically parallel post-alveolar fricative). 952 voiceless affricates and 1103 voiced affricates were found. Mean closure and frication durations are shown in Table 4 below.

affricate	closure (ms)	s.d. (ms)	release (ms)	s. d. (ms)
[tʃ]	43	18	86	28
[dʒ]	43	19	62	28

Table 4--affricate closure and release durations and standard deviations in ms

ANOVA determines there to be a significant effect of voicing on release duration. No effect of voicing on closure duration is found, in accordance with the findings for [t] and [d] closures. When we compare the closures for the two affricates with the two stop closures at the alveolar place, we find that the affricates have shorter closure durations. This can be seen in Figure 6.

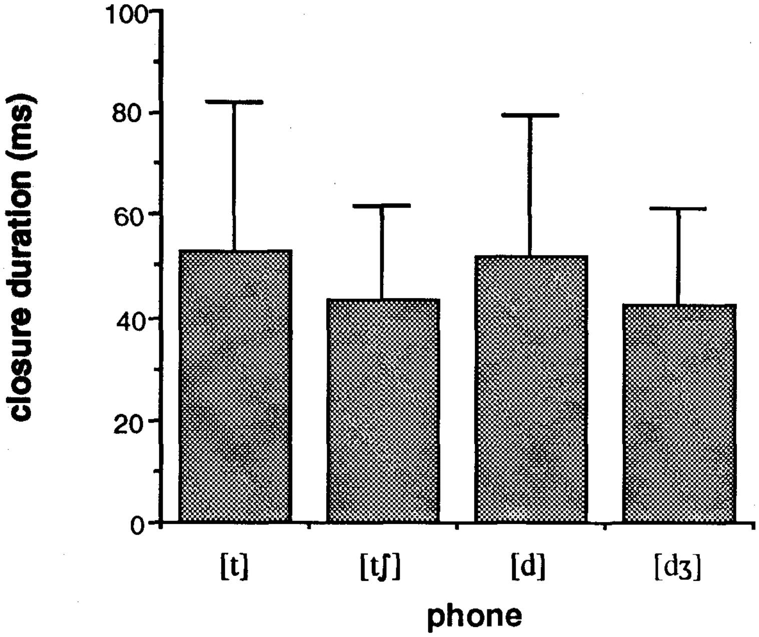


Figure 6: alveolar stop and affricate closure durations

ANOVA testing the four-level factor of closure shows there to be a significant effect (F(3,15064)=62.389, p=.0001) of this factor. A post-hoc Scheffé's-S test shows all pairwise

[1] Note that this result differs from that reported in the footnote of a preliminary study in Byrd (1992b); a more comprehensive study on this topic can be found in Byrd (1992a).

comparisons except that of the two affricate closures to be significantly different at the p<.001 level. If we consider the total (closure+release) of the mean affricate and alveolar stop durations, we find that the affricates are still considerably longer than the stops, approximately 26ms.

III. Nasals

The TIMIT database includes labels for both syllabic and non-syllabic nasals at the bilabial, alveolar, and velar places of articulation. Syllabic consonants are the result of complete reduction of the vocalic syllable nucleus. A total of 18,101 nasals occur; 16,913 of them are non-syllabic. The frequency and durations of each of the nasals are shown in Table 5. (Note that nasal flaps are transcribed separately and not included below.) As the reader may be interested in the distributional frequency of these nasals without the addition of tokens from the two calibration sentences which were read by all speakers, these values are given in parenthesis where different.

nasal	**number** (w/o calib. sen.)	**duration (ms)**	**s. d. (ms)**
m	5509 (4881)	62	26
n	9660 (8496)	55	23
ŋ	1744 (1568)	63	24
m̩	171	79	31
n̩	974 (846)	79	30
ŋ̩	43 (30)	81	24

Table 5-- syllabic and non-syllabic nasal durations and standard deviations in milliseconds, and number of tokens; calibrations sentences are excluded from the number shown in parentheses

As reported in Byrd (1992a), a chi-square test determined there to be no effect of sex on the distributional frequency of syllabic [m̩] and [ŋ̩]. However, the sex of the speaker did have a significant effect on the frequency of [n̩] (χ^2=12.632, p=.0004). Women had significantly fewer [n̩]'s than the men. As reduction in the environment of alveolar consonants is a particularly common process, it is important to note that men and women appeared to produce this and only this syllabic consonant with different frequency.

The syllabic nasals were on average 21ms longer than the non-syllabic nasals. Based on the start and end time of each label, word positions were calculated for each phone as initial, medial, final or unaffiliated Unaffiliated denotes a phone occurring outside the temporal extent of any word labels, such as might occur in a filled pause, or a phone which cannot be determined to belong to one of two abutting words. An ANOVA of the effects of nasal identity and word position, and their interaction, was calculated for the duration values for nasals in all TIMIT sentences (calibration sentences included). Nasal identity (F(5,18080)=8.636), p=.0001) and word position (F(3,18080)=4.070, p=.0067) both have a significant effect on duration. Post-hoc Scheffé's S-tests show the [n] to be significantly different in duration from all the other nasals, while [m] is different in duration from all but the non-syllabic velar nasal. The shorter duration of [n] parallels the findings for alveolar oral stop closures where it was suggested that the tongue tip was the most rapid articulator of the three, but this correspondence is not found in the syllabic case where there seems to be a strong tendency for these syllable nuclei to be realized with a consistent duration. None of the syllabic nasals were significantly different in duration from one another. The shorter duration of the alveolar relative to the other consonantal nasals does not carry over to the syllabic counterparts. The mean durations of the nasal stops in each word position are in Table 6 below. (Ten unaffiliated nasals are not included.)

word position	number	duration (ms)	s.d. (ms)
initial	3767	61	28
medial	7236	54	24
final	7088	64	24

Table 6--mean durations in milliseconds for nasal stops by word position ; (ten unaffiliated nasals are not included.)

Post-hoc tests showed duration to be significantly different in each word position. (Unaffiliated nasals were not significantly different from any other position.) However, these values appear to vary quite a bit depending upon the phone in question. ANOVA also showed a significant interaction of nasal identity with word position. Table 7 shows values for nasal duration at each word position for both the consonantal and syllabic nasals. (Note that initial syllabic velar nasals, of which there were only 15, were found in circumstances when a word phonemically beginning with a velar stop was nasalized initially due to a preceding nasal.)

nasal/position	word initial	word medial	word final
m	59	59	73
n	59	49	59
ŋ	88 (n=2)	54	65
m̩	80	77	81
n̩	84	75	80
ŋ̩	87	74 (n=3)	78

Table 7--nasal duration at each word position in milliseconds for both the consonantal and syllabic nasals

Initial and medial [m] are indistinct in duration, and relatively small differences exist between [m̩] in all positions.

IV. Flaps

Both oral and nasal flaps are transcribed in TIMIT. A total of 4980 flaps occurred; 3649 oral flaps and 1331 nasal flaps. Of these, 1557 were in one of the calibration sentences that included at least two potential flap sites--the word *water* and the phrase *suit in.* The analysis and results below were calculated both including and excluding the flaps from the calibration sentences. The results did not differ so data from all the flaps, including those in the calibration sentences, are presented below. The mean duration of flaps in the database is 29ms (s.d.=8ms). ANOVA testing for differences between oral and nasal flaps and between word positions shows there to be no effect of flap identity and a small but significant effect of word position (F(2,4974)=7.238, p=.0007). There is no interaction. Flaps increased in duration in 1ms steps from medial to final to initial. Zue and Laferriere state that "[a]s a phonetically defined group, flaps vary in duration from 10 to 40ms." (p. 1044) Their study with six speakers yielding 1484 flaps found mean durations of 26 to 27ms for oral flaps, with a range of 10 to 40ms. The range in the TIMIT data for flap duration was 9 to 73ms for oral flaps and 8 to 68ms for nasal flaps. Rimac and Smith found a mean flap duration of 36ms. The TIMIT mean is in close agreement with Crystal and House's (1988c) reported duration of 29ms and other like values reported in Fisher and Hirsch (1976) and Sharf (1962). Fox and Terbeek (1977) present mean durations of flaps in 40 words which appear from their graphical display to range from 21 to 33ms, but no overall mean is reported. Comparison of ranges should be done with caution as there is a high error rate in measuring the duration of flaps from spectrograms as noted by Fisher and Hirsch (1976). They state that flap duration is "on the order of 25 to 35 milliseconds and spectrographic measurements of voiced durations typically cannot be made with an accuracy greater than one period, or about ±5 milliseconds." (p. 191, also Klatt, 1971)

In the calibration sentences 99% of the speakers had a flap in *water* while only 19% had one at the word-final site. These two flaps were significantly different in duration (F(1,743)=72.185, p=.0001). The word-final flap had a mean duration of 33ms (s.d.=8ms), and the word-internal flap had a mean duration of 27ms (s.d.=6ms).

Effects of sex on the frequency of flaps are reported in Byrd (1992a) and Zue and Laferriere (1979). Women were found to have fewer flaps than men in both studies. In Byrd (1992a) chi-square tests found a significant effect of sex on the frequency of both oral and nasal flaps in the TIMIT data (χ^2=55.341, p=.0001 and χ^2=11.41, p=.0007, respectively). The women produced significantly fewer flaps than the men. A three-factor ANOVA with all two-way interactions testing for effects of speaker sex, flap identity and position on duration show there to be a trend for women to have shorter flaps than men (F(1,4970)=3.489, p=.0618). Women had a mean nasal flap duration of 28ms as compared to 29ms for the men. (The difference in oral flap durations was smaller (.6ms).) This is in spite of the fact that women spoke more slowly in the TIMIT calibration sentences (Byrd, 1992a). This may contrast with the finding by Zue and Laferriere (1979) of longer medial [t] and [d] durations for women in "seven out of eight cases" (p. 1047); they are not explicit about whether this includes the flaps. There was also a significant interaction of sex with word position (F(2,4970)=3.163, p=.0424). Most of the difference between male and female flap durations appears to occur in word final flaps which are approximately 2ms different. Flaps at the other positions have approximately the same duration.

V. Glottal Stops

While not phonemic in English, the glottal stop does occur in allophonic and stylistic variations, although little if any prior study has been done of its overall usage. 4834 glottal stops occur in TIMIT, including 1222 glottal stops from the calibration sentences. Ignoring occurrences in calibration sentences, the non-phonemic glottal stop of English occurred more frequently than the oral stop closures [p], [b], and [g]. Henton, Ladefoged, and Maddieson (1992) report that (phonemic) glottal stops generally have closure durations at least as long as other stops' closures. The glottal stops in TIMIT have a mean duration of 65ms (s.d.=32) making them longer than all the other oral stop closures except [p]. The glottal stop closure durations have a greater standard deviation than the oral stop closures as well. A post-hoc Scheffe's S-test determines that the glottal stop closure durations are significantly different from all the oral stop closure durations except those of [b].

When the word position of the glottal stop is considered we find that initial glottal stops made up 49% of the total, medial glottal stops 6% of the total, final glottal stops 16% of the total, and unaffiliated glottal stops 29% of the total. Note that when a glottal stop occurred between two vowels at a word boundary, it was considered unaffiliated. A two-factor ANOVA testing the effects of word position and speaker sex on glottal stop duration finds a significant effect of word position (F(3,4826)=84.745, p=.0001), no effect of sex, and a significant interaction of the two factors (F(3,4826)=4.554, p=.0034. In Figure 7 below, glottal stop duration is shown according to word position. The unaffiliated glottal stops are the longest, followed by the final glottal stops. The initial and medial glottal stops are the shortest and are the only pair not significantly different in a Scheffé's post-hoc test.

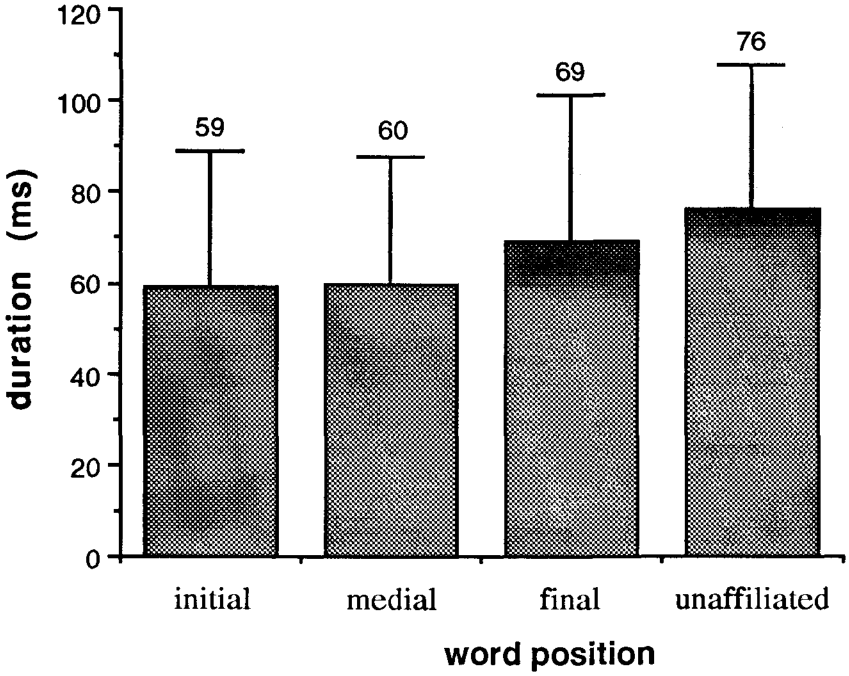


Figure 7: glottal stop duration by word position

The glottal stops produced by women had a longer mean duration by 2 to 4ms than those by the men in every position except medial where they were shorter by almost 7ms.

The effects of sex on the frequency distribution of glottal stop in TIMIT are reported in Byrd (1992a). In summary, a chi-square test shows that the women use significantly more glottal stops than the men, in every word position (p≤.0103). When we consider the small set of 57 glottal stops produced in place of the sentence-final [t] of the word "that" in one calibration sentence, we find with a chi-square test that the production of a glottal stop in this position is not significantly influenced by sex, although the distribution favors the direction demonstrated above. It is somewhat unexpected to find this relationship of sex to the frequency of glottal stop. In fact, women's voices are often characterized as more breathy than men's, and glottal closure is related to creakiness in the voice quality. It may be that the glottal stop is used as a devoicing mechanism more often by women or that it participates in allophonic patterns which are less productive for the men.

VI. Conclusion

The use of a large, commercially available database in the study of segmental duration has been described. The material includes 630 American speakers from a range of geographical locations reading 2342 different sentences. The corpus yields approximately 54,000 stops making this description of stops very comprehensive in terms of quantity and context and speaker diversity. TIMIT is not as useful for the investigation of effects of specific sentential context or the variability found in the productions by a single speaker. The data presented above is intended to provide a broad picture of durational characteristics of American stops. It has also offered results on the structure of stops, their distributional frequency, and certain effects of talker sex. Comparisons are made with findings previously discussed in the literature. A comparison of these findings with those reported in earlier studies is important as much of this work has used isolated words or words in carrier phrases and small groups of speakers. Such studies may not reflect the variability found in a larger population of the language's speakers and the limitation to carefully controlled test items may focus a speaker's attention on contrasts, thereby exaggerating them. The similarities and differences found in the TIMIT results inform us as to the extent to which we can generalize from experimental studies to corpuses including a great deal of variation in reading material and speakers. Of course an even broader scope of study would be provided by comparison of large read with non-read corpuses, something not now possible. As a large and diverse collection of labeled speech, TIMIT provides an interesting testing ground for the linguist to assess the accuracy of generalizations based on previous laboratory experimentation. In addition to being of academic interest, linguistic knowledge of reliable differences in segmental characteristics may help aid in the phonetic classification and speech recognition goals which TIMIT was designed to serve. The study presented above outlines many reliable segmental characteristics of American stops, nasals, affricates, flaps and glottal stops.

Acknowledgments

This research was supported by the National Science Foundation and the UCLA Department of Linguistics. Many thanks are due to Edward Flemming for designing and implementing the structure of a relational Macintosh database with which TIMIT is used at UCLA. I am grateful to Patricia Keating, Peter Ladefoged, and Ian Maddieson for their insightful comments on an earlier version of this paper.

References

Byrd, D. (1992a) Sex, dialects, and reduction. *Proceedings of the International Conference on Spoken Language Processing.*

Byrd, D. (1992b) Preliminary results on speaker-dependent variation in the TIMIT database. *JASA*, 92:593-596.

Chen, M. (1970) Vowel length variation as a function of the voicing of the consonant environment. *Phonetica* 22, 129-159.

Crystal, T.H.; House, A.S. (1982) Segmental durations in connected-speech signals: Preliminary results. *JASA* 72(3), 705-716.

Crystal, T.H.; House, A.S. (1988a) Segmental durations in connected-speech signals: Current results. *JASA* 83(4), 1553-1573.

Crystal, T.H.; House, A.S. (1988b) Segmental durations in connected-speech signals: Syllabic stress. *JASA* 83(4), 1574-1585.

Crystal, T.H.; House, A.S. (1988c) The duration of American-English stop consonants: an overview. *J. Phon.* 16(3), 285-294.

Fischer-Jørgensen, E. (1964) Sound duration and place of articulation. *Z. Phonet. Sprachwiss. Kommunikationsforsch.* 17, 175-234.

Fisher, W.M.; Doddington, G.R.; Goudie-Marshall, K.M. (1986) The DARPA speech recognition research database: specifications and status. *Proceedings DARPA Speech Recognition Workshop*, 93-99, 1986.

Fisher, W.M.; Hirsch, I.J. (1976) Intervocalic flapping in English. *Chicago Linguist. Soc.* 12, 183-198.

Fox, R.A.; Terbeek, D. (1977) Dental flaps, vowel duration and rule ordering in American English. *J. Phonetics* 5, 27-34.

Halle, M., Hughes, G.W., and Radley, J.-P.A. (1957) Acoustic properties of stop consonants. *JASA* 29(1), 107-116.

Henton, C.; Ladefoged, P.; Maddieson, I. (1992) Stops in the World's Languages. *Phonetica* 49, 65-101.

Keating, P.A. (1984) Phonetic and phonological representations of consonant voicing. *Language* 60, 286-319.

Keating, P.A.; Blankenship, B.; Byrd, D.; Flemming, E.; Todaka, Y. (1992) Phonetic analyses of the TIMIT corpus of American English. to appear in the *Proceedings of the International Conference on Spoken Language Processing.*

Klatt, D.H. (1975) Voice onset time, frication, and aspiration in word-initial consonant clusters. *J. Speech Hearing Res.*, 18, 686-706.

Kuehn, D.P. Moll, K. (1976) A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics* 4, 303-320.

Lamel, L.F., Kassel, R.H.; Seneff, S. (1986) Speech database development: design and analysis of the acoustic-phonetic corpus. *Proceedings DARPA Speech Recognition Workshop*, 100-109.

Lisker, L. (1957) Closure duration and the intervocalic voiced-voiceless distinction in English. *Language, 33*, 42-49.

Lisker, L.; Abramson, A.S. (1964) A cross-language study of voicing in initial stops: acoustical measurements. *Word* 20, 394-422.

Lisker, L. (1972) Stop duration and voicing in English. In *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*. (A. Valdman, ed.), Paris:Mouton, pp. 339-343.

Luce, P.A.; Charles-Luce, J. (1985) Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *JASA* , 78, 1949-1957.

Madebrink, R. (1955) The duration of the stops in the speech of deaf-mutes. *Folia Phoniatrica*, 7, 44-55.

Maddieson, I. (1984) *Patterns of Sounds.* Cambridge:Cambridge University Press.

Malécot, A. (1968) The force of articulation of American stops and fricatives as a function of position. *Phonetica*, 18, 95-102.

Ohala, J. (1983) The origin of sound patterns in vocal tract constraints. *The Production of Speech.* P. MacNeilage, ed., 189-216.

Pallet, D. (1990) Speech corpora and performance assessment in the DARPA SLS program. *Proceedings of the International Conference on Spoken Language Processing*.

Prosek, R.A.; House, A.S. (1975) Intraoral air pressure as a feedback cue in consonant production. *Journal of Speech and Hearing Research*, 18, 133-147.

Randolph, Mark. (1989) *Syllable-based Constraints on Properties of English Sounds.* MIT dissertation.

Scharf, D.J. (1962) Duration of post-stress intervocalic stops and preceding vowels. *Language and Speech*, 23(3), 297-307.

Slis, I.H. (1970) Articulatory measurements on voiced, voiceless and nasal consonants. *Phonetica* 21,193-210.

Slis, I.H. (1971) Articulatory effort and its durational and electromyographic correlates. *Phonetica* 23, 171-188.

Smith, B.L. (1978) Temporal aspects of English speech production: a developmental perspective. *Journal of Phonetics* 6.

Subtelny, J.D.; Worth J.H.; Sakuda, M. (1966) Intraoral pressure and rate of flow during speech. *Journal of Speech and Hearing Research* 16, 397-420.

Suen, C.Y.; Beddoes, M.P. (1974) The silent interval of stop consonants. *Language and Speech*, 17, 126-134.

Umeda, N. (1977) Consonant duration in American English. *JASA* 61, 846-858.

Zue, V.W. (1976) *Acoustic Characteristics of Stop Consonants: A Controlled Study.* Sc.D. Thesis (MIT, Cambridge, MA).

Zue, V.W.; Laferriere, M. (1979) Acoustic study of medial /t,d/ in American English. *JASA*, 66, 1039-1050.

Zue, V.W.; Seneff, S. (1988) Transcription and alignment of the TIMIT database. *Proceedings of the Second Meeting on Advanced Man-Machine Interface through Spoken Language.*

Zue, V.W.; Seneff, S.; Glass, J. (1990) Speech database development at MIT: TIMIT and beyond. *Speech Communication*, 9, 351-356, 1990.

Phonetic Units and American English [ow]

Kenneth de Jong

Introduction.

The present paper reports an investigation into the nature of phonetic primitives. It formulates a description of English mid back vowels which explains patterns of articulatory and acoustic behavior with different amounts of coarticulation. Two specific questions are addressed. First, what is the best way of describing the compound, diphthongal nature of English [ow]? And, second, what kind of primitives underly the dynamic action associated with diphthongs?

Traditionally, diphthongs are described as having two acoustic or articulatory vowel targets in succession. Gay (1968), having performed an a spectrographic analysis of American English diphthongs, described them as relatively invariant formant movements. That is, rather than consisting of two separate targets, diphthongs are better characterized as involving some formant movement in a particular direction at a particular rate in a particular area of the vowel space. Diphthongs, in this description, differ from other vowels in being inherently dynamic entities. Data in Kent (1970) supported this same conclusion.

Some recent theories of phonology and speech production -- specifically Gestural Phonology (Browman and Goldstein, 1986, 1989) and Task Dynamics (Saltzman, 1986; Saltzman and Munhall, 1990) have gone one step further in modeling speech as consisting entirely of dynamic units. In these models, the basic phonetic building blocks are articulatory gestures which are specified as having inherent dynamic properties (as argued for in Fowler, Rubin, Remez and Turvey, 1980). Each gesture, thus, not only includes a goal, a particular vocal tract configuration toward which the articulators move, but also an explicit time course for the movement. This is accomplished by using second order dynamic equations, such as that which describes the movement of an (over damped) spring mass system (Ostry, Keller, and Parush, 1983; Ostry and Munhall, 1985; Kelso, Vattekiotis-Bateson, Saltzman, and Kay, 1985; Browman and Goldstein, 1988, 1989). The time course of the movement is then usually specified as the stiffness term (k) in the dynamic equation.

Fundamental to these theories is the notion of a gesture. A gesture is the specification of speech activity in terms of abstract articulatory goals. Thus, for example, the specification of an alveolar consonant would not be stated in terms of muscle activity patterns, such as the contraction of the genio-glossus, nor in terms of actual articulator goals, such as advancing and raising the tongue away from the mandible, but rather in terms of constrictions along the vocal tract, in this case, the constriction of the oral cavity around the alveolar ridge. Gestures generally involve a complex of articulators which each can contribute to the attainment of the articulatory goal. Specifying gestures in terms of abstract articulatory goals allows task dynamics to capture the results of several studies which have shown compensatory behavior between jaw and tongue-tip in the attainment of alveolar constrictions (Kelso, Tuller, Vattekiotis-Bateson, and Fowler, 1984), between the jaw and the tongue body in the production of vocalic constrictions (Gay, Lindblom, and Lubker, 1981), and jaw and lip in the attainment of labial closures (Folkins and Abbs, 1975; Abbs, Gracco, and Cole, 1984). In each of these studies, perturbations of the position of one of the articulators (the jaw) was compensated for in the positioning of another articulator, which, provided the perturbation was not too large, allowed the speakers to obtain the articulatory goal.

An aspect of gestural theory that has been somewhat problematic, and has thus received a certain amount of attention is the specification of the temporal coordination between the various gestures which make up speech (see especially Browman and Goldstein, 1988, 1990). A recent approach is to specify the relative timing of speech gestures in terms of a phase angle, a value indicating how far through the execution of the gesture another articulatory event is to occur. Browman and Goldstein (1988), for example give a set of canonical phasing rules which give the temporal coordination of the vowel and neighboring consonants in a syllable. Regardless of the precise method of specifying the temporal relations between gestures, one of the most interesting aspects of gestural phonology is its ability to describe coarticulatory effects explicitly in terms of temporally overlapping gestures whose demands are "blended" using simple algebraic methods (Boyce, Krakow, Bell-Berti, and Gelfor, 1990; Munhall and Löfqvist, 1992). This technique has been used, for example, by Saltzman and Munhall (1991) to model the difference in the point of articulation of velar stops in the context of different vowels.

Considering Gay's (1968) results and description of diphthongs from the perspective of phonetic theories in which all phonetic primitives are dynamic suggests that the traditional description of diphthongs as consisting of two consecutive goals may be appropriate. Using gestural primitives, for example, diphthongs can be modeled as a complexes of gestures with relatively invariant relative timing relations. Gay's results -- a relatively invariant formant movement and direction -- can very likely be generated with a pair of gestures in sequence with a fixed timing between them. Two such gestures will give rise to a movement of the articulators from a point near the target configuration for the beginning of the diphthong and then a movement toward the offset of the diphthong at a relatively consistent rate of speed. Furthermore, if we model coarticulatory effects as arising from the temporal overlap of gestures and subsequent blending of the demands of the various gestures, coarticulatory effects will be most apparent at the edges of the diphthongal movement. That is, perseveratory effects (blending of the onset vowel with the previous consonant) will affect the starting position of the diphthongal movement, while anticipatory effects (blending of the offglide gesture with the following consonant) will affect the end point of the diphthongal movement, leaving a relatively unaffected -- relatively invariant -- portion in between the transition between the vowel and its off-glide.

This description of diphthongal structure demands especially two things. First, that speech be composed of a series of specifications with an explicit dynamic aspect. And second, that this dynamic aspect can be referenced in the timing of speech to yield fixed timing patterns. Such descriptions of phonetic structure are assumed in Gestural Phonology, and have been used to describe the structure of /s/ - stop syllable initial clusters in English, as well as pre-nasalized stops (Browman and Goldstein, 1986). In these descriptions, the phonetic entities are described as part of "gestural constellations," a combination of gestures with fixed temporal relations with one another. Such is the structure that would seem to be necessary here -- diphthongs comprise the class of vowels which are composed of gestural complexes in which oral components are non-synchronous. So , for example, the diphthong /ej/ might be described as oral opening and closing gestures and a palatal constriction gesture which are out of phase. That is, the palatal gesture is specified to occur during some phase of the closing gesture.

A related issue in phonetic theory concerns the nature of phonetic primitives. In gestural theory, the dynamics of speech are said to arise from the nature of phonetic primitives as gestures. The timing of acoustic events in speech fall out of the explicit specification of articulatory motions. The acoustic results of these motions, though they most certainly play a role in the evolution of a language over time (as noted in Goldstein, 1983, 1989) and are, of course, important for acquisition, are not needed in the specifications of a speaker's production system once the system has been acquired. To quote Browman and Goldstein (1989), "..gestural descriptions are purely articulatory," and Browman and Goldstein (1990), "[o]utput considerations do not ... appear to constrain actively the processes of variation in speech production." Similarly, the synthesis technique outlined by Saltzman and Munhall (1991) makes no use of the acoustic signal, neither in the specification of gestural primitives nor in the process of sequencing and combining the actions of the various articulators. Acoustic transformation is the last step in the process and has no feedback into the production system.

This approach differs from that taken by other speech researchers. For example, Nearey (1978) begins by saying that "the principal claim made by this work is that there is strong evidence that phonetic features specifying vowels are more directly related to acoustic rather than articulatory parameters." More particularly Perkell (1990) claims acoustic factors not only influence the choice of articulatory goals, but also the functioning of the system which seeks to attain the goals . He states,

> The programming of articulatory movements is a function of the sequence of articulatory goals, the moment-to-moment state of the vocal tract and the particular *acoustic* requirements of the individual utterances (p.264, emphasis mine)

and later,

> our approach also claims that some aspects of speech kinematics have to be determined on the basis of the communicative (acoustic) requirements of the sound sequences. (p. 266)

Thus, an alternative view to that of the present form of gestural phonology and task dynamics is that acoustic information actually is part of the goal of at least some gestures, and further also that acoustic requirements govern the manner in which the gestural primitives are to be coordinated, blended, and mapped onto articulator movements.

The present paper reports the results of an analysis of a corpus of X-ray microbeam records of three American English speakers producing the back vowel, [ow]. This diphthong is especially useful for two reasons. First, it, along with [ej] (the so-called "small diphthongs") showed the least amount of steady state of all of Gay's diphthongs. Second, it involves both dorso-velar and labial constrictions. While there is no obvious articulatory connection between these two activities, acoustic models of the vocal tract have long shown that dorso-velar and labial constrictions have similar depressing effects on the second formant (Chiba and Kajiyama, 1941; Stevens and House, 1953; Fant, 1960; Lindblom and Sundberg, 1971). Thus, if the production specification of the diphthong is purely articulatory in nature, the dorso-velar constriction and the labial constriction should be modeled as separate gestures -- two (possibly synchronized) components of a gestural constellation. One should, thus, not find the direct compensatory relationship between labial and dorsal activity that one finds between the jaw and the lip. Alternatively, if the dorsal and velar activities are better seen as manifestations of one single ("dorso-labial"; or to use an acoustic term, "grave") gesture, one would expect a compensatory relationship between the dorsal and velar constriction movements. That is, the motion of the lips should make up for any coarticulatory reduction of the tongue's contribution to a velar constriction.

Methods.

Procedure.

The Wisconsin X-ray microbeam system (Nadler, Abbs, and Fujimura, 1987) was used to record the movement trajectories of gold pellets attached to the tongue, jaw and lips of three American English speakers while producing monosyllabic words containing the vowel, [ow]. Three 2.5 mm pellets were attached to tongue at distances of 5, 15, and 25 mm. from the tongue apex, as measured with the tongue extended. Two 3 mm pellets indexed jaw movement, one placed on a mandible molar, one placed at the base of the mandible incisors. Two 3 mm. pellets indexed the movement of the upper and lower lip. Since a dental adhesive was used to fix the pellets to the articulators, the labial pellets could not be attached to smooth inferior surface of the upper lip and superior surface of the lower lip, but rather were attached to the vermilion border of both lips. Since labial movements tend to be quite small, and the mechanical connection between the various parts of the lips may be quite complicated, this positioning of the pellets may not index the motion of the internal surfaces of the lips very well. The tongue pellets and the lower lip pellets were sampled at 100Hz, while the other, slower moving pellets, were sampled at 50Hz After acquisition, the data was smoothed and oriented so that the x-dimension is perpendicular to the occlusal plane.

The subjects were three speakers of a northern midwestern dialect of American English. Each were undergraduate students at the University of Wisconsin who were paid for their participation in the experiment. Two subjects were male, SD and MB; one was female, TG. Given the difference in size of the male and the female subjects, the tongue pellets probably indexed a larger proportion of the female's tongue surface than the males'.

The words containing the diphthong were all monosyllables beginning with /t/ and ending with zero to three alveolar consonants. Thus the target words were *toe, toes, toad, toads, tote, totes, toast,* and *toasts.*, occurring in a frame sentence of the form, "I said, 'Put the _______ on the table.'" In addition to varying the coda consonants, the amount of stress placed the word was varied by giving the subjects miniature discourse settings in which the subjects were to respond to an imaginary hearer who misheard part of the sentence. The subjects consistently responded with nuclear (sentence) accent on the misheard portion. There were three elicitation conditions, nuclear accent on the target word, nuclear accent on the following *on* (in which subjects consistently placed a prenuclear pitch accent on the target word), and nuclear accent on *put* (which precludes placement of an accent on the target word). Thus, to obtain an utterance with the word *toad* and nuclear accent on the preposition, *on*, the subjects were given the following cue:

-- Did you say, "Put the toad under the table?"
-- I said, "Put the toad ON the table."

The subjects then said the second line of the dialogue. They were given a chance to practice the sentences before the pellets were attached.

The stress pattern of each of the utterances was verified by the experimenter performing an intonational analysis using the transcriptional system described in Pierrehumbert (1980) and developed in Pierrehumbert and Beckman (1988). The words in these three elicitation conditions were repeated six times in a random order by TG, and four times by SD and MB. This yielded the 8 target words by three levels of stress (nuclear accented, pre-nuclear accented, and unaccented), totaling 144 utterances for TG

TABLE 1. Number of utterances in each stress category

SUBJECT:	MB	SD	TG
Unaccented:	29	32	48
Prenuclear accented:	32	30	47
Nuclear accented:	29	31	47

and 96 for the other two subjects. Due to various mechanical and software problems in recording and porting the data onto PC's a small number of the tokens for each subject were rendered unanalyzable. The total number of tokens in each stress condition which were analyzed are given in Table 1.

Articulatory Analyses.

Articulatory analyses were performed using the XD program developed by Joan Miller implemented on a PC under MS-DOS. After inspecting a number of tongue pellet trajectories displayed in the sagittal plane, a criterion was set for marking dorsal events corresponding to the [o] and the [w] off-glide. Points chosen were ones at which the tongue dorsum began moving in both a posterior and superior direction (dorsal position for [o]), and the point at which that movement ended -- in this case, began moving in an anterior direction (dorsal position for [w]; see Figure 1). As is discussed below, there were two difficulties with the criteria. First the subjects did not always show the posterior-superior movement expected of the diphthong. In these cases a single point indicating the extremum in a posterior and inferior direction (defined as a 45 degree diagonal) was taken for both measurements (see top panel of Figure 1). Subjects TG and SD often posed another problem in that their dorsal trajectories did not always have a downward movement out of the onset [t]. Thus the first measurement would occur during the closure of the [t]. However, comparing cases with and without inferior motion out of the [t] showed that an inflection point in the superior-posterior movement in tokens without downward motion corresponded to the most inferior point in the tokens with a clear downward movement out of the [t]. This small dip, then, was taken as the dorsal position for [o].

In addition, two labial maxima were marked as well; one in the horizontal dimension of the upper lip pellet, and a second in the difference between the horizontal dimension of the lower lip pellet and the mandible incisor pellet. The incisor position was subtracted from the lower lip position to factor out the effect of jaw opening on the horizontal position of the lower lip.

Finally, to get a relatively clear index of the timing of the articulation of the consonants surrounding the vowel, two events in the movement of the tongue tip were noted as well. The tongue tip trajectories were rotated so that the new y-dimension is roughly perpendicular to the alveolar ridge. A velocity trace was generated from the new y-dimension tip trace, and the tip minimum position was extracted, along with the velocity maximum as the tip moved toward closure during the coda consonants. This technique is described in greater detail in deJong (1991a).

Acoustic Analyses.

In addition to articulatory measurements, the time-aligned acoustic records were submitted to an acoustic analysis to determine how the timing of the articulatory events corresponds to acoustic events noted in the literature. In addition, acoustic analyses were used to obtain some measure of the effect of the positioning of the different articulators on the acoustic signal.

Acoustic measurements were made using the CSL developed by Kay Elemetrics implemented on a PC under MS-DOS. Two times were measured from a spectrographic display generated with an effective bandwidth of 300Hz, the timing of the f1 maximum, and the subsequent timing of the f2 minimum. All tokens had an f1 maximum and a subsequent f2 minimum. In addition, f1 and f2 values at these times were obtained using a 12th order LPC analysis.

Results 1: Timing of articulatory and acoustic behavior.

Considering first the tongue dorsum movements, the most dramatic aspect of the present corpus is the amount of variation in the dorsal movement trajectories, especially for MB. Figure 1 plots four representative trajectories. In the upper token, there is no apparent diphthongal movement. The dorsum simply lowers for the vowel and then advances for the following consonant. At the bottom, however, is an example in which the tongue dorsum shows a lowering followed by an extensive diphthongal movement toward a velar constriction for the [w] off-glide.

The most straightforward interpretation of these differences would be a difference between a diphthong and a monophthong, between [ow] and [o]. That is, the off-glide has simply been deleted in the upper case. However, there are several reasons to reject this analysis. The first is apparent in an inspection of the rest of Figure 1. Between tokens with an extensive diphthongal movement and those with no diphthongal movement are a series of tokens with shorter and shorter diphthongal movements, illustrated here with two steps. Thus, the effect here cannot be adequately described as a categorical difference in the vowel produced, but rather seems to indicate an effect of a mapping of phonological elements onto continuously variable phonetic behavior.

One approach to describing this effect, given the gestural framework outlined above, would be to blend the [o] and the [w] gestures more and more until eventually arriving at one vocalic movement towards a position, higher than an [o], but lower than an [w]. Comparing the position of the dorsum during the [o] and [w] suggests, however, that if the trajectory differences are due to a blending of the vocalic gestures, the blending is strongly weighted toward the [o]. Figure 2 plots the position of the dorsum during [o] and [w]. Tokens are divided into three roughly equal groups by the duration of the diphthongal movement. As one can see, the truncation of the diphthongal movement involves mostly the realization of the velar constriction for the off-glide. The other two subjects showed the same pattern; the dorsum retracted less for the [w], while the position of the dorsum for [o] remained relatively unaffected in diphthongal movements of shorter duration.

A more effective interpretation of these patterns is that the shortening of the diphthongal gesture is an effect of the following alveolar consonants coming earlier relative to the vowel onset. With this interpretation, one would expect the coarticulatory effects of the alveolar consonants to be apparent especially in the tongue movements, not,

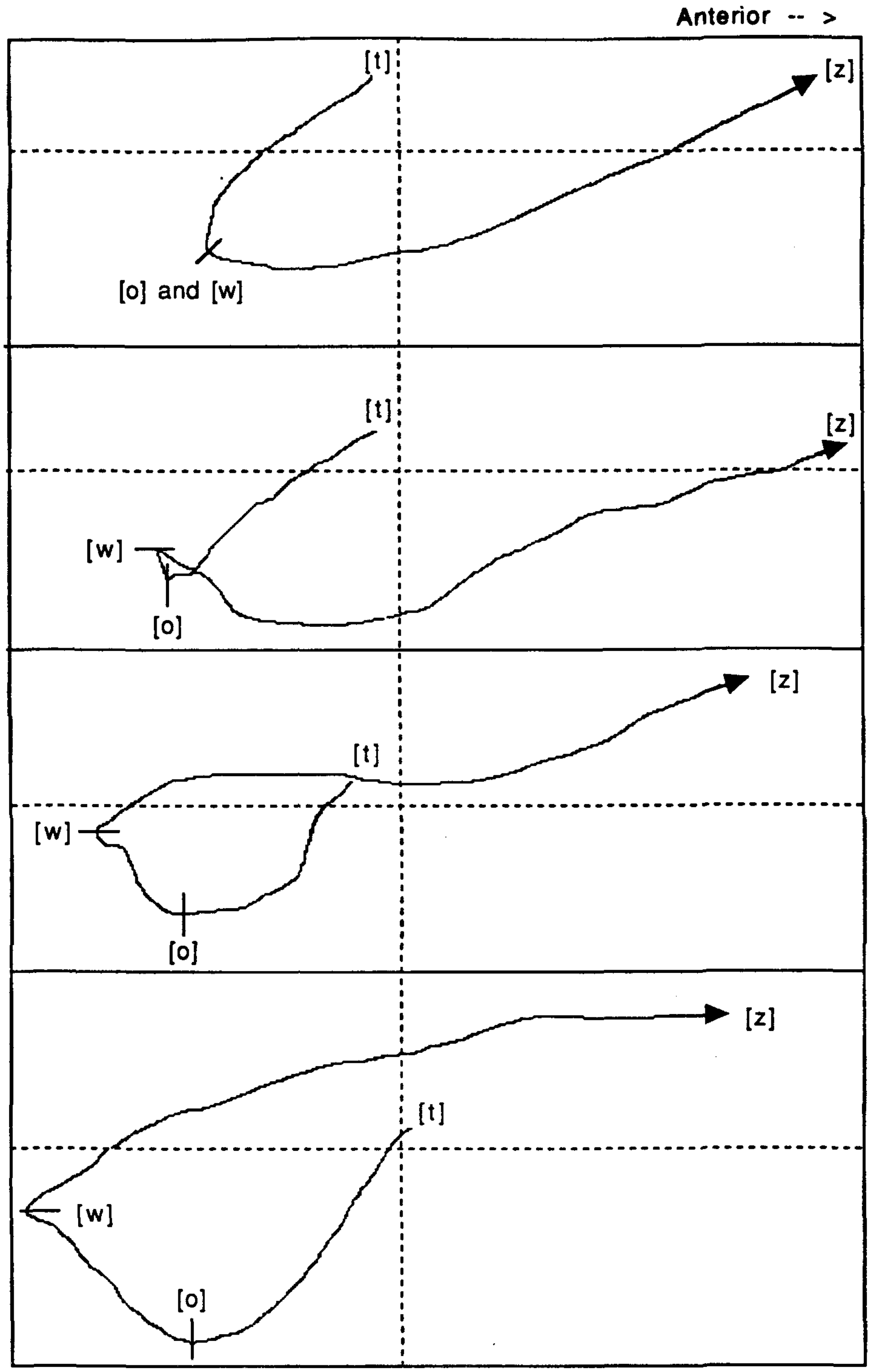


Figure 1. Four dorsal pellet trajectories from subject MB of the word, *toes*. Trajectories show dorsal flesh point movement from the position for [t] to the extreme position for [z]. Also shown here are example measurements of the dorsal position in [o] and [w] discussed in the text.

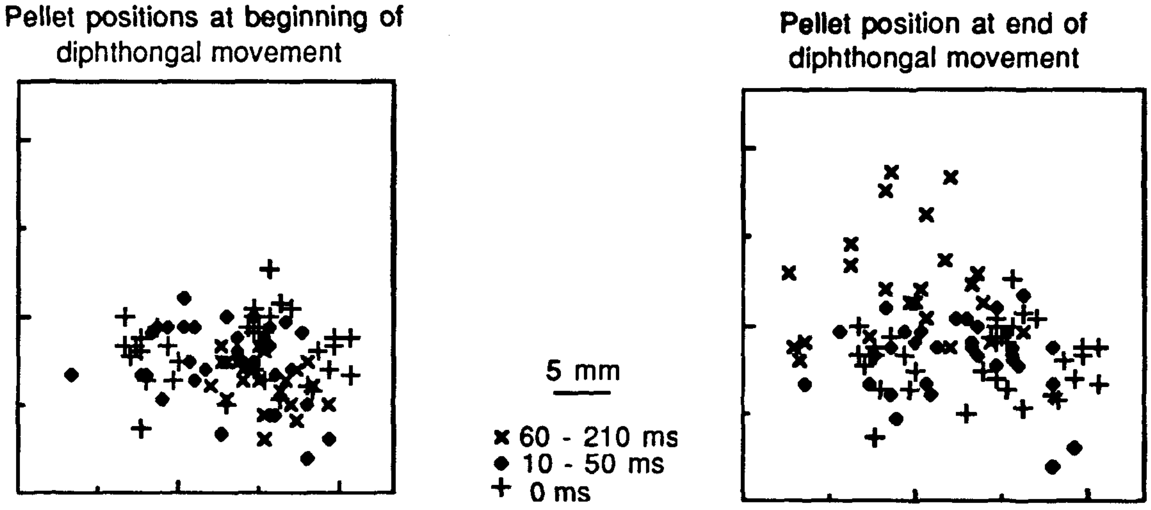


Figure 2. Tongue dorsum pellet positions at the beginning and end of the dorsal diphthongal movement for Subject MB. Tokens are coded by the duration of the diphthongal movement.

for example, in the lip rounding movements also specified as part of the [w] off-glide. The first piece of evidence that this is correct is that there always is a positive f2 minimum latency from the f1 maximum. That is, even though there may be no dorsal diphthongal movement, there is consistently an acoustic diphthongal movement.

Figure 3 plots the duration of the acoustic diphthongal movement against the latency of the beginning and end of the movement from the dorsal diphthongal movement. As Figure 3 shows, the f1 maximum does correspond in time to the dorsal event mark as corresponding to [o], at least for subjects MB and SD. Thus, the articulatory event which corresponds in time to the f2 minimum cannot always be the dorso-velar constriction maximum. This is pointed out even more clearly in Figure 4 which plots the latency of the f2 minimum from the dorsal [o] against latency of the velar constriction maximum from the dorsal [o]. In cases in which there is a long dorsal diphthongal movement (to the right), the f2 minimum does correspond to the dorso-velar constriction. In many cases where the diphthongal movement into the velar constriction has been shortened, the f2 minimum comes considerably later than the velar constriction. Figure 5 suggests why this is. At least for SD and MB, the cases in which the diphthongal movement has been shortened are those cases in which the labial protrusion maxima come late with respect to the dorsal velar constriction. Thus, the temporal deviation of the f2 minima from the velar constrictions is strongly correlated with the deviation of the labial protrusion maxima from the velar constrictions. In cases in which the dorsal diphthongal movement towards a velar constriction is shortened, the labial gestures remain behind, causing a lowering of the f2 for the [w]. Thus, it seems that the disappearance of the dorsal diphthongal movement is the result of coarticulatory activity centered in the tongue; coarticulation of the diphthong with the following lingual consonants.

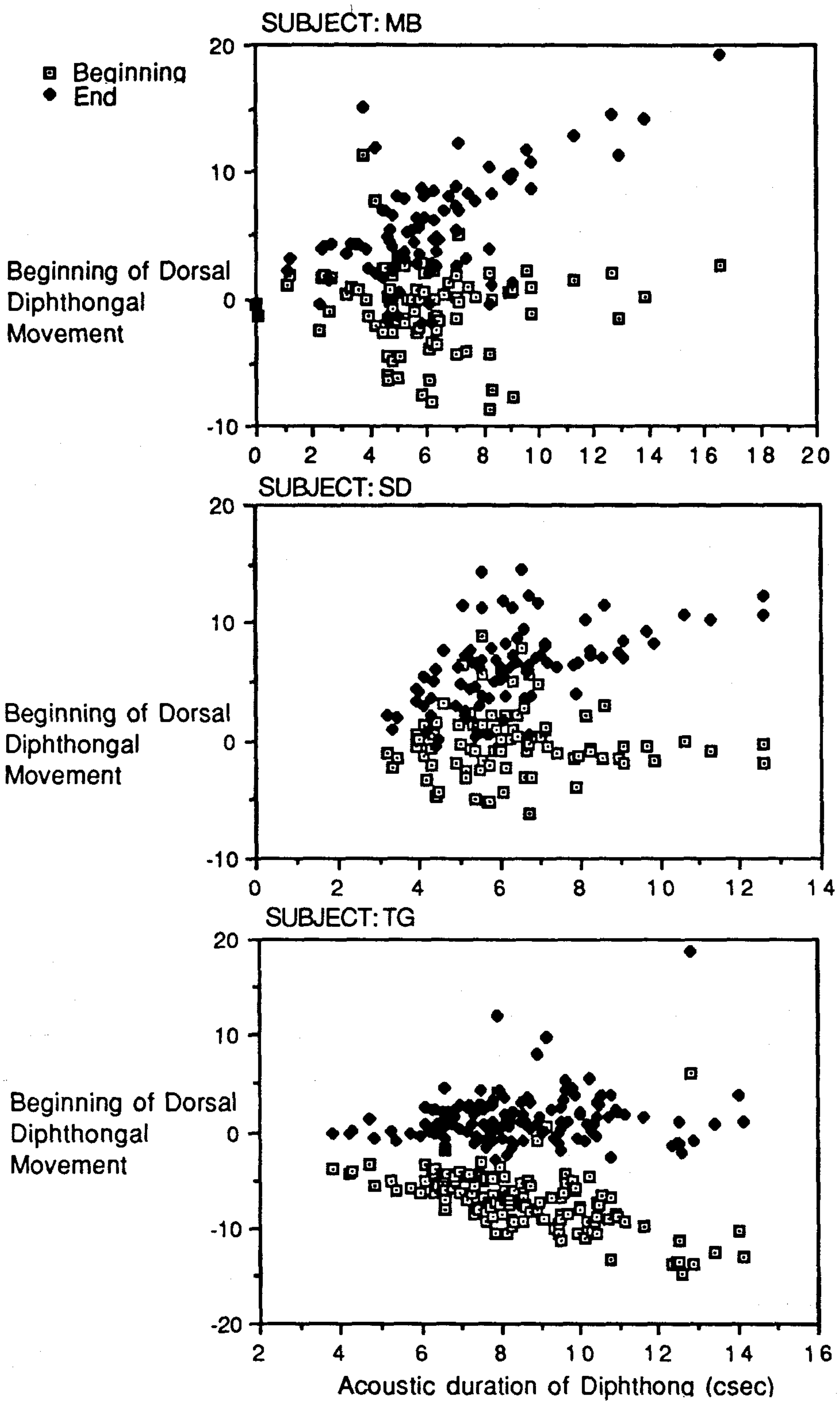
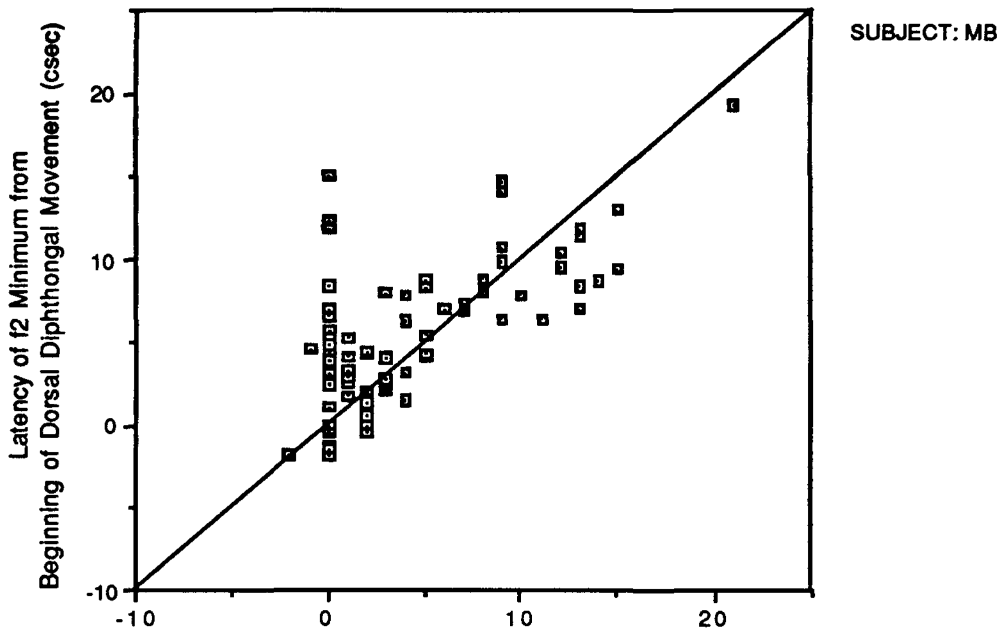


Figure 3. The acoustic duration of [ow] (time of f2 minimum minus time of f1 maximum) plotted against the latency of the beginning and end points from the dorsal measure of the beginning of the diphthong. Thus 0 on the horizontal axis indicates the timing of the minimum position of the dorsal pellet and time proceeds upward.



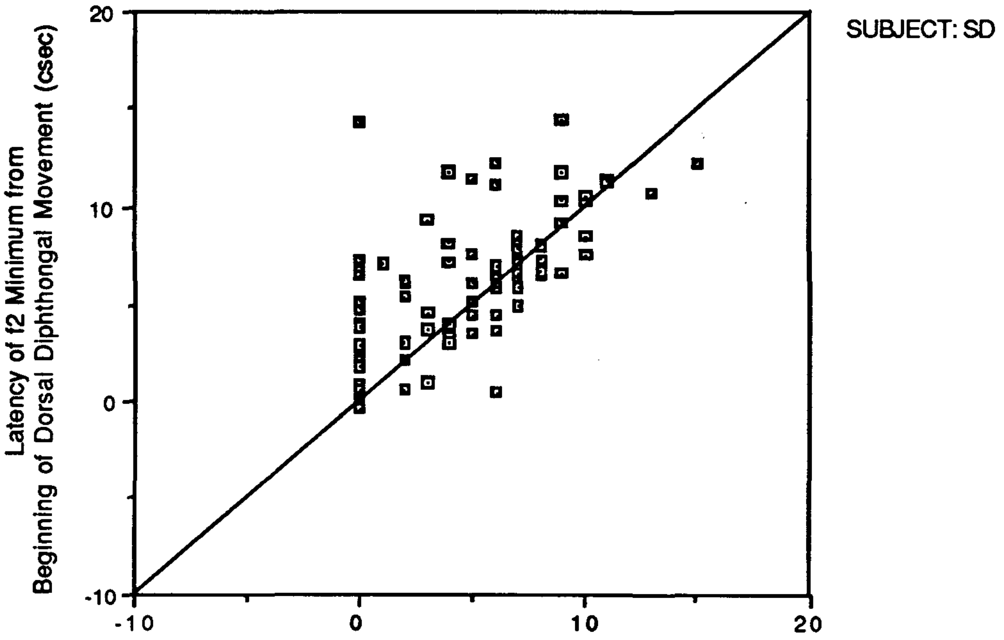


Figure 4 (cont'd next page).

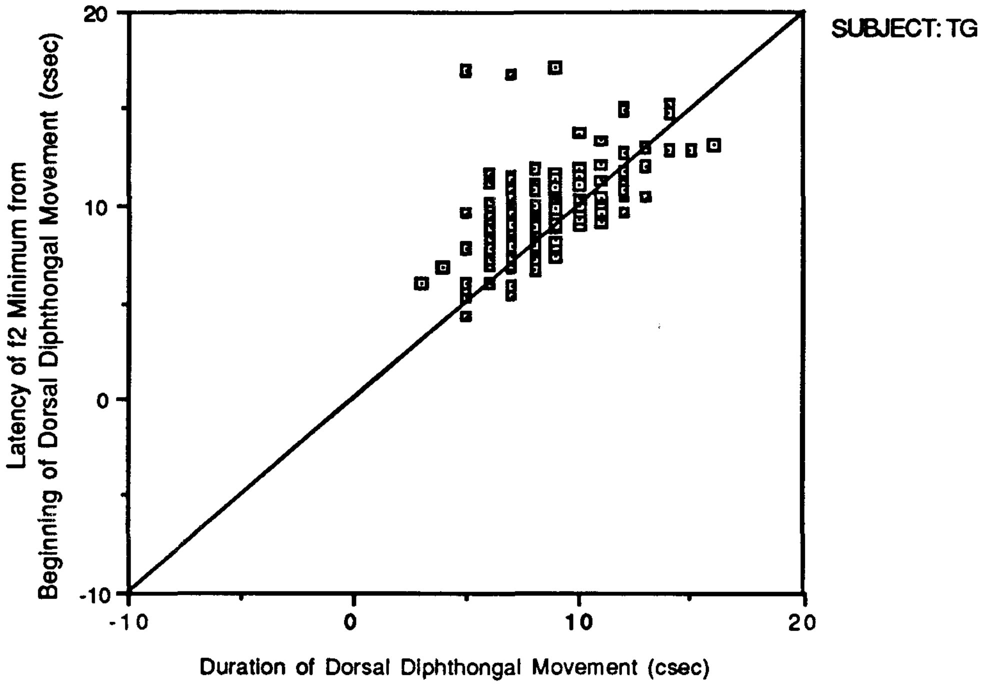


Figure 4. The latency of the f2 minimum from the dorsal minimum (beginning of dorsal diphthongal movement) plotted against the latency of the maximum velar constriction from the dorsal minimum.

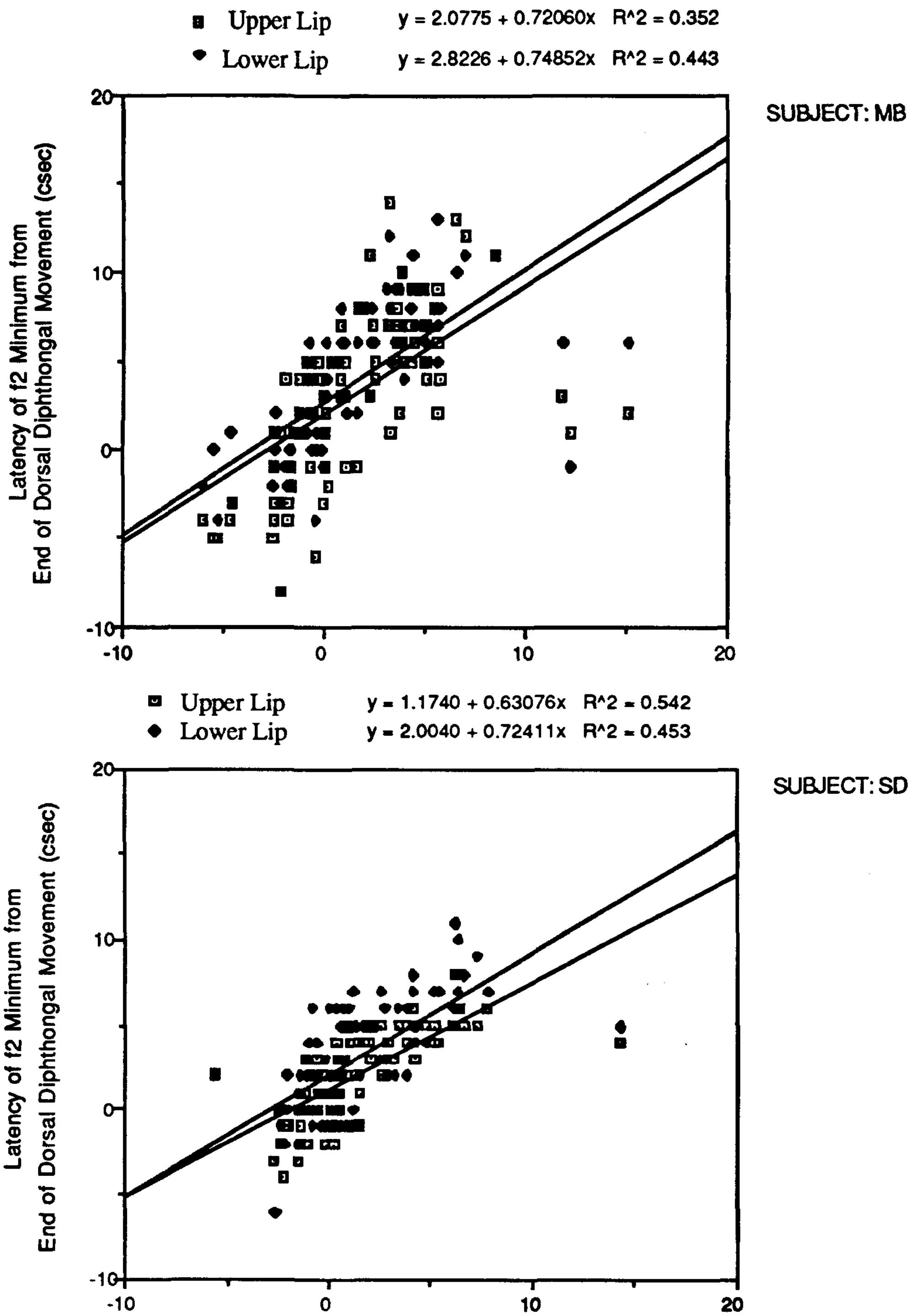


Figure 5 (cont'd next page).

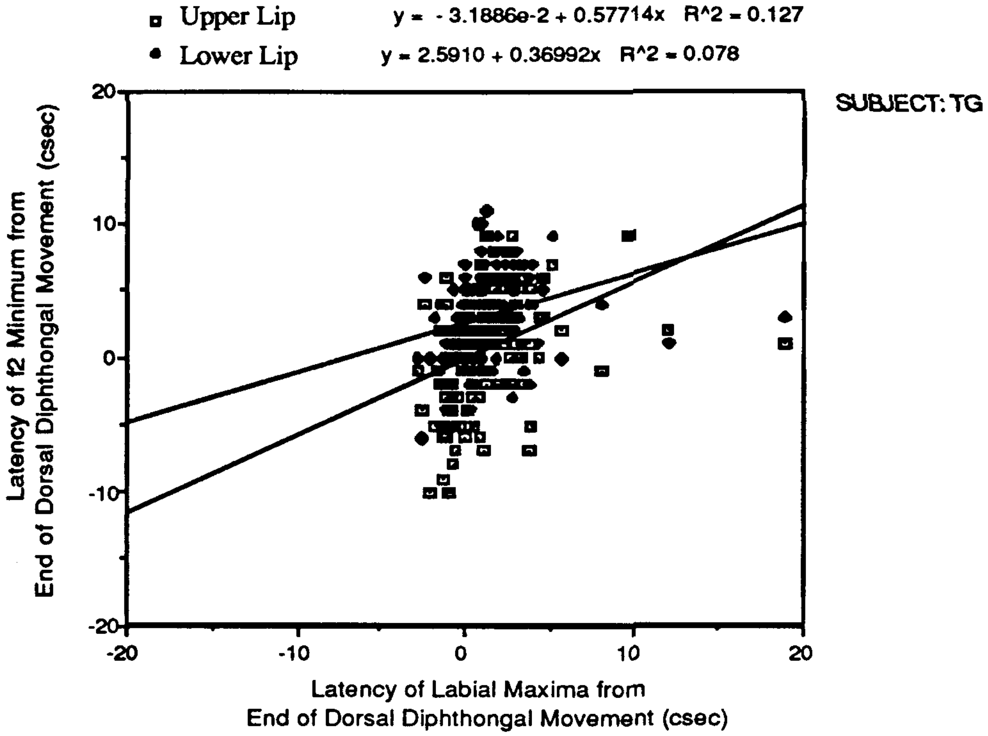
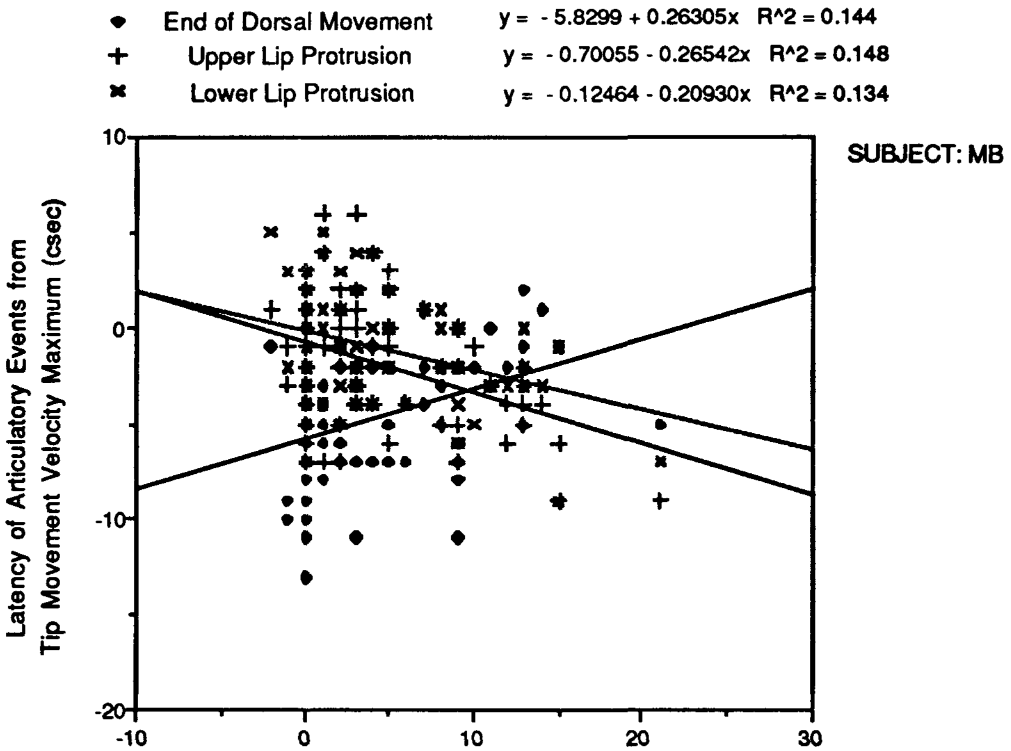


Figure 5. The deviation of the timing of the f2 minimum from the velar constriction maximum plotted against the deviation of the timing of the labial protrusion maxima from the velar constriction maximum.

One final piece of evidence also suggests this interpretation. Different degrees of gestural overlap would cause differences in the relative timing of the tongue tip events (an index of the timing of the consonants) and the dorsum and lip events. If the alveolar consonant specification and a hard mechanical connection between the tongue tip and tongue dorsum are causing the shortening of the dorsal diphthongal movement, one would expect a time locking of the velar constriction with tongue tip events, such as the moment of peak upward velocity for the alveolar coda consonants. By contrast, the labial protrusion maxima should occur later with respect to the tip events in cases where the consonant gesture comes earlier with respect to the vowel.

Figure 6 shows some evidence that this scenario is, at least, partially correct. Figure 6 plots the timing of the velar constrictions and labial protrusion maxima with respect to the tip velocity maxima against the duration of the dorsal diphthongal movement. As with Figure 3, time proceeds upward. The overlap hypothesis predicts a negative correlation between labial timing with respect to consonant articulation and diphthong duration -- short diphthongs associated with late labial maxima. No correlation is expected for dorsal timing, since the timing of the velar constriction is supposed to be caused by the timing of the consonant articulation. For all three subjects, the same general pattern obtains. Labial maxima tend to come later when there's greater coarticulation of the consonants and the vowel, while the velar constriction maxima tend to come earlier. The labial patterns are expected, while the dorsal patterns are not. The earlier velar constriction maxima probably indicate that the mechanical coupling between the dorsum and tip is far more complicated than is assumed above.



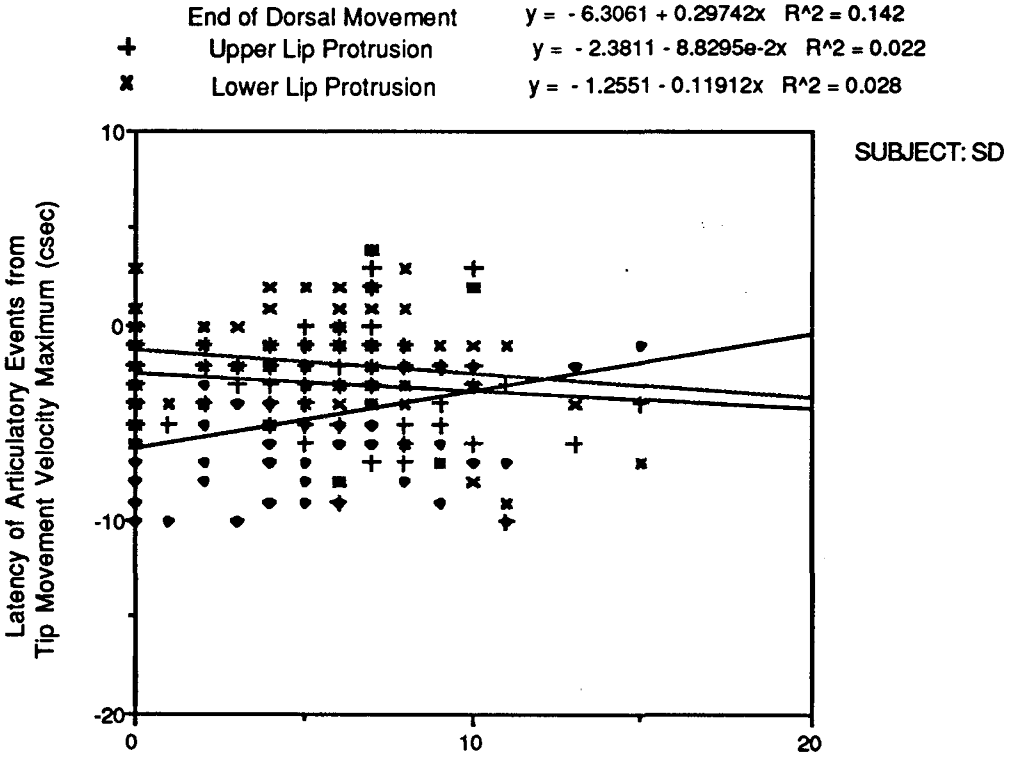


Figure 6 (cont'd next page).

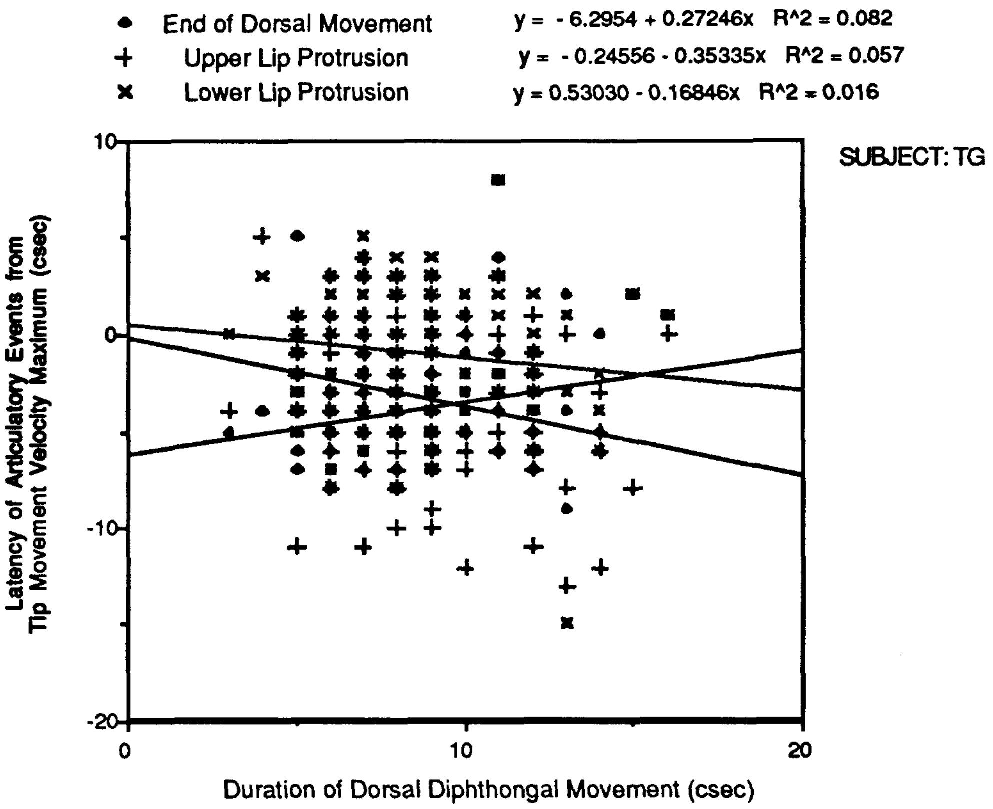


Figure 6. The latency of articulatory events associated with [w] from the moment of tongue tip movement velocity (as an index of consonant articulation) plotted against the duration of the dorsal diphthongal movement.

Results 2: Spatial relationships between articulators.

The present study has discovered a situation in which the results of coarticulatory overlap on the acoustic signal are mitigated by articulatory redundancy. Specifically, even though coproduction of the velar off-glide with neighboring alveolar consonants strongly affects the movement of the tongue body, the added degree of articulatory freedom afforded by the specification of roundness allows the maintenance of at least some of the acoustic correlates of the [w] off-glide. So far, however, no evidence determines whether this is just a property of the gestural inventory of English, or whether there is some aspect of the functioning of the production system itself which contributes to the maintenance of particular acoustic goals. Specifically, should the velar and the labial motions best be seen as the manifestation of one (super) gesture aimed at an acoustic effect, or should they be seen as the manifestation of two separate gestures, which are specified to occur simultaneously?

If the dorsal and labial movements are best described as two gestures, one would expect no on line compensation between the dorsal and labial gestures, since they comprise separate, but simultaneous gestures. If (at least some) gestures should be specified in terms of the acoustic output of the gestures, one might expect the labial and dorsal activities to be the reflex of a single, dorso-labial gesture. If this is the case, they should compensate for perturbations of one another. In the present case, the dorsal component of the diphthong is affected rather strongly by coarticulation with the following alveolar consonants, while the labial component is free to be manifested in rounding at the appropriate time following the articulation of the [o]. Lip rounding should be increased in those cases in which the tightness of the velar constriction has been reduced by coarticulation.

One confounding factor, however, must be controlled. If one varies the amount of stress on the tokens, one would expect a positive correlation between the amount of velar constriction and the amount of lip protrusion. Here I assume the model of stress presented in deJong (1991b) and in deJong, Beckman and Edwards (to appear), that stress is phonetically realized as a hyperarticulation localized to the syllable. In this model stress enhances phonemic distinctions within the stressed syllable. The presence of accent should increase the magnitude of articulations leading to distinctions -- in this case the dorsal retraction movement and the labial protrusion movement. Thus, across stress conditions one would expect a positive correlation between the amount of labial protrusion and the amount of dorsal retraction.

Figure 7 plots the most protruded horizontal position of the upper lip pellet against the most retracted horizontal position of the dorsal pellet for [w]. The tokens are separated by the amount of stress on them. Since anterior is to the right in this figure, dorso-labial compensation would result in positive correlations, since tokens with less retraction would have more anterior dorsal positions and also more anterior lip pellet positions. The results vary from subject to subject. MB shows the positive correlations expected, while SD has a negative correlation. TG shows no relationship between the two measures at all. The results of similar correlations with the lower lip and with another measure of dorsal position (vertical position of dorsum minus horizontal position), are given in Table 2. The pattern of results across subjects obtains regardless of which articulatory measures are paired up, thus the pattern shown in Figure 7 is not idiosyncratic to the articulatory measures used.

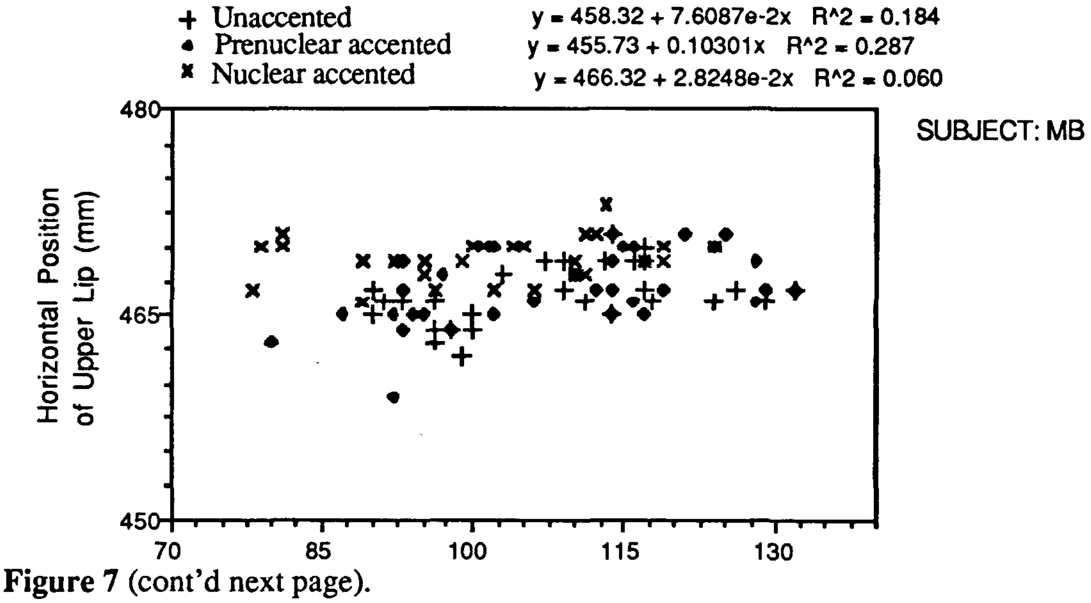


Figure 7 (cont'd next page).

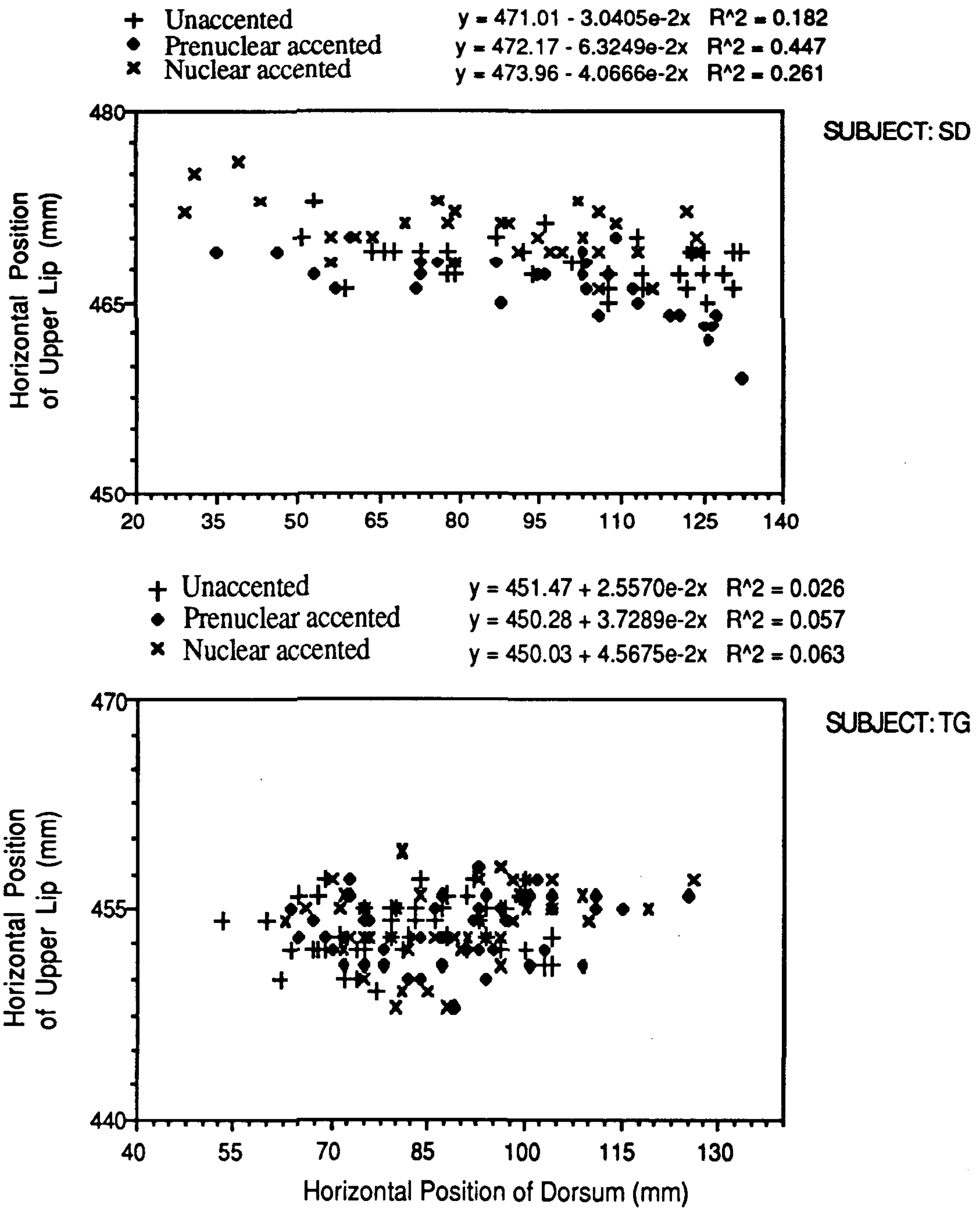


Figure 7. The most protruded position of the upper lip pellet plotted against the most retracted position of the tongue dorsum pellet. Dorso-labial compensation should be manifested in a positive correlation between the two positions.

TABLE 2. Correlations between dorsal and labial dimensions.

SUBJECT:	MB		SD		TG	
	r^2	m	r^2	m	r^2	m
ALL TOKENS						
Horizontal position of dorsum X						
Upper lip - horiz.	0.072	+0.05	0.251	-0.05	0.050	+0.04
Upper lip - vert.	0.062	+0.05	0.004	-0.01	0.056	-0.06
Lower lip - horiz.	0.185	+0.09	0.002	+0.02	0.000	0.00
Diagonal position of dorsum X						
Upper lip - horiz.	0.023	-0.05	0.249	+0.05	0.056	-0.04
Upper lip - vert.	0.032	-0.03	0.026	0.00	0.006	+0.01
Lower lip - horiz.	0.064	-0.04	0.000	0.00	0.009	+0.01
UNACCENTED TOKENS						
Horizontal position of dorsum X						
Upper lip - horiz.	0.184	+0.08	0.182	-0.03	0.026	+0.03
Upper lip - vert.	0.002	+0.01	0.002	-0.01	0.149	-0.09
Lower lip - horiz.	0.231	+0.11	0.264	-0.04	0.042	-0.03
(with reference to incisor)						
Diagonal position of dorsum X						
Upper lip - horiz.	0.036	-0.03	0.142	+0.02	0.032	-0.03
Upper lip - vert.	0.025	+0.03	0.003	0.00	0.010	+0.02
Lower lip - horiz.	0.060	-0.05	0.203	+0.03	0.064	+0.03
PRENUCLEAR ACCENTED TOKENS						
Horizontal position of dorsum X						
Upper lip - horiz.	0.287	+0.10	0.447	-0.06	0.057	+0.04
Upper lip - vert.	0.113	0.00	0.022	+0.02	0.013	-0.02
Lower lip - horiz.	0.355	+0.13	0.404	-0.05	0.014	+0.01
Diagonal position of dorsum X						
Upper lip - horiz.	0.243	-0.09	0.520	+0.06	0.128	-0.05
Upper lip - vert.	0.025	-0.02	0.030	-0.02	0.000	0.00
Lower lip - horiz.	0.214	-0.09	0.434	+0.05	0.000	0.00
NUCLEAR ACCENTED TOKENS						
Horizontal position of dorsum X						
Upper lip - horiz.	0.060	+0.03	0.261	-0.04	0.063	+0.05
Upper lip - vert.	0.128	+0.07	0.078	-0.03	0.082	-0.09
Lower lip - horiz.	0.125	+0.06	0.034	+0.09	0.001	0.00
Diagonal position of dorsum X						
Upper lip - horiz.	0.252	-0.04	0.231	+0.03	0.040	-0.04
Upper lip - vert.	0.161	-0.06	0.043	+0.02	0.039	+0.06
Lower lip - horiz.	0.164	-0.05	0.010	-0.04	0.003	+0.01

Discussion and Conclusion.

The present study has two general results. The first is that diphthongs can be described as gestural complexes with relatively fixed temporal coordination between their component gestures. In the present case, the relative timing of the vowel and coda consonants is found to vary considerably more than the relative timing of the vowel and its diphthongal offglide. Evidence for this comes from the effects of coarticulation of the following alveolar consonants with the offglide, which vary considerably, having a marked effect on the motion of the tongue body. This effect is somewhat mitigated by the labial component of the off-glide, which remains distant from the vowel nucleus, shifting "over" the movements for the following consonant.

Given this pattern of results, it is reasonable to ask what is the best way of describing the timing relationship of the various gestures to one another. It is clear that the timing relationships between all of the gestures are not fixed (as suggested by the kinds of canonical phasing rules written in Browman and Goldstein, 1988). This is evident indirectly in the present study -- it is difficult to see how the change in the dorsal trajectories for [ow] can be derived other than by gestural blending which is the result of continuous changes in the relative timing of the gestures. Timing changes, especially consonant-vowel timing changes have been explicitly tested for and found by Beckman, Edwards and Fletcher (1991), and using different techniques in the present database by deJong (1991a).

This being the case, a somewhat different general model of timing may be appropriate from the perspective of the phonological and phonetic specifications of a language. Intergestural timing may be linguistically unspecified, but constrained by the prosodic closeness of the units being articulated. Claimed instances of gestural cohesion (relative invariance in timing relationship in various conditions, as in Saltzman and Munhall, 1990) include the [s] - stop clusters and pre-nasalized stops mentioned above (see Browman and Goldstein, 1986). The temporal coordination between laryngeal and oral gestures in the production of voiceless obstruents is also remarkably consistent (Lőfqvist, 1980; Lőfqvist and Yoshioka, 1981, 1984). In these cases, the gestural cohesion is so strong, perturbations of the oral gesture which cause a delay in the attainment of the obstruent closure also cause a delay in the glottal gesture (Munhall, Lőfqvist, and Kelso, 1986; also Saltzman, Löfqvist, Kinsella-Shaw, Rubin, and Kay, 1992). Each of these cases involves gestural entities which share the onset of the syllable. The cohering gestures are closely bound in the prosodic structure of the utterance. Similarly in the present study, the gestures comprising the [o], indexed here by the dorsal minimum, and those comprising the [w], indexed here by dorsal retraction and labial protrusion, can be seen as inhabiting the nucleus of the syllable together (see Kaye and Lowenstamm, 1984, for phonological arguments to this effect). Thus, the various components of the syllable, onset, nucleus, coda, and possibly rhyme as well, can be seen as production units within which the timing pattern is relatively fixed. Timing relations between such units are relatively free to vary, and thus are good candidates for the implementation of stress contrasts (as found by Edwards, Beckman, and Fletcher, 1991), and also for certain lexical contrasts such as that between voiceless and voiced obstruents in English (deJong, 1991a).

The second result of the present study is that the task dynamic framework should be modified in two ways. The first involves the coarticulatory effects of the alveolar consonants on the articulation of the diphthongal offglide. The task dynamic framework as presented in Saltzman and Munhall, (1990) involves two mappings from the underlying gestural units to the actual articulatory motions. The first compiles the individual gestures into a common gestural score which expresses a temporally aligned series of abstract articulatory goals. At this stage, tuning takes place. Tuning is the process of blending together incommensurate goals, such as the demands of a velar stop and a palatal vowel on the goal positioning of the tongue body. Technically, the effect of the alveolar consonants on the articulation of the diphthongal offglide shown in the present study would not take place during this mapping, since the [w] and the alveolar consonants would be specified on separate articulatory channels -- the tongue body and the tongue tip. The results of the present study suggest that the conditions under which tuning must be executed should be broadened to include such cases as extended examples of incommensurate demands on the same articulator (the tongue body in this case -- see also Lindblom, Pauli, and Sundberg, 1974, for explicit modeling of the coupling between the tongue body and the tip).

A second possible change may have to be made as well to be able to capture the dorso-labial compensation effects shown by the one subject. Presumably, this should be done during the second mapping -- that which goes from the (tuned) abstract articulatory goals to the articulators themselves. It is in this mapping that the compensatory behavior of the jaw and other articulators is simulated. Capturing dorso-labial compensation would simply entail the addition of an abstract articulatory goal which can be attained by the tongue body and jaw complex as well as the lip and jaw complex, a bit of a complication in dimensionality of the present model.

It seems quite possible to extend the present task dynamic framework in an ad hoc fashion to cover the extra dimension of articulatory coordination which links labial and velar activity. However, the theoretical point should not be missed, the coordinative structures which at least one language learner in this experiment set up is constructed around relationships defined not only in terms of physical couplings between the articulatory structures, but also acoustic couplings between various roughly equivalent vocal tract configurations. As such, output considerations can affect the on-line functioning of the speech production system.

References.

Abbs, J.H., Gracco, V.L. and Cole, K.J. (1984), "Control of multimovement coordination: Sensorimotor mechanisms in speech motor programming, *Journal of Motor Behavior*, 16(2): 195-231.

Boyce, S.E., Krakow, R.A., Bell-Berti, F., and Gelfor, C.E. (1990), Converging sources of evidence for dissecting articulatory movements into core gestures, *Journal of Phonetics*, 18: 173-188.

Browman, C., and Goldstein, L. (1986), Towards an articulatory phonology, *Phonology Yearbook*, 3: 219-252.

Browman, C., and Goldstein, L. (1988), Some notes on syllable structure in articulatory phonology, *Phonetica*, 45: 140-155.

Browman, C., and Goldstein, L. (1989), Tiers in articulatory phonology: Some implications for casual speech. In *Papers in Laboratory Phonology: Between the Grammar and the Physics of Speech* (J. Kingston and M.E. Beckman, editors), pp. 341 - 376. Cambridge: Cambridge University Press.

Browman, C., and Goldstein, L. (1990), Gestural specifications using dynamically-defined articulatory gestures, *Journal of Phonetics*, 18: 299-320.

Chiba and Kajiyama (1941). *The Vowel: Its Nature and Structure.* Tokyo.

deJong, K.J. (1991a), An articulatory study of consonant-induced vowel duration changes in English, *Phonetica*, 48: 1-17.

deJong, K.J. (1991b). *The Oral Articulation of English Stress Accent*, unpublished doctoral dissertation, Ohio State University.

deJong, K.J. (to appear), The interplay of prosody and coarticulation, *Language and Speech.*

Edwards, J.R., Beckman, M.E., and Fletcher, J. (1991), The articulatory kinematics of final lengthening, *Journal of the Acoustical Society of America*, 89: 369-382.

Fant, G. (1960). *The Acoustic Theory of Speech Production*. The Hague: Mouton.

Folkins, J.W., and Abbs, J.H. (1975), Lip and jaw motor control during speech: Responses to resistive loading of the jaw. *Journal of Speech and Hearing Research*, 18: 207-220.

Fowler, C.A., Rubin, P., Remez, and Turvey (1980), Implications for speech production of a general theory of action. In *Language Production, Vol. 1: Speech and Talk* (B. Butterworth, editor), pp. 373-420. New York: Academic Press.

Gay, T. (1968), Effect of speaking rate on diphthong formant movements, *Journal of the Acoustical Society of America*, 44: 1570-1573.

Gay, T., Lindblom, B.E.F., and Lubker, J. (1981), Production of bite-block vowels: Acoustic equivalence by selective compensation, *Journal of the Acoustical Society of America*, 69(3): 802-810.

Goldstein, L. (1983), Vowel shifts and articulatory-acoustic relations. In *Abstracts of the Tenth International Congress of Phonetic Sciences, Utrecht., 1 - 6 August, 1983* (A. Cohen and M.P.V. van den Broeke), pp. 267 -173. Dordrecht: Foris.

Goldstein, L. (1989), On the domain of quantal theory, *Journal of Phonetics*, 17: 91-97.

Kaye, J.D. and Lowenstamm, J. (1984), De la syllabicite. In *Forme Sonore du Langage: Structur des Representations en Phonologie* (F. Dell, D. Hirst, and J.R. Vergnaud, editors), pp. 123-160.

Kelso, J.A.S., Vattekiotis-Bateson, E, Saltzman, E. and Kay, B. (1985), A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling, *Journal of the Acoustical Society of America*, 77: 266-280.

Kent, R.D. (1970). *A Cineflourographic-spectrographic Investigation of the component gestures in lingual articulation*, unpublished doctoral dissertation, University of Iowa.

Lindblom, B.E.F., Pauli, S. , and Sundberg, J. (1974), Modeling coarticulation in apical stops. In *Proceedings SCS-74: Speech Communication* (G. Fant, editor). Stockholm: Almqvist and Wiksell.

Lindblom, B.E.F., and Sundberg, J. (1971), Acoustical consequences of lip, tongue, jaw and larynx movement, *Journal of the Acoustical Society of America*, 50: 1166-1179.

Lőfqvist, A. (1980), Interarticulator programming in stop production, *Journal of Phonetics*, 8: 475-490.

Lőfqvist, A., and Yashioka (1980), Laryngeal activity in Swedish obstruent clusters, *Journal of the Acoustical Society of America*, 68: 792-801.

Lőfqvist, A., and Yashioka (1981), Interarticulator programming in obstruent production, *Phonetica*, 38: 21-34.

Munhall, K.G., and Lőfqvist, A. (1992), Gestural aggregation in speech: Laryngeal gestures, *Journal of Phonetics*, 20: 93-110.

Munhall, K.G., Lőfqvist, A., and Kelso J.A.S. (1986), Laryngeal compensation following sudden oral perturbation, *Journal of the Acoustical Society of America*, 80, suppl. 1: s109.

Nadler, R., Abbs, J.H., and Fujimura, O. (1987), Speech movement research using the new X-ray microbeam system, In *Proceedings of the 11th International Congress of Phonetic Sciences, Tallinn, 1-7 Aug., 1987*, vol. 1, pp. 221-224. Tallinn: Academy of Sciences of the Estonian Soviet Socialist Republic.

Nearey, T.M. (1978). *Phonetic Feature Systems for Vowels*, available through the Indiana University Linguistics Club.

Ostry, D.J., Keller, E., and Parush A. (1983), Similarities in the control of the speech articulators and the limbs: Kinematics of tongue dorsum movement in speech, *Journal of Experimental Psychology: Human Perception and Performance*, 9: 622-636.

Ostry, D., and Munhall, K. (1985), control of rate and duration of speech movements, *Journal of the Acoustical Society of America*, 77: 640-648.

Perkell, J.S. (1990), Testing theories of speech production: Implications of some detailed analyses of variable articulatory data In *Speech Production and Speech Modelling, NATO ASI Series D: Behavioural and Social Sciences, Vol. 55*, (W.J. Hardcastle and A. Marchal, editors), pp. 263-288.

Pierrehumbert, J.B.(1980), *The Phonology and Phonetics of English Intonation*, unpublished doctoral dissertation, MIT.

Pierrehumbert, J.B., and Beckman, M.E. (1988), *Japanese Tone Structure, Linguistic Inquiry, Monograph 15*. Cambridge, Mass.: MIT Press.

Saltzman, E.L. (1986), Task dynamic coordination of the speech articulators: a preliminary model, In *Generation and Modulation of Action Patterns* (H. Heuer and C. Fromm, editors), pp. 129-144. New York: Springer.

Saltzman, E.L., Lőfqvist, A., Kinsella-Shaw, J., Rubin, R.E., and Kay, B. (1992), A perturbation study of lip-larynx coordination. In *ICSLP 92 Proceedings: 1992 International Conference on Spoken Language ProcessingI,* Addendum (J.J.

Ohala, T.M. Nearey, B.L. Derwing, M.M. Hodge, and G.E. Wiebe, editors), pp. 19-22. Edmonton, Canada: Personal Publishing.

Saltzman, E.L. and Munhall, K.G. (1990), A dynamical approach to gestural patterning in speech production, *Ecological Psychology*, 1: 333-382.

Stevens, K.N., and House, A.S. (1953), Development of a quantitative description of vowel articulation, *Journal of the Acoustical Society of America*, 27(3): 484-493.

Consonantal evidence against the Quantal Theory*

James B. Long (University of California, Berkeley)
and
Ian Maddieson

Background

To reach an understanding of the basis for the organization of phonological space has been a goal of increasingly sophisticated theories in Linguistics. While organizational notions such as 'vowel space' and 'series' of consonants are taken for granted by most linguists, a few have attempted to provide the scientific underpinnings for language users' exploitation of the phonetic facts. Primarily working with acoustic and articulatory data on the production and perception of vowels, Stevens (1989) attempts such an explanation, but there is evidence that his Quantal theory makes the wrong prediction for Acehnese consonants.

Quantal theory

Stevens (1989) reports that in the production of certain vowels there are ranges of movement of the articulators involved within which there is very little change in the associated acoustic parameter. The relationship between articulatory range and acoustic realization is said to be quantal in that such acoustically stable states or plateaus are separated by regions of rapid transition from one plateau to another as the relevant articulator traverses its normal range of movement. From such evidence, Stevens makes the claim that "certain ranges of acoustic and articulatory parameters are preferred over others in ... language" (1989: 40). In Figure 1 these preferred ranges are shown as State A and State B.

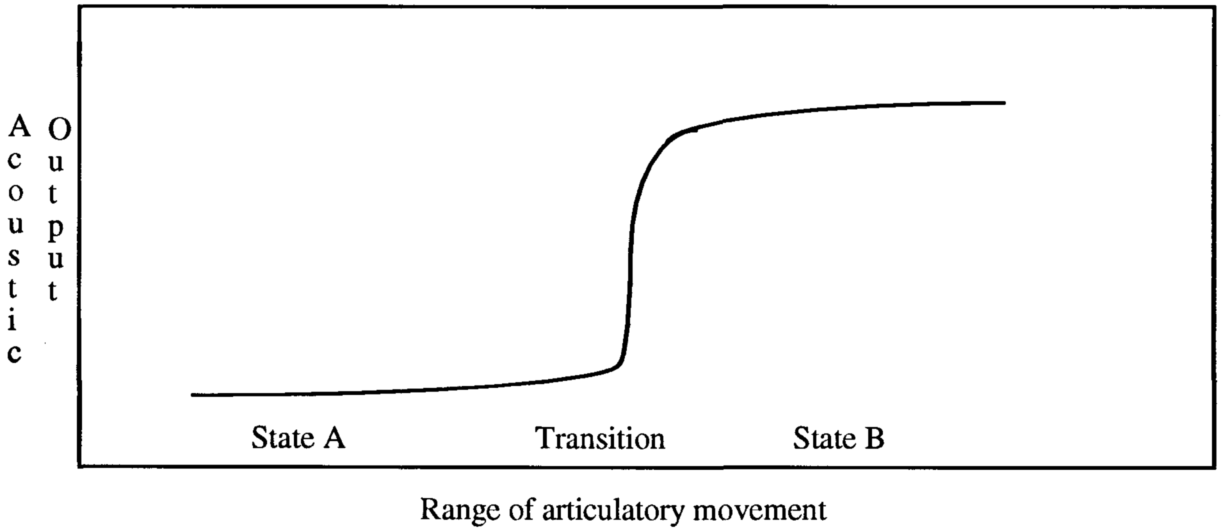


Figure 1. Schematic representation of a non-linear relationship between articulatory movement and acoustic output (after Stevens (1989).

Treatment of consonants

The ideas behind the Quantal Theory were originally developed with the structure of the vowel space in mind (Stevens 1972), and only recently has much thought been given to how they would apply to consonants. In his discussions of several classes of consonants, Stevens draws attention not so much to the stability of the acoustic signal (as he does for vowels) but rather to the abrupt change of state at the transitional edge of a consonant. However, what is common to both classes of sounds is the absence of a requirement for articulatory precision. The stable acoustic regions in the vowel space, and the abrupt changes of state associated with consonants share the property of being producible by a range of articulatory gestures. For example, it does not matter that a stop burst occurs at a precise point in time, only that it occurs at the right point in a sequence. It is enough, then, that the articulators are launched in the right direction toward separation. They do not have to follow a precise course.

'Funny' nasals

In Acehnese, an Austronesian language spoken on the northwest tip of Sumatra, there are series of oral stops, nasal stops, and a third series which we will call 'orally-released nasals'. These are the so-called 'funny nasals' of Durie (1985). We indicate the third series with small superscript stop symbols after a nasal as in Table 1 (cf. Maddieson and Ladefoged 1993). Orally-released nasals are reported in the literature on Acehnese as early as one century ago (Snouck Hurgronje 1892, reported in Durie), and similar phenomena may be present in other Austronesian languages of the area.

Table 1. Four consonant series involving oral closure in Acehnese

Plain stops	b	d	g
Plain nasals	m	n	ŋ
Orally-released nasals	m^b	n^d	ŋ^g
Nasal + stop	mb	nd	ŋg

In addition to these three series, Acehnese permits the concatenation of nasal plus oral stop, such as /mb/ in Table 1.. This series is orthographically identical with the orally-released nasals, but constitutes a fourth contrast, phonemically and phonetically distinct from both the orally-released nasals and the plain nasals. The distinction between orally-released nasal segments and sequences of nasal + stop can be established on the basis of acoustic phenomena, specifically the length of the single segment compared to the length of the sequence, and the absence or presence respectively of a burst of acoustic energy characteristic of a stop. That is, the sequence is longer in duration than the single segment, and the single segment lacks such a burst.

We illustrate these contrasts with spectrograms (from the Signalyze program) of the alveolar place set of phonemes. Figure 2, *bada* 'fried banana', exemplifies the plain stop /d/. The spectrogram shows the attenuation of energy during the stop closure and a clear burst from the stop's release. Figure 3, *pateurana* 'place to put piper beetles', exemplifies a plain nasal: /n/. The spectrogram shows the expected formant-like pattern throughout the segment. We will demonstrate below with a different display that such nasal segments also induce perseverative nasalization on the following vowel. Figure 4, *mandum* 'all', exemplifies a sequence of nasal + stop segments: /nd/. The nasal segment and the closure for the oral stop are apparent, as well as a burst at the release, indicating that the nasal passage has closed at some time beforehand. Figure 5, *banda* 'port', exemplifies an orally-released nasal. It shows nasal characteristics throughout the segment, with some attenuation of higher frequencies prior to the release, but no acoustic burst at the oral release of the segment. This word illustrates the unusual case, the orally-released nasals. Note that the time duration of the nasal + stop sequence in Figure 4 compared with the duration of the single segment in Figure 5 is substantially longer. Several tokens of each of the two words *mandum* and *banda* were recorded and the duration of the sequence of nasal + stop averaged 185 msec; whereas the average duration of the single orally-released segment was 129 msec.

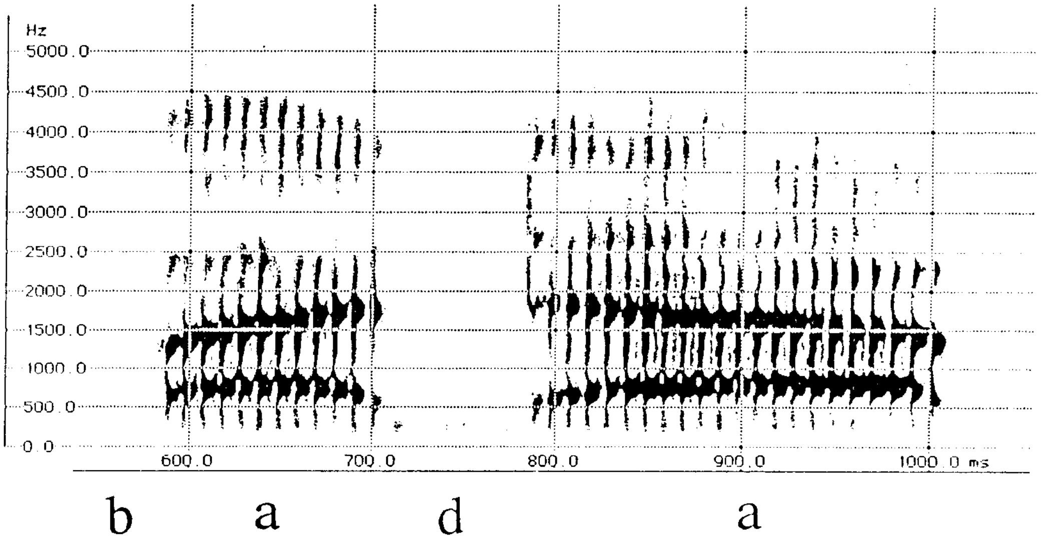


Figure 2. Spectrogram illustrating /d/ in ***bada***

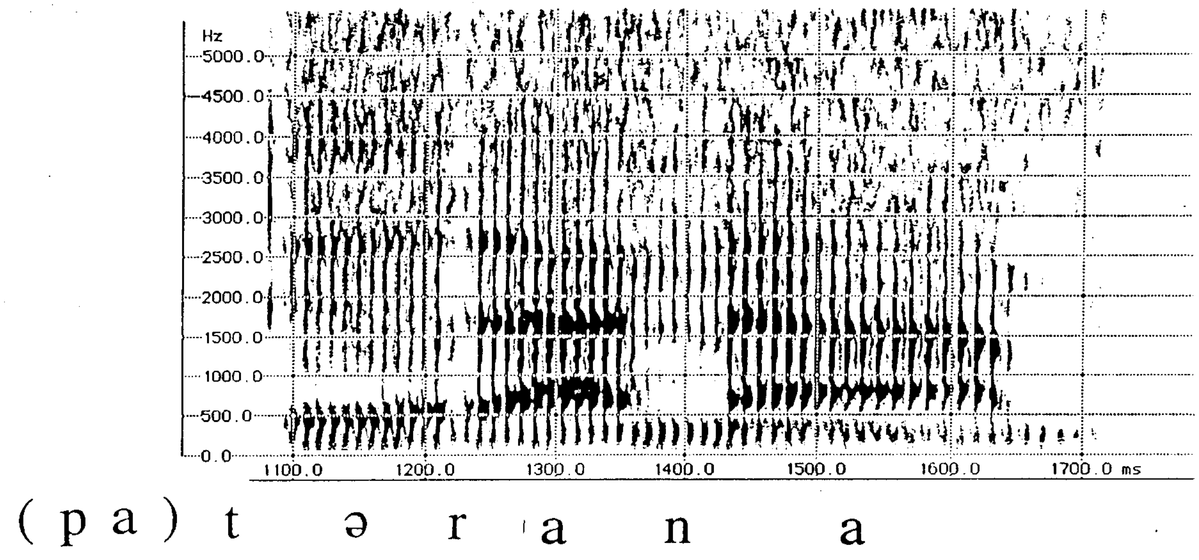


Figure 3. Spectrogram illustrating /n/ in ***(pa)teurana***

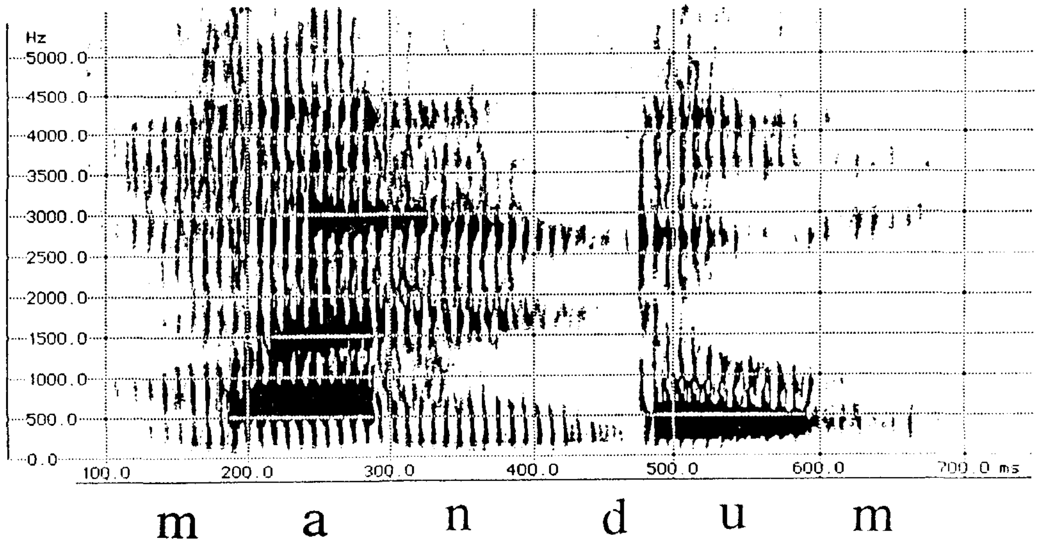


Figure 4. Spectrogram illustrating /nd/ in ***mandum***

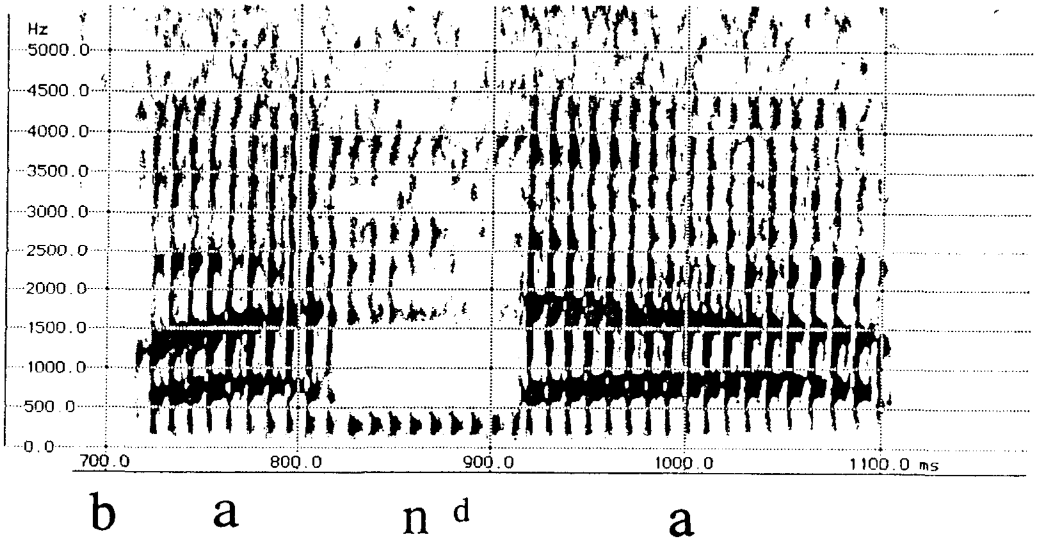


Figure 5. Spectrogram illustrating /n^d/ in ***banda***

A more detailed understanding of how these orally-released nasals are produced can be obtained from examining aerodynamic records, obtained with special portable equipment designed at UCLA. Figure 6, *cama̱* 'sea-mew', shows another example of a plain nasal: /m/. The four traces show: on line 2, oral flow volume, collected via a mask placed over the speakers mouth; on line 3, intraoral pressure, such as that which would build up before a typical oral stop release, sensed via a tube in the speaker's mouth, which is connected to a pressure transducer; on line 4, airflow through one nostril, collected via a 'nasal olive' inserted in the nostril and connected by a tube to a transducer (the other nostril is pinched shut during the recording); and, on line 1, an audio recording made from outside the mask, hence its degraded quality. It does, however, serve as a place finding device.

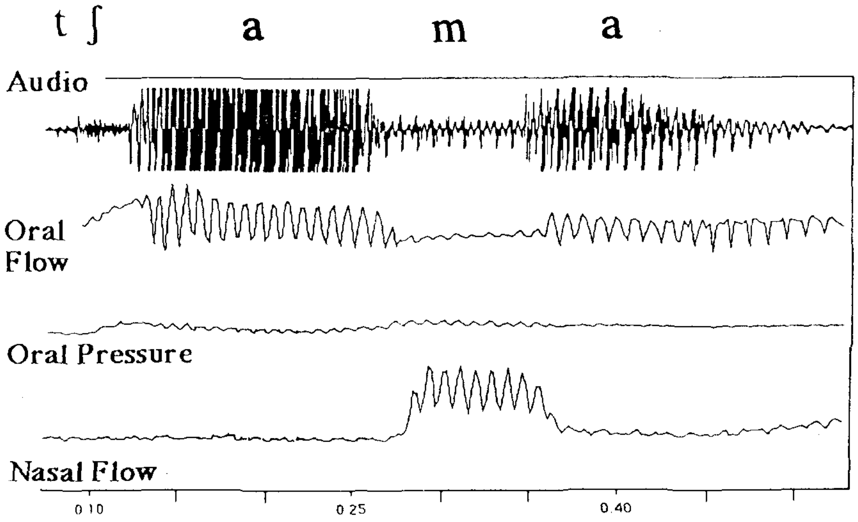


Figure 6. Aerodynamic records illustrating /m/ in ***cama*** . Perseveration of nasal flow into following vowel is seen as continued elevation of the nasal flow trace above the baseline level after the onset of the second vowel.

During a VNV sequence containing an ordinary nasal, such as in *cama*, the flow of air is transferred from the mouth to the nose and back again, but note that the nasal flow continues after *the oral flow resumes. In other words, no particular care is taken to ensure any strict coherence in* time between the moment of separation of the lips and the velum raising gesture. This is in line with Quantal theory, since the labial release itself creates an abrupt change of state, sufficient to mark the difference between the nasal consonant and the following vowel.

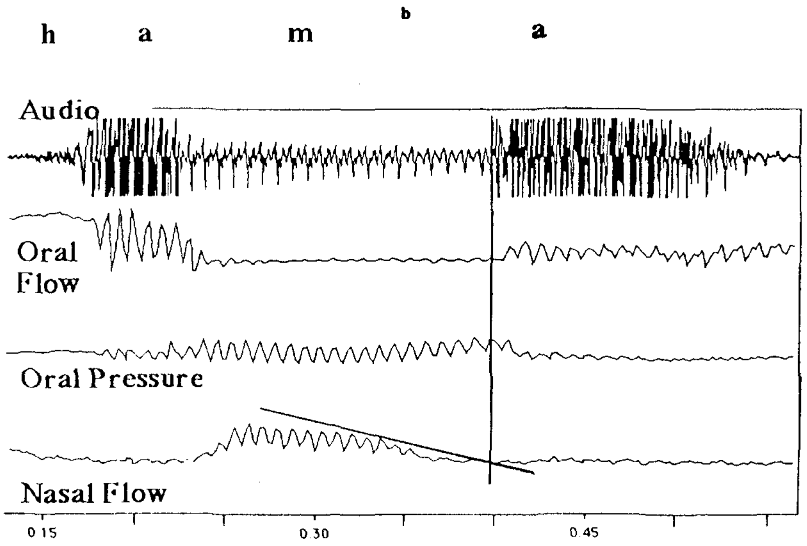


Figure 7. Aerodynamic records illustrating /m^b/ in ***hamba.*** The construction of the lines drawn on the record is described in the text.

Figure 7 shows similar records of the word *hamba*, 'slave', containing an orally-released nasal. In contrast with plain nasals, during orally-released nasals the velum raising shows every sign of being managed so that the velic closure and oral release are quite precisely aligned in time. The slow decline of the nasal flow trace shown here indicates such "management". We indicate the precision achieved in the timing of the two movements involved by constructing a line to mark the oral release and fitting a second line between a peak in the nasal airflow and the point at which the nasal flow is judged to have reached baseline level (this can be considered a way of indicating the declination of the nasal flow during the consonant). These lines intersect at the baseline level of the nasal flow, indicating that cessation of nasal flow and oral release for the vowel onset coincide in time. Note that, unlike in Figure 6, there is no nasal air flow at the beginning of the vowel following the orally-released nasal segment; and that also there is also no sharp rise in intraoral pressure which would have been expected if the velum had been fully closed at any time prior to the release. Similar patterns were seen in the majority of aerodynamic records of other tokens.

We deduce that the timing of the nasal and oral articulations of these orally-released nasals is such that the achievement of velic closure is simultaneous with the oral release. Thus there is no perseverative nasalization, nor any acoustic burst, since the velo-pharyngeal port remains at least partially open at the critical time during which pressure would otherwise build up in the oral cavity. It can be readily seen that great precision of timing is necessary to execute these segments accurately. If the timing were such that the oral release preceded the velum raising, a plain nasal would be the acoustic result, with perseverative nasalization on the following vowel. If the oral release followed the velic closure, a sequence of nasal plus stop would result. As shown earlier, segments of both these other types are contrastive in Acehnese, and free variation of this kind would lead to loss of lexical distinctiveness.

Conclusion

Our conclusion is that the degree of articulatory precision required to correctly produce these Acehnese phonemes, the 'orally-released nasals', is exactly the wrong characteristic of a consonant according to Stevens' Quantal Theory. The theory indicates that such segments should be disfavored. That these segments have been maintained in Acehnese for at least 100 years constitutes direct evidence against the extension of Quantal Theory—as currently understood—to consonants.

Having said that, we must report that there is sufficient variation in repeated utterances of orally-released nasals by our Acehnese consultant for incipient acoustic bursts to be detected in some tokens of words containing orally-released nasals, due to early raising of the velum. One such token is shown in Figure 8. Our consultant does not show any tendency to delay raising the velum, which would produce the plain nasal. Thus the contrast between plain nasal and orally-released nasal is maintained. Nevertheless, it may be that, despite the fact that Acehnese speakers identify the ability to correctly produce orally-released nasals as an important indicator of native speaker fluency, they do find the degree of articulatory precision necessary for the production of these segments difficult to maintain. These segments may be in the process of changing and it may turn out that Quantal Theory has, in the long run, made the correct prediction after all.