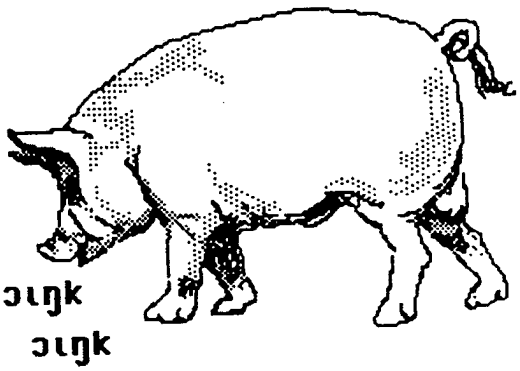
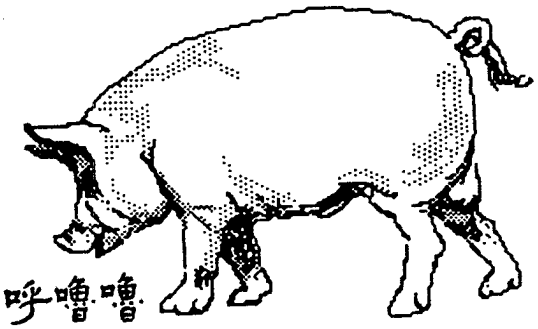
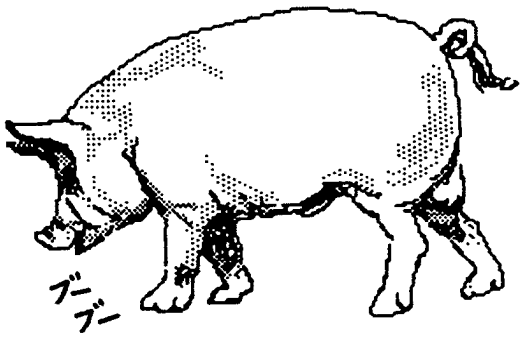
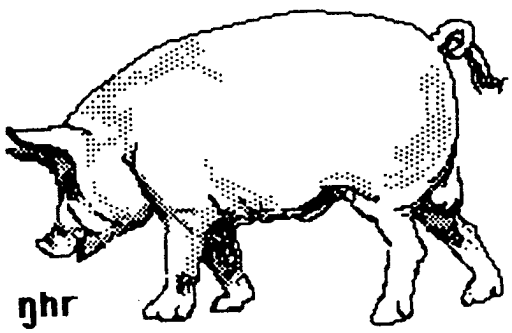
UCLA Working Papers in Phonetics

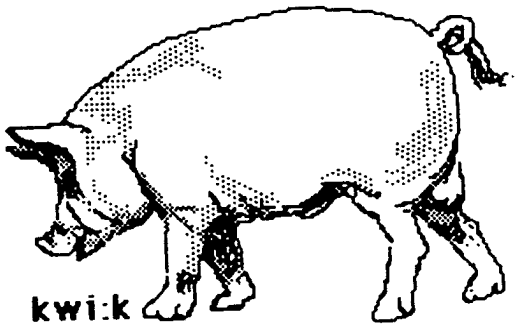


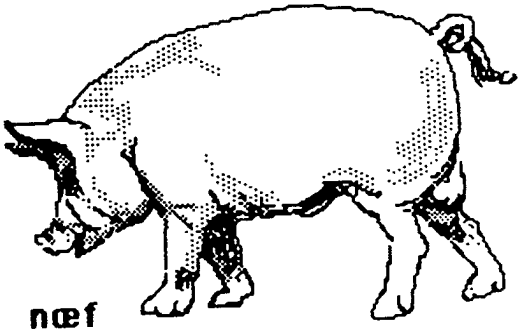




Number 61, July 1985







The UCLA Phonetics Laboratory Group.

Stephen R. Anderson
Alice Anderton
Norma Antoñanzas-Barroso
Abby Cohn
Sarah Dart
Bill Dolan
Karen Emmorey
Vicki Fromkin
Bruce Hayes
Marie Huffman
Michel Jackson
Pat Keating
Paul Kirk
Jody Kreiman
Peter Ladefoged
Jenny Ladefoged
Karen Lau
Mona Lindau
Ian Maddieson
Yoko Mimori
Kristin Precoda
Ren Hong-Mo
Lloyd Rice
Mika Spencer
Henry Teheranizadeh
Diana Van Lancker
Dawnelle White
Wanda White

As on previous occasions, the material which is presented in this volume is simply a record for our own use, a report as required by the funding agencies which support the Phonetics Laboratory, and a preliminary account of research in progress for our colleagues in the field.

Funds for the UCLA Phonetics Laboratory are provided through:

NSF grant BNS-23110
USPHS grant 1 R01 NS18163-03
USPHS grant 1 R01 NS22726-01
and the UCLA Department of Linguistics.

Correspondence concerning UCLA Working Papers in Phonetics should be addressed to:

Phonetics Laboratory
Department of Linguistics
UCLA
Los Angeles CA 90024
(U.S.A.)

UCLA Working Papers in Phonetics is edited by Ian Maddieson.

UCLA Working Papers in Phonetics 61

July 1985

Measures of Phonation Type in Hmong

Marie K. Huffman

1. Introduction

Variation in voice quality is one means by which languages indicate phonological contrasts (Ladefoged 1983). In addition to "normal" phonation, speakers use breathy and creaky phonation types. Languages employ these phonation types distinctively, though to date no language is known to use more than three phonation types in contrast. In recent years phoneticians and speech scientists have been very interested in determining the acoustic and physiological properties crucial to distinctions in phonation type. Two interrelated lines of research are being pursued. One is the identification of distinctive characteristics of phonation types which might be compared across speakers and languages (Fischer-Jørgensen 1967, Laver 1980, Bickley 1982, Ladefoged 1983). The other is identification of the best means of quantifying different aspects of phonation, from the acoustic signal to vocal fold vibration (Fourcin 1974, 1981; Rothenberg 1981; Javkin and Maddieson 1983; Kirk, Ladefoged, and Ladefoged 1984.)

As linguists we want to ascertain whether there are restrictions on what may be used as contrastive phonation types in language. That is, we are interested in knowing if there is a limited number of phonetic categories available for use in phonological phonation contrasts, as, for example, Keating (1984) asserts is the case for the phonetic representation of stop consonant voicing. To investigate whether phonation types show restrictions on their phonetic realization, we need to know the phonetic dimensions involved in phonation contrasts.

With respect to acoustic dimensions, linguistic concerns lead us to expect a limited number of acoustic cues to phonation distinctions. Since a speaker of a language with contrastive phonation types has to be understood by other speakers of that language, he must signal voice quality distinctions so that there is a consistent difference between contrasting items. Furthermore, as Kirk, Ladefoged and Ladefoged (1984) propose, we would expect that all speakers of such a language would signal changes in phonation type "in much the same way" (p. 103).

As for what the acoustic dimension(s) might be, recent studies using spectral analysis of normal audio recordings (Fischer-Jørgensen 1967; Bickley 1982; Ladefoged 1983; Kirk, Ladefoged, and Ladefoged 1984) have found that measures of the amplitude of some higher spectral component in relation to the amplitude of the fundamental was a distinctive acoustic difference between contrastive phonation types. For example, Fischer-Jørgensen (for Gujarati vowels) and Ladefoged (for !Xóõ vowels) found that the difference in intensity of the fundamental and the first formant or lower harmonics is a fairly reliable correlate of the breathy/normal phonation contrast. Bickley (1982) confirmed these findings and presented results of a perception test which support the notion that listeners are sensitive to this sort of spectral balance when they judge degrees of breathiness. Finally, Kirk et al's (1984) study of the three-way phonation type distinction in Jalapa Mazatec found that the relative difference between the intensity of the fundamental and the first formant was successful in distinguishing creaky, modal (normal) and breathy vowels. Thus,

phonation types might be characterized by values along a continuum of spectral balance.

As for the physiological domain, there have been several proposals in the literature that phonation types can be characterized in terms of glottal stricture. Catford (1977) proposes a detailed schema for categorizing possible phonation types. He uses two parameters: stricture type and stricture location. Stricture types have characteristic degrees of aperture. In order, from most open to most closed, his basic stricture types are voiceless, whisper, voice, and creak. The additional parameter of stricture location, anterior versus posterior, allows for combinations of basic stricture types. So for example, creak produced in the anterior part of the vocal folds, with whisper in the posterior region of the folds, together form whispery creak. The net yield of such combinations is a large inventory of phonatory possibilities, for only some of which Catford mentions linguistic uses. Laver (1980)'s system for characterizing phonation type production has a slightly different focus. He describes the types of muscle tension and compression that are possible in different regions of the vocal folds given the musculature of the larynx. The resultant set of phonatory possibilities is of a size comparable to Catford's.

Ladefoged (1973) proposes a continuum of linguistic phonation types based on degrees of glottal stricture alone, with values ranging from closed glottis (glottal stop) to spread (strong aspiration). Five of these values denote qualities we would treat as different modes of vocal fold vibration; namely breathy, lax, normal, tense and creaky voice. However, as Ladefoged notes, the number of values is arbitrarily chosen; there might be more, or less.

To test the notion of restricted dimensions underlying phonation contrasts, we need quantitative characterizations of phonation types in individual languages, so that we can make meaningful cross-language comparisons. Towards this goal, we focus here on the phonetic characterization of phonation types in Hmong, a Southeast Asian language which uses breathy and normal phonation contrastively. Trying to quantify acoustic differences between phonation types is a more straightforward task than quantifying laryngeal gestures producing these differences. As was mentioned earlier, previous studies suggest that measures of prominence of the fundamental relative to the first formant or lower harmonics are fairly reliable correlates of phonation contrasts. These spectral properties can be measured from normal audio recordings.

On the other hand, data on laryngeal gestures are not easily obtained because the larynx is delicate and somewhat inaccessible. However, there are indirect indicators of laryngeal behavior, e.g. the glottal air flow. Some investigators (e.g. Bickley 1982, Javkin and Maddieson 1983) have used the technique of inverse filtering to study flow characteristics of linguistically contrastive phonation types. The procedure allows the recovery of glottal flow from oral airflow.

Vocal fold closing speed and closure duration are reflected in properties of the glottal flow waveform (Rothenberg 1981). Thus variations in these physiological properties can be inferred from measures of these properties. Below we will outline the physiological and aerodynamic requirements for voicing, then we will consider the effect which glottal stricture could have on vocal fold closing speed and closure duration. The kinds of flow waveform measures likely to reflect these physiological properties will also be outlined.

The general physiological and aerodynamic conditions necessary for normal voicing are described by van den Berg (1958). The vocal folds themselves must be somewhat tensed and partially adducted. The elastic properties of the vocal fold tissues help restore the folds to this starting position. The essential aerodynamic requirement for voicing is that there be a sizable pressure drop across the glottis. This pressure drop must be large enough that air will pass through the glottis at a rate sufficient for the Bernoulli effect to suck the vocal folds together. Once the folds are together, air pressure below the folds builds up until it pops them open, at which point the cycle begins again.

Given these general specifications for voicing, what is changed when a particular vibration type is the goal? Control of rate of vibration is based primarily on the mass and stiffness of the vocal folds, and subglottal pressure (Hollien 1984). In contrast, as was noted above, descriptions of phonation type make crucial reference to changes in glottal stricture (Ladefoged 1973, Catford 1977). Speaking very generally, breathy voice is considered to be generated by laxer than normal vocal folds which are abducted more than for normal voice. The folds may never completely close, or they may vibrate fairly normally in the anterior portion but always remain open to a degree between the arytenoid cartilages. Creaky voice is produced with the glottis more tightly closed than normal, with air escaping in (sometimes irregular) bursts at the forward end of the vocal folds.

How will the glottal stricture differences described for breathy and creaky voice affect vocal fold closing speed and closure duration? Let us consider first what might happen if the vocal folds were less closely approximated than in normal voice, as is reportedly the case in breathy voice. Given a comparable trans-glottal pressure difference that is fairly constant, when the airflow meets the less severe constriction at the vocal folds, velocity of air as it passes through the glottis will increase less. The result is a weaker Bernoulli effect and thus a less sharp closure of the vocal folds. Furthermore, if the folds are far enough apart, they might never close completely, though a weak Bernoulli effect could still pull them towards each other.

When the vocal folds are more closely approximated than in normal voice, as may be true in creaky voice, we predict the opposite results. In this case, if there is an initial glottal opening, it will be relatively small. Therefore we would expect a particularly strong Bernoulli effect, with a correspondingly sharper than normal vocal fold closure. In addition, we would predict that when the vocal folds are held together, there will be strong resistance to their separation by subglottal pressure and strong tissue recoil force pulling them together after they do separate. Thus the closing of the glottis should be fast, and closure duration as a proportion of the total pitch period would probably be longer than normal.

Observations of glottal waveforms given by Bickley (for Gujarati and !Xóõ) and Javkin and Maddieson (for Burmese) suggest that properties of the glottal flow waveform are distinctive for different phonation types in ways we might expect given the preceding discussion. Bickley notes, for example, that in Gujarati, glottal flow pulses of breathy vowels are more symmetrical during the open phase, and have shorter closed phases, than normally voiced vowels. These observations are supported by evidence from laryngographic studies of breathy and normal phonation. For example, Fourcin (1974) reports that in breathy voice, though vocal fold closure is more rapid than release, there is less difference between the two in breathy than in normal voice. Javkin and Maddieson

investigated a number of measures of flow pulse shape and duration differences in Burmese. They report that phonation types in Burmese were distinguished by the slope of the line between what were judged as endpoints of the rising branch of the glottal flow pulse. Sometimes the comparable measure of the falling branch of the pulse was also successful. However, they note that amplitude variations tended to interfere with the discriminatory power of their slope measures.

What might measures of glottal flow properties tell us about the acoustic cues to phonation contrasts? The studies of spectral differences between phonation types mentioned earlier make reference to variations in spectral balance in terms of relative strength of the fundamental frequency. Since changes in voice quality are attributed to the glottal source, a logical place to look for the basis of these spectral differences across phonation types is in frequency characteristics of the glottal flow waveform.

Van den Berg (1958) discusses the properties of the glottal spectrum with reference to glottal pulse shape. Although he is talking about differences in mode of singing voice or changes in intensity of normal voice, rather than addressing phonation types, his comments are still helpful in predicting what spectral differences to expect for different voice qualities. Van den Berg says that a "weak and relatively broad puff" (open phase) of the glottis, characteristic of falsetto and low intensity normal voice, will produce few higher harmonics of the fundamental. He contrasts this with a "strong and relatively short puff," characteristic of high intensity normal voice, which produces many high harmonics.

Fant (1980) makes similar claims about the relation of spectral properties to glottal pulse shape. The voice source model described there was developed to account for relationships between glottal spectra and glottal pulse shapes observed for normal phonation produced at different intensities. He notes that pulses that have a less sharp closing, akin to van den Berg's "weak puff," will have spectra dominated by the fundamental, with weak higher harmonics. In contrast, pulses with proportionately sharper closing phases will have weaker low harmonics and stronger higher frequency harmonics.

To the extent that the vocal folds are sucked together weakly in breathy voice, we predict that the glottal spectrum will be comparable to that of van den Berg's "broad puffs," with a prominent fundamental and few strong higher harmonics compared to normal voice. Similarly, creaky voice, which we predicted would have a sharp closing gesture, should produce many high harmonics, even more than van den Berg's strong, short "puff" which corresponds to a kind of normal voice. Thus our investigation of acoustic cues to the breathy/normal phonation contrast in Hmong will be addressed to possible differences in glottal spectra as well as waveshape properties.

We have identified several properties of glottal flow waveforms which might be distinctive for phonation types. In the following, we will discuss how well these measures characterize phonation contrasts in Hmong. While doing so, it is important to bear in mind that phonation type differences used contrastively in natural languages are properties of segments that have many other attributes, including duration and pitch. If our measures of glottal waveforms capture only phonation contrasts despite the possible influence of these other properties, then they can be considered that much more effective indicators of significant physiological and acoustic differences. Knowing which measures consistently distinguish phonation differences in Hmong will help us isolate phonetic

characteristics of phonation contrasts which might be studied cross-linguistically.

2. Materials

Hmong is a Sino-Tibetan language which has been described as having seven tones. Tones in Hmong have both a particular pitch and a characteristic voice quality. Five of the seven tones--high, rising, mid, low, falling--are reportedly produced with normal voice. Of the remaining two tones, one is described as low, produced with what Lyman (1974) calls "voiced aspiration" or "breathiness." Smalley (1976), similarly, characterizes this tone as breathy, and speculates that this quality is "possibly caused by an enlarged laryngeal cavity and doubtless by a special configuration of the vocal cords" (p.100). The other tone is said to be slightly falling, with a reported rising variant which, Lyman notes, occurs in utterance final position and very careful speech. Only falling tokens of this tone were observed in the present study. This seventh tone is characterized by Lyman as having a quality of "glottalization" or "creaky" voice (p.38). Smalley, on the other hand, attributes no special voice quality to this tone; rather, he describes it as shortened and "terminated by a glottal stop" (1976:100). As will be discussed below, Hmong data collected for the present study support Smalley's description; i.e. this seventh tone does not seem to involve a phonation contrast. We will refer to this tone as the checked tone.

Table 1 illustrates this seven-way contrast. Consistent with Smalley's description, we transcribe checked tone syllables with a final glottal stop.

Table 1

Tone (quality)	Word and Gloss	
high (normal)	tau^{55}	"pumpkin"
rising (normal)	tau^{35}	"to dam up (water)"
mid (normal)	tau^{33}	"to be able"
low (normal)	tau^{22}	"axe"
checked (normal)	$tau?^{31}$	"bean"
falling (normal)	tau^{42}	"sp. of grass"
low (breathy)	tau^{32}	"to follow"

Since our measures of glottal flow waveforms are most meaningful when they reflect only phonation type differences, we want to compare the "breathy" and checked tones with normal voice tones of similar pitch range and contour. Consultation with a native speaker revealed that the falling and low normal voice tones would be appropriate for this purpose. The low tone is in fact slightly falling. Paradigms were developed with these tones on several different syllables. The main word list used in this study appears in Table 2. A tentative segmental transcription is included for each syllable.

Table 2. List of Hmong words.
Dashes mark nonsense words added to complete paradigms.

Low	Checked	Falling	Breathy
pa^{22}	pa?^{31}	pa^{42}	pa^{32}
stick	blanket	flower	-
tau^{22}	tau?^{31}	tau^{42}	tau^{32}
axe	bean	sp. of grass	to follow
taɯ^{22}	taɯ?^{31}	taɯ^{42}	taɯ^{32}
to emit light	to go out	-	firewood

Three male speakers of Hmong between the ages of sixteen and eighteen were recorded reading each of the words in Table 2 twice. All from Laos, two of the speakers were Hmong Daw, that is, belonging to the White Hmong tribe; one was Hmong Njua, or Blue Hmong. Dialectal differences between these two groups, however, are small and generally do not involve the tones of interest to this study.

Combined oral and nasal airflow was recorded for each speaker using a special mask of the type described by Rothenberg (1973), in conjunction with the UCLA Portable Instrumental Phonetics Station the relevant components of which are illustrated in Figure 1. The flow recording then served as the input to inverse filtering, described in Section 3. The FM recording system had a frequency response flat (± 3 dB) from DC to 2000 Hz. The frequency response of the flow recording system was not calibrated, but, as the transducer itself has a flat frequency response up to at least 2000 Hz, the overall response is presumably limited by the characteristics of the mask. Rothenberg reports that the frequency response of the mask is flat to about 1000 Hz (1973:1637).

During recording, the signal from the mask was sent through a preamplifier, then through an FM system, and on to a Nagra IV-L tape recorder. For each speaker, a test sample was made during which the FM signal was recorded on the Nagra, and was then reproduced, and displayed on the chart recorder for comparison with the unmodulated signal from the preamplifier. Thus the recording and play back system was checked for each speaker, to ensure a minimum of distortion by the FM system. A cassette tape recording was also made of most of the recording session, to facilitate identification of tokens during analysis. A final audio recording was made without the interference of the mask. At this time the words in Table 2 were said once by each of the speakers, recorded onto the Nagra using a Sennheiser condenser microphone.

3. Analysis Procedure

The first step in analysis was to digitize the recorded flow data by sampling with an LSI 11/23 computer. After demodulation, the recorded signal was passed through a low pass Bessel filter with attenuation of -25 dB at 9,000 Hz., and digitally sampled into the computer at a rate of 18,000 samples per second. Individual words were then identified and stored separately for further

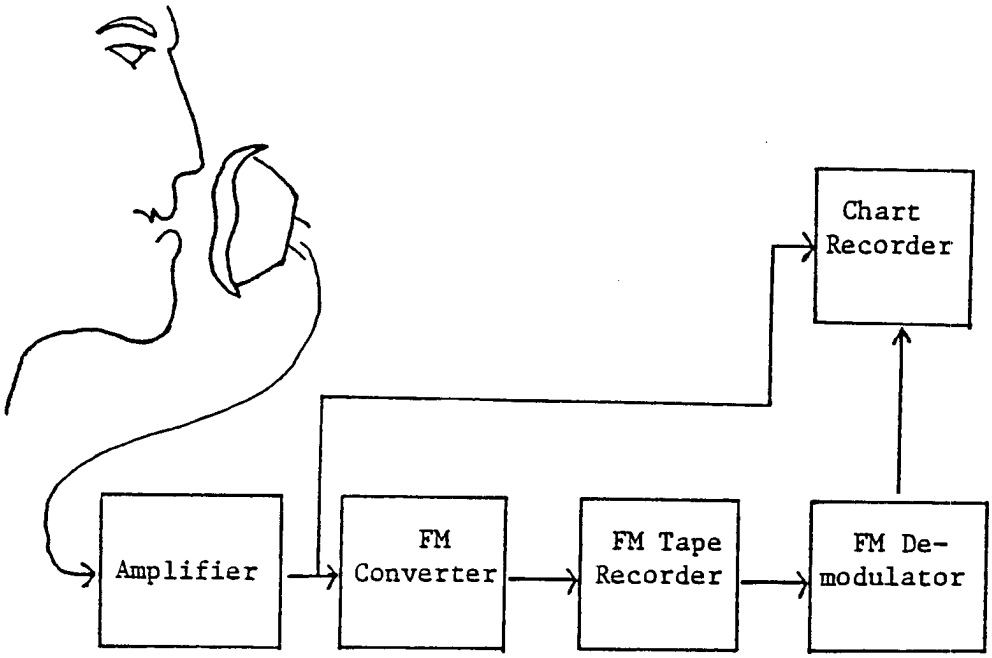


Figure 1. Airflow recording system

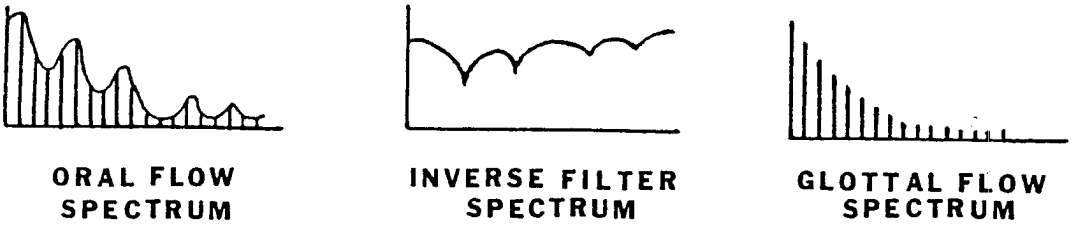


Figure 2. Spectral effects of inverse filtering

analysis. Another preliminary step was production of wide and narrowband spectrograms, and power spectra, from the air flow recording of each vowel token, using a Kay Digital Sonagraph. Wideband spectrograms were used to determine vowel duration and to observe the general formant properties of the vowel tokens. Power spectra, taken during steady-state portions identified on the wideband spectrograms, also offered insight into formant structure. Narrowband spectrograms were used for pitch measurements.

In measuring duration from the wideband spectrograms, the beginning of the vowel was taken to be the moment of voice onset after release of the (unaspirated) stops. The end of the vowel was judged as the last point at which there was energy in two of the first three formants. To ascertain pitch from the narrowband spectrograms, measurements were made at onset, midpoint, and offset of the highest well-defined harmonic (usually the fourth or the sixth) for each vowel token. Fundamental frequency was then calculated from these values.

As mentioned earlier, a major goal of the analysis was recovery of the glottal flow waveform from the airflow recording. The technique of inverse filtering was used, employing a computer program developed at UCLA; for a more detailed description of the procedure see Javkin, Antoñanzas-Barroso and Maddieson (1985). The general method of inverse filtering aims to filter out the resonance effects of the vocal tract and radiation at the lips in order to determine characteristics of the acoustic signal that can be attributed to the glottal source. Lip radiation effects present in oral pressure waveforms are not a factor in volume velocity measured with the Rothenberg mask. Thus, in the present study, only the resonance effects of the vocal tract needed to be filtered out.

The inverse filtering process can be viewed most easily in terms of changes in spectra. As illustrated schematically in Figure 2, the procedure involves filtering the original sound (as represented by flow recordings) using a filter with zeroes at points in the spectrum corresponding to poles of the vocal tract; that is, at spectral peaks induced by supraglottal resonances. The result is a spectrum of the glottal source. In the time domain, the result is a fairly smooth waveform representing flow at the glottis. Of the nine zeroes applied by the inverse filtering program, the first four were specified by the experimenter; the other five had fixed values set at intervals of 1000 Hz beginning with 4500 Hz.

LPC formant analysis was used primarily to identify the frequencies of the poles in the spectrum where the first four zeroes should be applied. To facilitate determination of formants centered at higher frequencies, for the purposes of LPC formant analysis the flow signal was differenced twice. This resulted in a preemphasis of 12 dB per octave, the roll-off characteristic commonly reported for the glottal source. This procedure has very little effect on the determination of the frequencies of the poles, although it will have a serious effect on the determination of the bandwidths.

Next a Hamming window was stepped through each token and linear prediction coefficients were computed, at 10 msec intervals. Trial analyses calculating formants via 14 coefficients and 16 coefficients proved unable to consistently separate the first four formants, so for this data 18 coefficients were calculated. These were then converted to formant frequencies and bandwidths, and written to files for editing. Editing involved verifying frequencies and bandwidths appropriate for the first four formants of the vowels [a], [au], and

[aɯ] with reference to the wideband spectrograms and power spectra mentioned earlier. LPC formant analysis occasionally generated obviously spurious values. These were corrected, and values for formants higher than the fourth were deleted so that the files could serve as input to the inverse filtering program.

A first run of the inverse filtering program was then performed. There was considerable variation in the smoothness of waveforms produced by this first pass of inverse filtering. The multiplicity of factors affecting glottal flow -- oral and subglottal pressure, vocal fold mass, tension and symmetry, to name a few -- lead us to expect gross variations in flow waveshape across speakers and phonation types as well as flow oscillations during the course of any single glottal pulse. However, inverse filtering should remove fine flow oscillations attributable to resonances of the vocal tract. Before it was possible to determine the significance of waveshape variations, it was necessary to determine whether incomplete filtering might be occurring.

A portion of each vowel token was selected for this and further analyses. The location within the vowels of the chosen portion was motivated by properties of the inverse filter. The inverse filtering procedure assumes that the vocal tract has relatively little effect on the glottal flow, i.e. that there is not coupling between vocal tract resonances and the fundamental. This means that when the first formant is low and thus close to the fundamental frequency, the inverse filter will be using an inaccurate estimation of vocal tract resonances. Therefore a high first formant is needed.

For the Hmong vowels, this requirement was met by using the non-front low monophthong [a] and the non-front low onset of the diphthongs [au] and [aɯ]. A steady-state portion of six glottal pulses was selected in the middle of each vowel token by consultation with LPC outputs, and in some cases, with spectrograms as well. In the case of the monophthongs, this portion was near the middle of the vowel; for the diphthongs the steady portion was at about one-third of the duration of the vowel.

Figure 3 shows a sample of a glottal flow waveform, with arrows marking oscillations possibly attributable to incomplete filtering. For this and other tokens, the frequency of such oscillations was measured from the computer display. If they corresponded to formant frequency values for the token, then it was assumed that the formant to be cancelled out by inverse filtering was of greater amplitude than originally allowed for by the bandwidths supplied by LPC formant analysis.

Original bandwidths from LPC analysis of the airflow data were on the order of 120-200 Hz for F1, 130-300 Hz for F2, 300-580 Hz for F3, and 300-640 Hz for F4. The poor frequency response of the mask is surely responsible for the attenuation and therefore greater bandwidths in the higher frequency ranges. To attempt to provide for improved filtering of formants, the bandwidths in the LPC output file were decreased, and the sample inverse-filtered again to see if more complete filtering resulted. This cycle of varying bandwidths and re-inverse filtering was repeated with different amounts of bandwidth reduction, until the best filtering with the least possible bandwidth reduction was achieved. In general, for formant one or formant two, effective bandwidth reduction was on the order of 10-50 Hz. For formants three and four, reduction was on the order of 100-200 Hz. On a few occasions examination of an incompletely filtered signal suggested that the frequency values calculated via LPC analysis for F1 or F3 might be centered slightly high. In these cases lowering the first or second

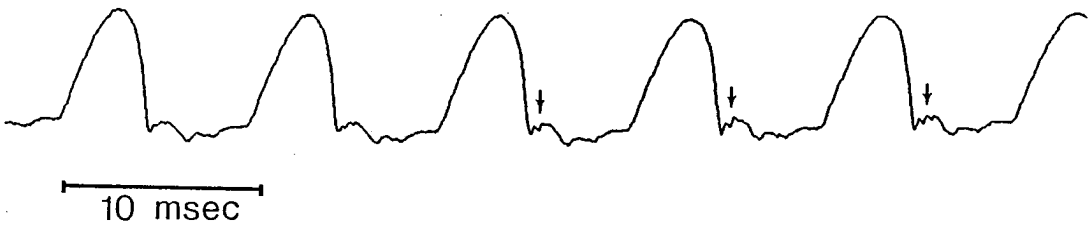


Figure 3. Incompletely filtered waveform

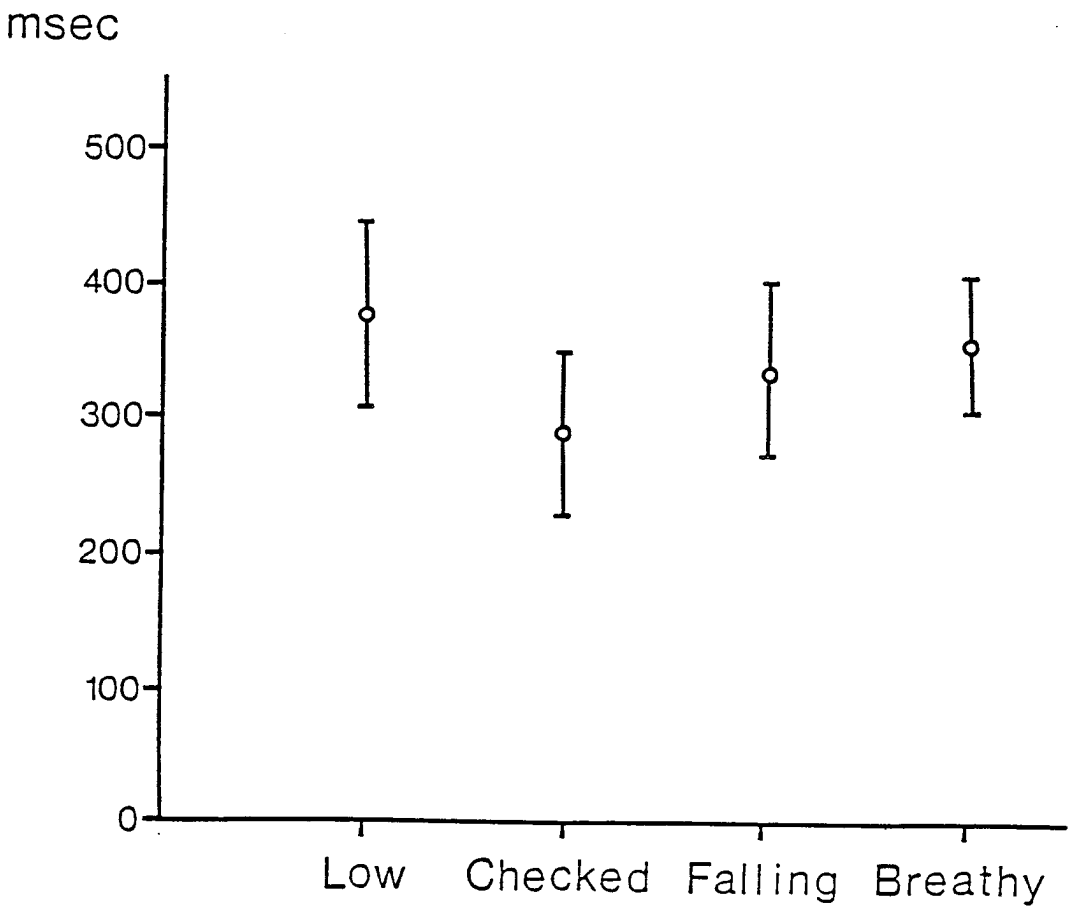


Figure 4. Duration of vowels

formant by 20-40 Hz improved the output of inverse filtering. In these cases lowering the first or second formant by 20-40 Hz improved the output of inverse filtering.

4. Results

4.1 Duration and Pitch

Before examining the outputs of inverse filtering, we need to consider the duration and pitch of the vowel tokens. Duration is of interest because it might be playing a role in signalling the phonemic contrasts in question. The same concern applies to pitch. Another reason to examine pitch is that changes in vocal fold tension and subglottal pressure that contribute to pitch changes can affect the glottal flow and hence also influence our measures of these properties of the glottal waveform which could distinguish phonation types.

Figure 4 shows average duration of vowels spoken with the four Hmong tones of interest to this study. Data are pooled for the 3 speakers. As was noted by Lyman (1974), the checked vowels are shorter than the normal (low, falling) and breathy ones. Notice also that among these latter three tones, the low tone is on average longer than both the breathy and falling tones. An analysis of variance on the pooled data found that of the four tones studied here, the checked tone had a significantly different duration from the low and breathy tones; there was no significant difference in duration between checked and falling tones or between low, falling and breathy tones. (Unless stated otherwise, the significance level is at least .01, determined using Tukey's studentized range test, which controls for the experimentwise error rate for multiple comparisons). So checked tones are usually shorter than the other tones, but duration is not a completely consistent cue to the tone differences.

Turning now to pitch, consider Figure 5, which shows average onset, midpoint and offset values of F0 for each tone type, for each of the three speakers. For all three speakers, the low and breathy tones are in fact falling. The low tone falls only slightly. Its onset is below those of the other three tones, but in falling so little, it ends with a final value higher than the others, except for speaker 2, whose low and checked tones have very close offset values. For all speakers, the falling, breathy and checked tones often share onset or offset pitches, but they are always distinguished at the second endpoint, and often, at the midpoint as well. So for example, the falling tone starts higher than the other tones but drops to about the same final pitch as the breathy tone. Thus the falling tone can be assigned a numerical tonal contour of 42, while the breathy tone is designated 32. The checked tone has an onset value very close to that of the falling tone, but its final value is somewhat lower, thus it is specified as 31. The low tone is specified as 22. It should be noted that these descriptions and tonal contour specifications are only very general characterizations of the facts in Figure 5. There is a considerable amount of speaker variation, both in frequency range employed, and in the amount of difference between tonal contours. For example, the breathy and falling tones of speaker 2 are very similar in contour and frequency range; the difference in onset values of these two tones is less than for the other speakers, even taking into consideration the smaller absolute range of frequencies being used by this speaker.

To summarize this section, a number of important points should be noted. First, the checked tone is often significantly shorter than other tones, but its

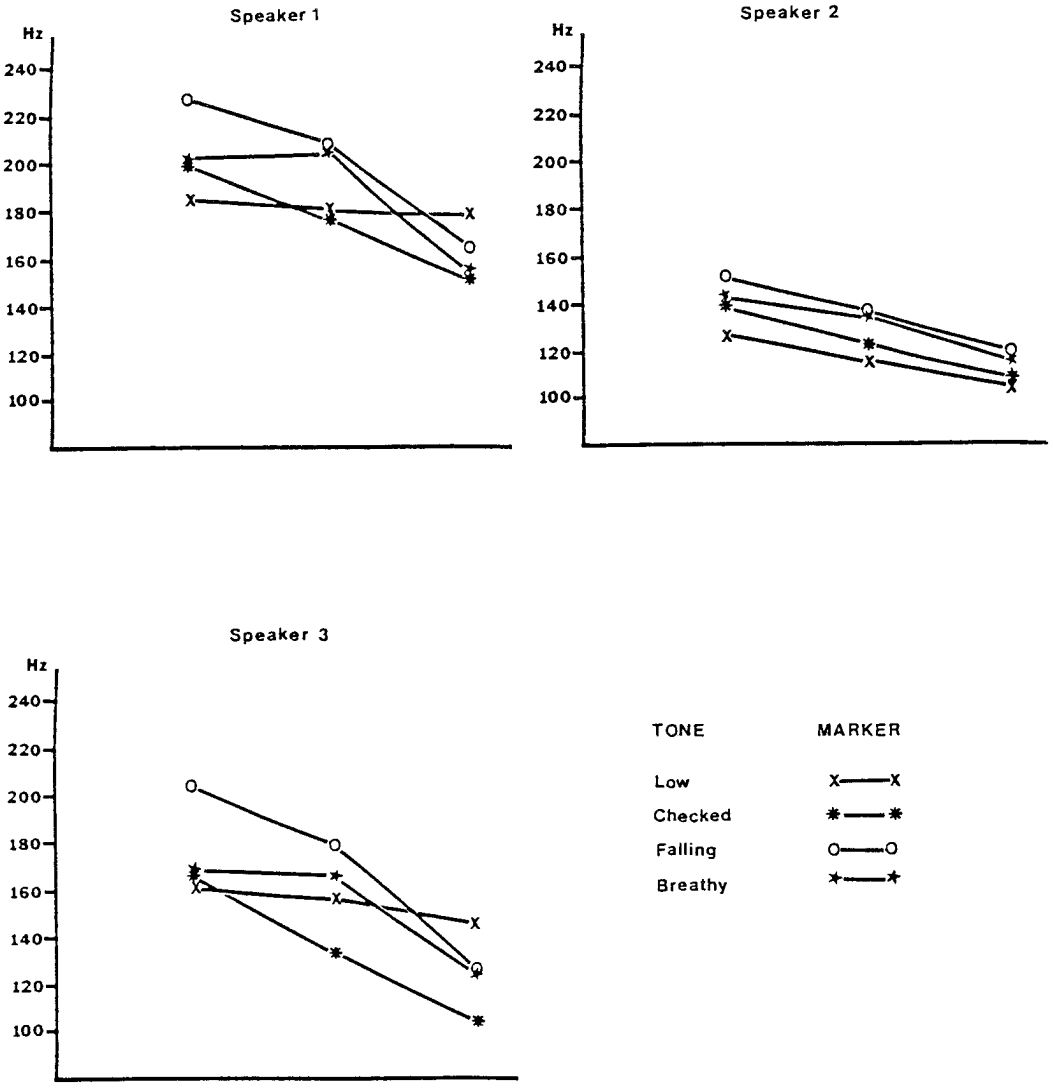


Figure 5. Pitch contours measured at onset, midpoint and offset, averaged over six tokens per tone, for each speaker.

duration is not consistently distinctive. Secondly, pitch contour as measured at onset, offset and midpoint, is generally different for the four tones studied here. However, individual onset, offset or midpoint values of the four tones are sometimes very similar. Since laryngeal adjustments for pitch variation may affect the glottal flow waveform, pitch differences may be reflected in our measures of phonation types. This possible interference by pitch provides a good test of the effectiveness of those measures, as will be seen below.

4.2 Inverse Filtering

The output of the inverse filtering procedure is a waveform representing airflow through the glottis. In the introduction we outlined ways of characterizing differences in glottal flow waveforms. The following two sections present results of analyses of glottal waveshape and spectral properties of the Hmong vowel tokens.

4.21 Waveshape Properties

One way of looking at properties of the glottal source is to consider glottal flow in the time domain. As was mentioned earlier, recent linguistic studies using inverse filtering (Bickley 1982, Javkin and Maddieson 1983) suggest a number of measures of glottal flow waveforms which can distinguish breathy, normal and creaky voice qualities. Some of these are properties of timing such as proportions of the pulse taken up by open and closed phases of the glottal pulse. Others are measures of amplitude changes over time, such as estimates of sharpness of opening and closing branches of a glottal flow pulse. For a general idea of the way pulse shapes can vary, see Figure 6, where samples of normal and breathy voice waveforms are presented.

We considered two measures likely to reflect glottal waveshape differences attributable to phonation contrasts. The first measure used was the ratio of the duration of the closed phase of the glottal flow pulse to the total duration of the pitch period, henceforth the closure duration ratio. Bickley (1982) reports that in Gujarati normal voiced vowels, the closed phase occupies about one-third of the total pitch period, but a smaller proportion in breathy vowels. As was discussed above, such differences in closure duration can be taken as indications of degree of glottal stricture.

Our review of the literature also suggested that vocal fold closing speed as reflected in the flow pulse falling branch could be taken as an indirect measure of glottal stricture differences across phonation types. However, the problems which Javkin and Maddieson report with their falling branch slope measure suggest that this is not a dependable measure of glottal flow differences between phonation types. Therefore, for our second measure we considered the ratio of slopes of the falling and rising branches of the open phase of the glottal pulse.

While this measure allows control for differences in flow pulse amplitude, it gives us no direct indication of falling branch slope. It is in effect a measure of symmetry of the open phase of the glottis. As was mentioned earlier, open phase symmetry is a property of glottal flow waveforms which has been reported to vary with the breathy/normal phonation contrast (Bickley 1982). Furthermore, Fant (1980)'s voice source model treats symmetry as a factor in production of differences in prominence of low harmonics, one spectral property reported to differ between breathy and normal phonation.

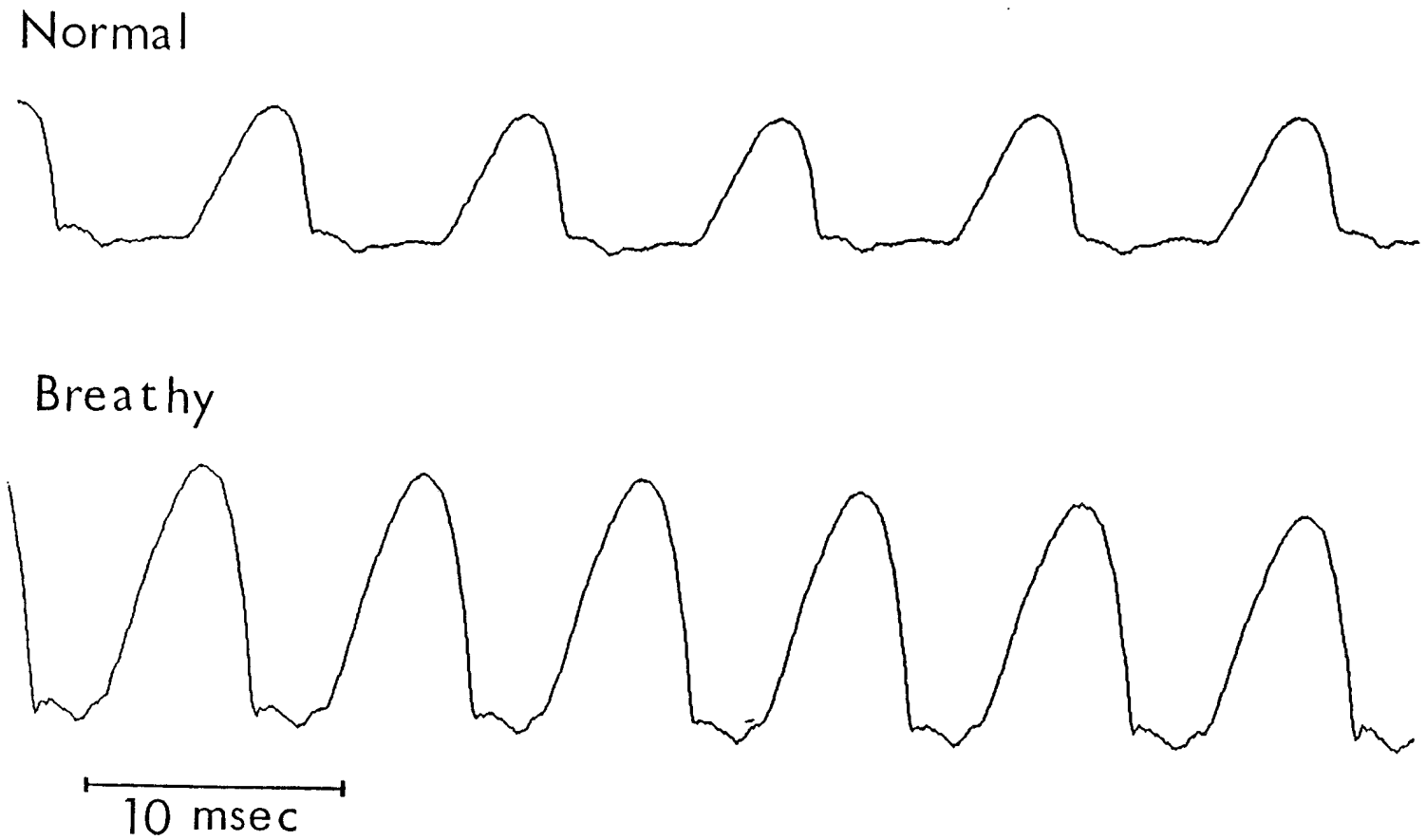


Figure 6. Sample glottal flow waveforms

Glottal flow waveforms were measured with the aid of the Glotta program developed at UCLA (see Javkin et al 1985, for more discussion). Before being input to the Glotta program, the flow waveform files were reviewed and the opening and closing points of the glottal pulse were identified for the six steady state pulses chosen earlier. Figure 7 illustrates the marking process on a normal voice vowel. The program outputs several parameters, including the slopes of the falling and rising branches of the open phase, and the durations of the closed phase, the rising branch and the falling branch. The rising and falling slopes were determined from the slopes of tangents to the rising and falling branches of the flow pulse. These tangents were estimated using the average derivative over 7 points in the middle of the rising branch and falling branch. The closure duration ratio and the slope ratio were calculated for five pulses of each Hmong token. From these values, mean duration and slope ratios were computed for each token.

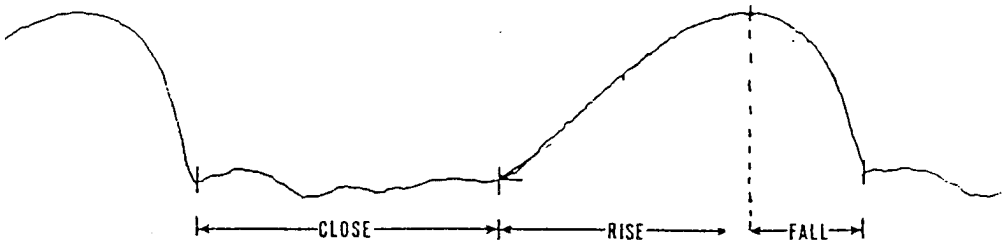


Figure 7. Waveform marking

Figure 8 shows means and standard deviations of the duration ratios pooled across speakers for all four tones. The non-breathy tones are very close in value, with a closed phase duration at about 40% of the pitch period. The breathy tone, on the other hand, has a closure duration about half this, at slightly less than 20% of the pitch period. Analysis of variance found the relative closed phase duration to be a significant indicator of differences between phonation types, again to a significance level of at least .01. Thus while none of the non-breathy tones were distinguishable from each other on this measure, all had values significantly different from the breathy tone. The same is true for speakers 2 and 3 taken individually. For speaker 1, while the non-breathy tones were not distinguished from each other on this measure, only the checked and falling tones had closure durations significantly different from the breathy tone.

The ratio of falling to rising slopes was a less effective diagnostic of phonation differences. Figure 9 shows mean values and standard deviations for each tone, across the three speakers. Contrary to our expectations, the breathy tone does not have a shape particularly more symmetrical than the non-breathy tones. An analysis of variance showed that none of the tones were significantly different from each other on this measure. The slope ratio measure is equally ineffective when taken as a variable in analyses of variance for individual

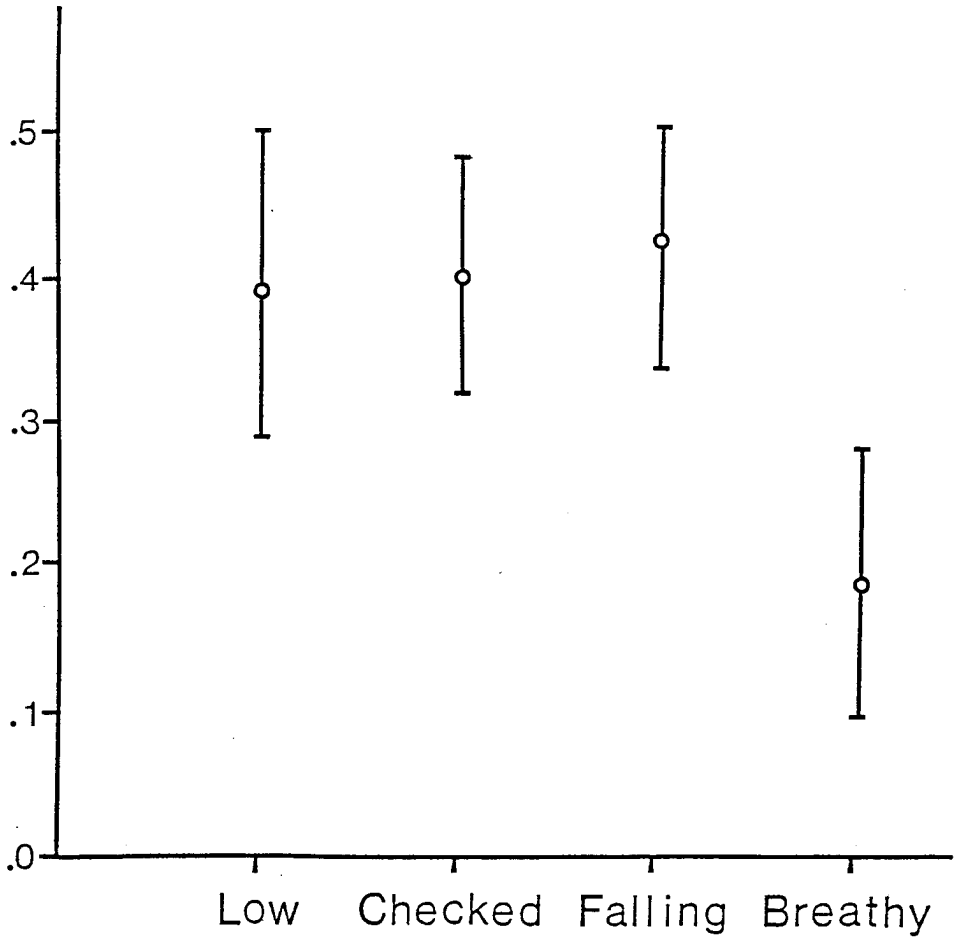


Figure 8. Closure duration ratios

Figure 9. Slope ratios

speakers, and thus less tokens. On the same significance criterion, for no speaker did the ratio of falling to rising slopes distinguish any of the four tones.

4.22 Discussion

What can these results tell us about physiological parameters of phonation contrasts? Symmetry of the glottal pulse open phase as measured here was not significantly different between the breathy and non-breathy tones. The same lack of significant differences was evident for individual speakers as well. On the other hand, the ratio of closed to total glottal pulse duration has been shown to be successful at distinguishing Hmong breathy and non-breathy vowels. By its nature, the measure controls for several other linguistic attributes of the vowels that a speaker could have varied independently of phonation types. As with other measures of the output of inverse filtering, vowel formant differences are controlled for. Pitch differences (differences in total pitch period duration) are controlled for by the fact that the measure is a ratio, of closed to total duration. In addition, vowel duration differences are irrelevant to this measure. The success of the ratio of closed to total duration in distinguishing phonation differences suggests that it is a property to consider in looking for evidence of a physiological parameter of phonation types.

Recall that the ratio of closed to total duration was about .40 for normal voice and .20 for breathy voice. At a low fundamental frequency of 100 Hz. this would mean a difference in closure duration of approximately 2 msec. Van den Berg (1958) says that it requires on the order of 5 msec for a complete chain of motor and sensory impulses to pass between the brain and the larynx. Clearly, a speaker cannot be controlling the duration of the closed phase directly. However, a speaker could be controlling one or more physiological variables which would produce this effect. We have argued above that changes in glottal stricture could contribute to such differences in closure duration of the vocal folds.

As for the question of consistency in phonation contrast production, Figure 8 indicates that there is quite a bit of variation in flow pulse closure durations. However, before we can interpret the importance of this variation we need to know how direct the relation is between vocal fold stricture and flow pulse closure duration. Rothenberg (1981) suggests that there are close parallels between properties of glottal flow waveforms and laryngographic measures of vocal fold closure. Models of vocal fold vibration and vocal tract aerodynamics should prove helpful in this regard.

It should also be noted here that the Hmong checked tone was not distinguished from the other three tones by either the pulse symmetry measure or the closure duration ratio. With respect to these measures, the checked tone acts like the other non-breathy tones, which are produced with normal voice. In addition to this fact, there is spectrographic evidence that we are justified in considering the checked tone to be produced with normal phonation. Examination of wideband spectrograms reveals that checked tone syllables have final glottalization. Recall that Smalley (1976) described the checked tone as ending in a glottal stop. This leads us to expect the first part of the vowel to have normal voice quality, with only the end of the vowel being glottalized; i.e. characterized by low frequency, often irregular taps of the vocal folds as they come together for the glottal stop. This is in fact how the checked tone sounds in the pronunciation of our 3 speakers. Spectrograms of the 18 checked tone

tokens show 14 of them ending with two to six glottalized pulses, where "glottalized" describes widely and often irregularly space voicing striations. In contrast, only 4 of the 54 non-checked tokens showed even slight glottalization, of one or two pulses at the end of the vowel.

4.23 Spectral Properties

In previous sections we have discussed how degree of glottal stricture might affect glottal closure duration and strength of the glottal closing gesture. In addition, we argued that strength of the glottal closing gesture should be reflected in the glottal flow spectrum. As has been noted, a number of investigators have found, using audio (sound pressure) rather than flow recordings, that the prominence of the fundamental frequency (first harmonic) relative to the second or other low harmonics served to distinguish breathy, creaky, and normal phonation types in (a few) natural languages. We expect that similar differences of spectral balance will be detectable in spectra of glottal flow waveforms of Hmong breathy and non-breathy vowels. In the introduction we predicted that, compared to normal voice, breathy voice tokens would have a more prominent fundamental and weaker higher harmonics. This notion is similar to van den Berg's contrast between weak, broad "puffs" which generate few higher harmonics, and strong, short "puffs" which generate many higher harmonics.

Two kinds of measures were chosen for analysis of glottal spectra. The first was a general measure of spectral tilt. Calculated as the slope of a line fitted to spectral peaks, spectral tilt was measured for two frequency intervals: 0 to 1000 Hz and 1000 to 2000 Hz. The second kind of measure was a more localized indication of spectral structure, namely the difference in relative amplitude of the fundamental and the first harmonic. The first step in spectral analysis was to calculate Fast Fourier Transforms (FFT's) for the glottal waveforms of all of the Hmong vowel tokens. A Hamming window 25.6 msec long was fitted over the middle of the 6 glottal pulses chosen earlier. Care was taken to position the window over the pulses so as to minimize the difference in the amplitude of the signal at the beginning and end of the window (i.e. to minimize discontinuities in what FFT analysis will treat as a repeated series of pulses). Because of differences in fundamental frequency, the number of pulses analyzed varied from 3 to 6. Figure 10 shows sample spectra of a normal and a breathy voice token.

Both of our measures were determined from the output of the FFT analysis. FFT analysis outputs spectral components at fixed intervals. These points may or may not correspond to actual harmonic peaks. Therefore peaks were estimated using a parabolic interpolation routine. The routine first finds a local maximum, determines that there is only a single reversal of slope over a range of five points centered on this maximum, and then fits a parabola to the three central points.

Our general measure of spectra refers to spectral tilt between 0 and 2000 Hz. Given the low range of flat frequency response of the Rothenberg mask, it is unlikely that results for frequencies much above 2000 Hz can be considered accurate. Within the two 1000 Hz intervals chosen, a linear regression routine was used to fit lines to the harmonic peaks identified by the parabolic interpolation routine described earlier. The literature suggests that the spectrum of the glottal source exhibits logarithmic roll-off (for example, van den Berg 1958, Fant 1980). This implies that a correlation should be detectable between intensity and frequency on logarithmic scales. To test this hypothesis,

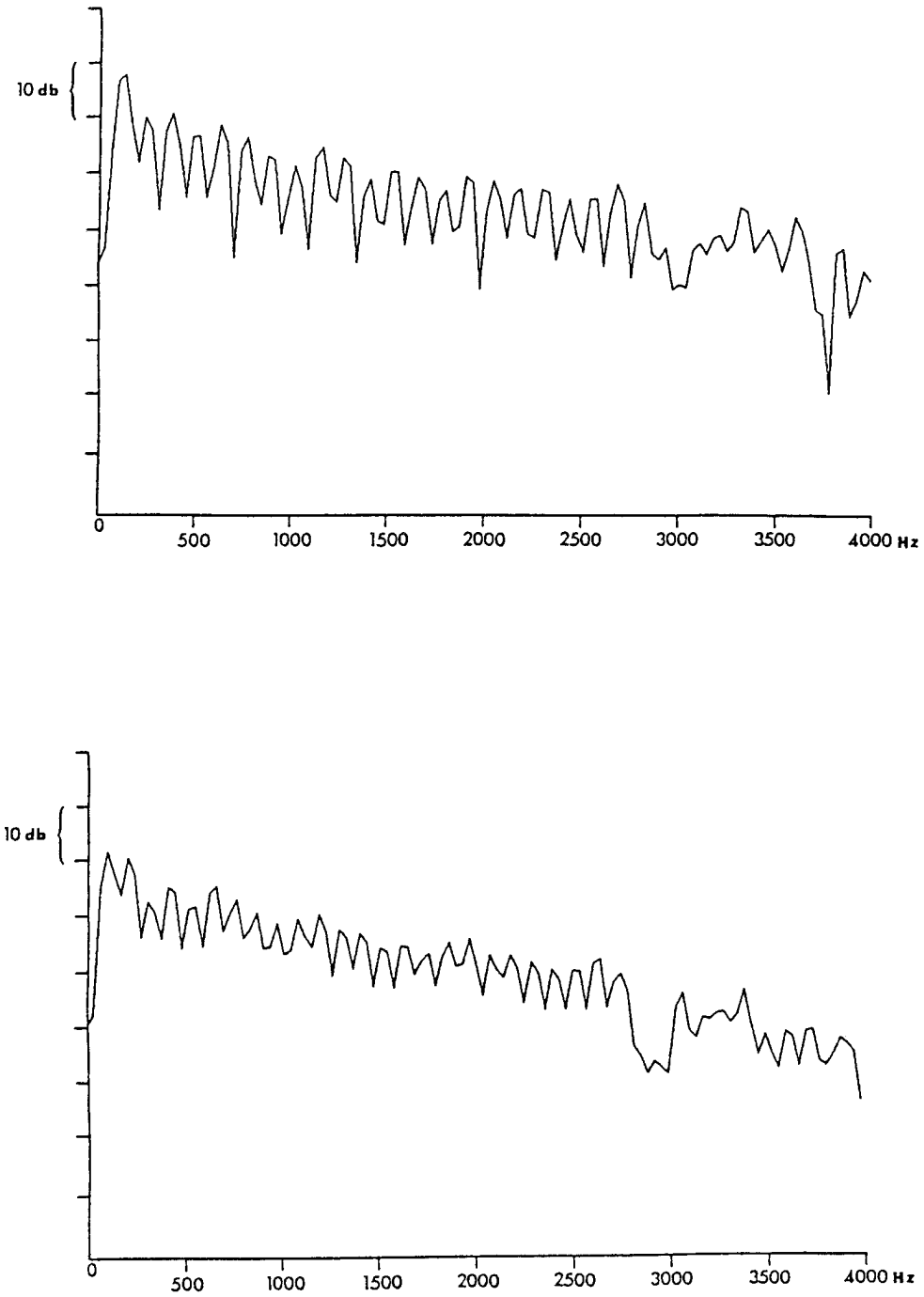


Figure 10. Glottal spectra

the lines fitted to spectral peaks had slopes defined with frequency on both linear and logarithmic scales. The result of a two-tailed T-test for linear relation was also computed for each of these lines.

With frequency on a linear scale, while there is a significant correlation between intensity and frequency 68% of the time for peaks within the 0-1000 Hz interval, this is true only 22% of the time for peaks within the 1000-2000 Hz interval. The spectral tilt measure does not show any sensitivity to phonation type. Thus for the 0-1000 Hz interval, the 23 of 72 tokens showing a lack of a linear relation include all breathy and falling tokens for speaker 1, as well as one or two tokens of each tone type for speakers 2 and 3. Since slopes could be calculated reliably for only about two-thirds of the tokens for peaks between 0 and 1000 Hz, this measure is not a good candidate for characterizing data of this sort.

With frequency on a logarithmic scale, this measure fared even worse. Within the 0-1000 Hz interval, T-tests for linear relation were significant for only 19 of 72 tokens, roughly 26%. As before, the tokens lost show no obvious pattern -- for the three speakers, tokens of all four tone types failed to be characterized by this measure.

Figure 11 shows mean values and standard deviations averaged over speakers for another measure of spectral balance, namely the difference in amplitude of the first and second harmonics (F0-H2). The non-breathy tones have very similar values, all clustering around slightly over 2 dB. The breathy tone shows a much higher difference between F0 and H2, with a mean of 9.48 dB.

An analysis of variance found the F0-H2 measure to be as successful at distinguishing phonation differences as the closed phase duration ratio. While F0-H2 was not significantly different among the non-breathy tones, the breathy tone was distinguished from each of the non-breathy tones on this measure. Analyses of variance on this variable for individual speakers had the same result.

4.24 Discussion

In testing for spectral differences of glottal flow between phonation types, we found that a general measure of spectral tilt over 1000 Hz intervals could not characterize all breathy and normal tokens reliably. With this procedure, lines could not be fitted to these peaks with confidence, whether frequency was considered on a linear or a logarithmic scale, for breathy or normal voice vowels. These results point out that care should be taken in accepting generalizations in the literature about the logarithmic roll-off characteristics of the glottal spectrum. They are at best estimations.

Our other spectral measure focussed on the prominence of the fundamental relative to the second harmonic. This F0-H2 measure was very successful at distinguishing breathy and non-breathy tones of Hmong. Like the closure duration ratio, the F0-H2 measure controls for pitch, duration and quality of the vowel tokens analyzed. Pitch, meaning here the fundamental frequency, or rate of vibration of the vocal folds, affects this measure directly only to the extent that it influences the number of glottal pulses analyzed with the 25.6 msec window for FFT analysis. However, if pitch were having an influence here, we should see a differentiation between normal voice low and higher pitch tones. The F0-H2 measure does not show any such effect. Instead, the categories of

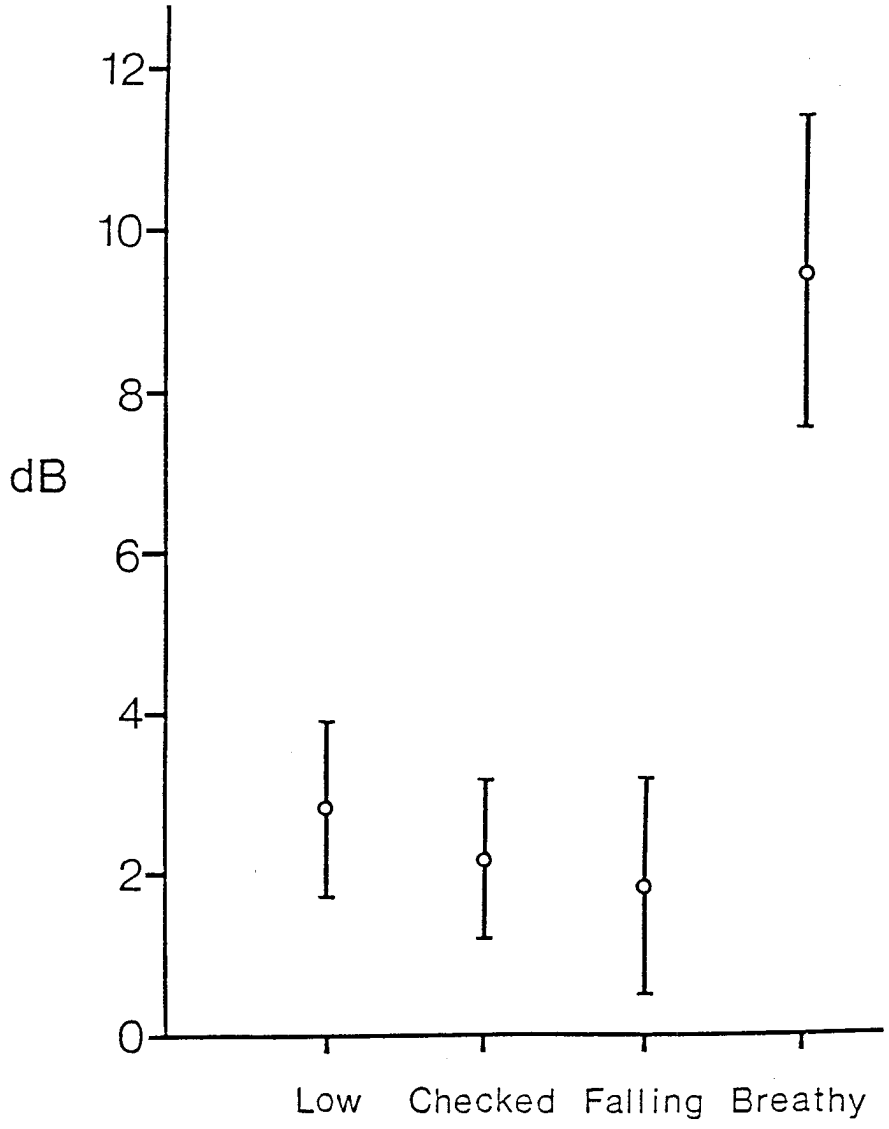


Figure 11. Amplitude difference between F0 and H2

F0-H2 values correspond exactly to the breathy/normal phonation contrast within the Hmong tokens.

5. Conclusions

In the Introduction we outlined several properties which could be distinctive for phonation types and thus relevant to their phonetic characterization. What have we learned about Hmong in this regard? In the Hmong data, two of the properties we discussed are effective in distinguishing breathy and normal voice vowels, and two are not. On the one hand, there is both a spectral difference and a pulse shape difference in glottal flow waveforms of contrastively phonated vowels. That is, the prominence of the fundamental relative to the second harmonic is a very significant acoustic correlate of the breathy/normal phonation distinction, and the proportion of the glottal flow pulse taken up by the closed phase is also significantly different for the breathy and normal tokens. These two quantities distinguish phonation types in Hmong without significant interference from pitch differences between the tones. On the other hand, flow pulse symmetry is not significantly different for breathy and normal voiced vowels in Hmong, and spectral tilt could not be used to make meaningful comparisons between phonation types because reliable measures could not be obtained for many breathy and normal tokens.

Having narrowed the field of possible distinctive measures of phonation types in Hmong, we will discuss what the successful and unsuccessful measures outlined above can tell us about the phonetic characterization of breathy and normal phonation in Hmong, with an eye to implications for cross-language investigation.

To adequately characterize a phonetic phenomenon we need to relate its acoustic attributes to the physiological gestures which produce them. One way to do this for phonation types is to relate their acoustic attributes to laryngeal gestures via investigation of the glottal flow waveform. We will first consider the relationship between acoustic properties and the glottal flow waveform. Previous studies of linguistic phonation contrasts identified spectral prominence of the fundamental as an important cue to phonation type. The Hmong data show thàt, for the breathy/normal voice contrast, there are quantifiable and significant differences in prominence of the fundamental in glottal spectra as well. Furthermore, these spectral contrasts co-occur with significant differences in flow pulse closure duration, such that when closure duration is smaller, the fundamental is more prominent. The implied inverse relationship between flow pulse closure duration and prominence of the fundamental relative to the second harmonic is the beginning of a link between the glottal flow waveform and acoustic attributes of phonation types.

Recall that there is already in the literature a model of the relation between waveshape and spectral properties of glottal flow for normal voicing. Fant (1980)'s voice source model was developed to account for relationships between glottal waveforms and spectra based on evidence from normal phonation at different intensity levels. We will discuss this model briefly in light of the Hmong data on waveshape and spectral properties of breathy and normal phonation.

Fant's model is based on three parameters: peak flow, glottal frequency (equivalent to the inverse of twice the duration of the rising branch of the flow pulse) and k, a measure of symmetry between closing and opening speeds of the vocal folds. Quantities like closure duration fall out from the interaction of

these other variables. Variation in the three parameters also changes the relative strength of different spectral components. So for example, if glottal frequency and peak flow are held constant, a low value of $\underline{k}$ (making the flow pulse more symmetrical, i.e. less skewed to the right) results in a short relative closure duration and weak higher harmonics.

To derive the results we reported for Hmong, a model like Fant's needs to be able to produce amplitude differences between F0 and H2, while generating the appropriate closure durations for the two phonation types. However, these amplitude differences cannot be considered dependent on flow pulse symmetry, since symmetry is non-distinctive across tones in Hmong. So for example, varying $\underline{k}$ as just described is not an appropriate way of accounting for the Hmong facts. There is no way to vary relative closure duration without also varying symmetry.

In Fant's model there are ways of generating differences in spectral balance without essential reference to symmetry. The strength of low frequency harmonics can be increased by increasing the "total air volume of the vocal pulse" (Fant 1980:2). As illustration, Fant suggests that a high amplitude (peak flow) pulse that is also more symmetrical (thus broader) is consistent with a relative spectral dominance of the fundamental. For the Hmong data, we might say that a relatively high amplitude pulse that is also proportionately broad (but not necessarily more symmetrical) has a dominant fundamental. Varying the model's parameters in this way would only work for Hmong if resultant changes in the dominance of the fundamental hold with respect to the second harmonic in particular. If Fant's model can be extended in this way, then it is possible that it can be used to characterize waveshape and spectral properties of flow for breathy and normal phonation across languages.

What about the relationship between flow waveshape properties and laryngeal gestures? Fant's model is of little assistance in this respect in that the parameters used are arbitrary properties of waveshape that cannot be simply related to physiological quantities. The physiological dimension(s) underlying flow pulse closure variations are not yet known. In this regard the Hmong results are of interest because, though the differences in duration ratio for breathy and normal voice vowels in Hmong are statistically significant, they are of an order of magnitude so small that they could not be controlled directly. It is possible that, as has been discussed in the literature for a long time, variations in glottal stricture are contributing to these closure duration differences. Modelling studies of vocal fold vibration may clarify the effect which glottal stricture and other factors like vocal fold tension can have on flow pulse closure in phonation contrasts like those in Hmong.

Finally, considering Hmong in conjunction with what we know about the breathy/normal phonation contrast in other languages, what are the chances of identifying a cross-language measure of the breathy/normal contrast? The phonetic characteristics found successful at distinguishing phonation types in Hmong are basically consistent with previous findings. For example, the proportionately shorter closed phase found for Hmong breathy versus normal vowels has also been reported for Gujarati (Bickley 1982). Laryngographic studies of non-linguistically significant breathy versus normal vocal fold vibration (Laver 1980, Fourcin 1981, Rothenberg 1981) also note this same pattern in closure duration.

Similarly, the strong difference in amplitude of F0 and H2 in breathy vowels of Hmong (9 dB on average) is paralleled in non-inverse filtered !Xóõ data

(Bickley 1982). There, despite the possible influence of the supralaryngeal tract on the spectrum, average F0-H2 is 7 dB. The smaller difference in F0-H2 found for Hmong normal voice vowels (about 2 dB) is of the same magnitude as that found for !Xóõ normal voice vowels (-2.98 dB). The essential properties of the F0-H2 measure of phonation differences are that the value will be positive for breathy vowels, and greater in magnitude for breathy vowels than for normal voice vowels.

Because they have parallels in other languages, both the spectral measure and the flow pulse closure measure used in analyzing Hmong are strong candidates for dimensions of phonation type which can be compared cross-linguistically. They are quantifiable means of comparing phonetic properties of breathy versus normal phonation, and thus their significance can be tested. In the Hmong and !Xóõ data vowel quality influences on the spectrum do not significantly affect how well the F0-H2 measure distinguishes phonation types. However, Maddieson and Ladefoged (1984) show that for one speaker of Yi, when comparing breathy and more tense /e/, the influence of supraglottal resonances on the spectrum was strong enough that not only was there a greater F0-H2 for the tenser vowel, the relation of F0-H2 values for tense and lax (breathy) vowels went in the opposite direction of what we would expect. That is, H2 was higher than F0 for the breathy vowel. So it may be that we need the normalizing effect of inverse filtering if we want to use F0-H2 to compare measures of breathiness across languages. Inverse filtering is clearly required for studying flow pulse closure duration. This is a valuable line of research to pursue, since the results can be related to laryngographic measures of vocal fold contact duration (Rothenberg 1981). Thus this measure brings us one step closer to being able to determine the nature of vocal fold behavior in producing phonation contrasts. In conclusion, the analysis of phonation types in Hmong suggests that we are on the right track in our attempts to characterize essential acoustic and physiological properties of breathy and normal phonation contrasts in language.

ACKNOWLEDGEMENTS

This work was supported by NIH grant 18163-02 to Peter Ladefoged. Members of the UCLA Phonetics Laboratory Group have been of assistance on many fronts; particular thanks to Peter Ladefoged and Ian Maddieson for help in data collection.

References

Berg, J. van den. (1958). "Myoelastic-aerodynamic theory of voice production. Journal of Speech and Hearing Research 1: 227-244.

Bickley, C. (1982). "Acoustic Analysis and Perception of Breathy Vowels," MIT Working Papers in Speech Communication, Vol. 1: 71-80.

Bless, D. and J. Abbs, (eds.). (1983). Vocal Fold Physiology. San Diego: College Hill Press.

Catford, J.C. (1977). Fundamental Problems in Phonetics. Indiana University Press: Bloomington.

Fant, G. (1980). "Vocal source dynamics." Speech Transmission Laboratory-Quarterly Progress and Status Report, 2-3: 17-37.

Fischer-Jørgensen, E. (1967). "Phonetic Analysis of Breathy (Murmured) Vowels in Gujarati." Indian Linguistics, 28: 71-139.

Fourcin, A. J. (1974). "Laryngographic Examination of Vocal Fold Vibration." in B. Wyke (1974): 313-333.

Fourcin, A. J. (1981). "Laryngographic Assessment of Phonatory Function." in C. L. Ludlow and M. O. Hart (1981): 116-124.

Hollien, H. (1984). "In Search of Vocal Frequency Control Mechanisms," in D. Bless and J. Abbs (1983), pp. 361-367.

Javkin, H., Antoñanzas-Barroso, N., and I. Maddieson (1985) "Digital Inverse Filtering for Linguistic Research." UCLA Working Papers in Phonetics, 60:87-100.

Javkin, H. and I. Maddieson (1983). "An Inverse Filtering Study of Burmese Creaky Voice." UCLA Working Papers in Phonetics, 57: 115-125.

Keating, P. A. (1984). "Phonetic and Phonological Representation of Stop Consonant Voicing." Language 60, 2: 286-319.

Kirk, P., Ladefoged, P. and J. Ladefoged (1984). "Using a Spectrograph for Measures of Phonation Type in a Natural Language." UCLA Working Papers in Phonetics 59: 102-113.

Ladefoged, P. (1983) "The Linguistic Use of Different Phonation Types," in D. Bless and J. Abbs (1983), pp. 351-360.

Laver, J. (1980). The Phonetic Description of Voice Quality. Cambridge: Cambridge University Press.

Ludlow, C. L. and M. O. Hart (eds.). (1981). Proceedings of the Conference of the Assessment of Vocal Fold Pathology (Bethseda, Maryland; April 1983). The American Speech-Language-Hearing Association, Rockville, Maryland.

Lyman, T. (1974). Dictionary of Mong Njua. The Hague: Mouton.

Maddieson, I. and P. Ladefoged. "'Tense' and 'lax' in four minority languages of China." UCLA Working Papers in Phonetics 60: 59-83.

Rothenberg, M..(1973). "A new inverse filtering technique for deriving the glottal air flow waveform during voicing." Journal of the Acoustical Society of America. 53: 1632-1645.

Rothenberg, M. (1981). "Some Relations between Glottal Air Flow and Vocal Fold Contact Area." in C. L. Ludlow and M. O. Hart (1981): 116-124.

Smalley, W.A. (1976). "The Problems of Consonants and Tone: Hmong (Meo, Miao)," in W.A. Smalley, ed. Phonemes and Orthography: Language Planning in Ten Minority Languages of Thailand. Pacific Linguistics Series C, No. 43, Canberra: Australian National University.

Wyke, B. (1974). (ed.) Ventilatory and Phonatory Control Systems. London: Oxford University Press.

Testing a model of intonation in a tone language.

Mona Lindau

Department of Linguistics, University of California, Los Angeles, Ca 90024
and
Phonetics Institute, University of Lund, Lund 223 62, Sweden

Abstract: Fundamental frequency curves of simple statements and questions are generated for Hausa, a two tone language of Nigeria, using a modified version of an intonational model developed by Gårding and Bruce. In this model, rules for intonation and tones are separated. Intonation is represented as grids of (near)parallel lines, inside which tones are placed, and modifications on tonal placements are made. All points inside a grid are concatenated into the final F0-curve. The slope of the grid depends at least on sentence type (statement or question), sentence length, and tone pattern. The model was tested by reference to data from nine speakers of Kano Hausa.

This study reports a test of a formal model of the relation between tone and intonation, using Hausa, a Chadic language spoken in Nigeria. Hausa has a relatively simple tone system of basically two tones, High and Low. There is also a falling tone that can be analyzed as a combination of a High and a Low, but it will not be considered further here (Meyers 1975, Cowan and Schuh 1976). In this text low tone is marked with a grave accent, ` , and high tone is unmarked. The goal of this study is to test a model of intonation by matching fundamental frequency curves with curves generated by rule for Hausa statements and questions on two different tone patterns. The intonational model is based on that developed in Sweden (Gårding and Bruce 1981, Gårding 1981). The results will be useful in speech synthesis of Hausa utterances. Figure 1 shows the application of the model to an ideal Hausa sentence with alternating high and low tones. The input is a phonetic transcription with major boundaries and tones marked. The model describes intonation as grids of baselines and toplines. In this description the grids are stored lexically as a basic slope for a particular number of syllables for each type of utterance, and are generated by rule. The boundary marks for the beginning and end of the sentence, as well as the Highs and Lows are inserted onto the grid lines. Local rules may apply to modify the underlying patterns. The last step is to concatenate the maxima and minima into a smooth curve.

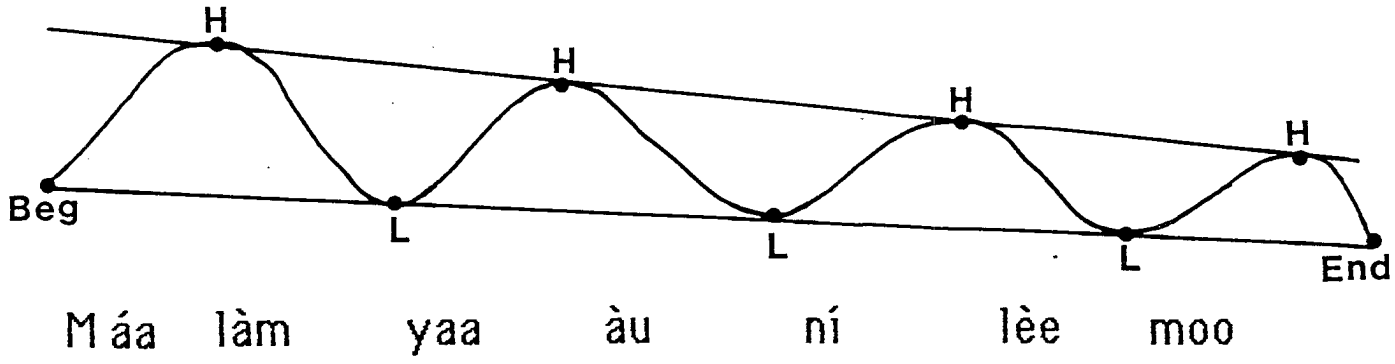


Figure 1. A model Hausa utterance.

PROCEDURE

The model can be tested by the degree to which it will account for tonal and intonational patterns used in common by different speakers. The data consist of tape recordings from nine male speakers of Kano Hausa, recorded in Kano. The speech material was designed to illustrate intonation in statements and questions in appropriate contexts, and consisted of sentences on High tones only, and on alternating High and Low tones. Each tone pattern occurred in a short and a long version. There were two short basic sentences on High tones only, "Muudii yaa zoo gidaa" (Muudii came home), and "Muudii yaa ga zoomoo" (Muudii saw a hare), and one long sentence, "Uwargidan Muudii taa ga zoomoo bayan gidaanaa" (Muudii's senior wife saw a hare behind my house). The basic short sentence on strictly alternating High and Low tones was "Maalàm yaa àuni lèemoo" (The teacher weighed the oranges), and the long sentence on this tone pattern was "Maalàm yaa àuni lèemoo gàban gàrinmùu" (The teacher weighed the oranges in front of our town). In addition there was one sentence of Highs and Lows with two adjacent High tones, "Maalàm yaa rabà naamàa", HLHHLHL, (The teacher divided the meat). Basic sentences on Low tones only were excluded by the fact that Hausa does not have any low tone verbs. The sentences were organized in blocks of one theme. Each block began with an introductory statement about a given topic; there then followed various question versions of the statement, each followed by a proper reply. This study examines a subset of the data, consisting of the introductory statements, the yes/no questions, and a question with the question-word "waa?" (who?).

The sentences were read by the speakers for the recording. The style of speech is fairly formal, but a non-colloquial style was intended, so as to maximize contrasts at this exploratory stage of the investigation. The speakers did not always produce the designated prosodic patterns on each utterance, so the results may reflect less than nine speakers. The number of speakers analyzed is specified in each case in describing the results. Particularly the long sentences were affected in that some speakers preferred breaking them up into smaller phrases with pauses. Only the long sentences without such pauses were included in the analysis.

The recorded material was analyzed from wide and narrow band spectrograms made on a digital sound spectrograph. Wideband spectrograms were used to aid segmentation and for durational measurements. The fundamental frequency curves were traced from the 10th harmonic. One advantage of tracing pitch on spectrograms, rather than from pitch extracted by computerized methods, is that a pitch curve is visible even when the intensity drops at the end of an utterance, or at creaky phonation. Several of the speakers had quite creaky phonation, particularly at the end of sentences.

Measurements for all speakers were made of the fundamental frequency at the beginning points (BEG), the end points (END), and at all turning points (H, L) in statements and yes/no questions. This points are illustrated in figure 1. The model predicts that it should be possible to specify tones in terms of an intonational grid. Accordingly, straight lines were fitted through the High and Low turning points to form a grid. Each line was drawn so as to fit as many maxima or minima as possible in the utterance. Local rules predict that certain turning points may fall outside the grid, and in these cases the grid lines were drawn so as to exclude these points. The local rules considered for drawing grids on sentences include the following: In utterances with alternaing Highs and Lows the last High is often lowered to a point inside the grid. In the long sentences

on alternating Highs and Lows the penultimate High is often raised above the top grid line. In yes/no questions the last High is considerably raised above the topline. The rates of slope of the grid lines were measured and calculated in % per second. In addition, the total duration for each statement and corresponding yes/no question was measured.

RESULTS AND DISCUSSION

Consider the statements first. Figure 2 shows one speaker's pitch curves of three statements with different tone patterns and lengths.

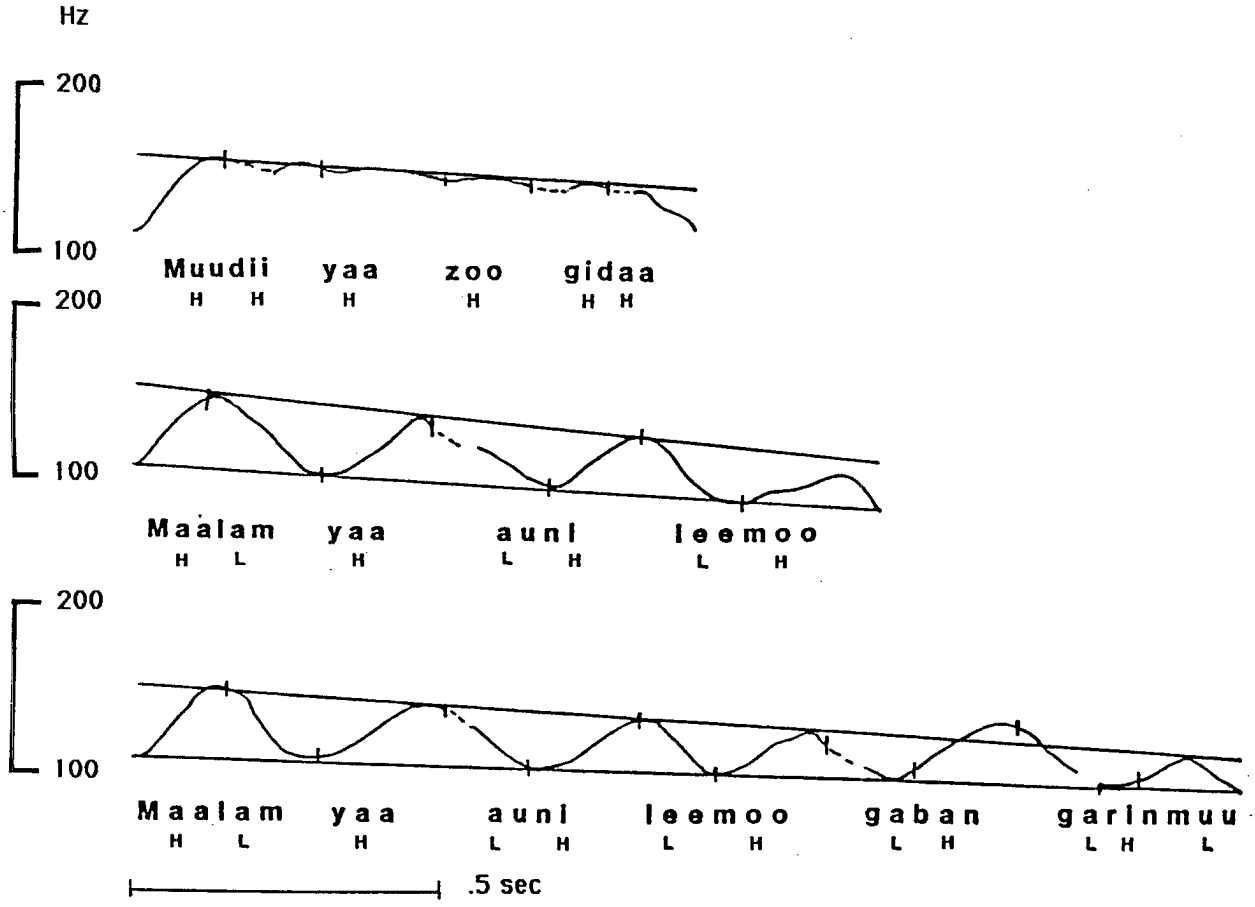


Figure 2. Pitch curves of statements on different tone patterns of one speaker. The vertical bars indicate syllable boundaries.

The statement intonation describes a downward slope, even in sentences on High tones only. In the utterances with alternating Highs and Lows the two grid lines are in a fairly constant relationship to each other throughout the utterance, except that, as predicted by the local rules, in the second sentence the final High is lowered, and in the third, long sentence the penultimate high is raised. The stable relationship between gridlines makes it possible to derive one line from the other. As there are no sentences on Low tones only, the topline is taken as basic, and the bottom line as derived. The parameters that need to be set in order to generate pitch curves for statements thus include:

1. the slope of the topline for short and long sentences.
2. the frequency of the first High.

	short sentences	long sentences
H only		
slope (% per sec.)	13.89 (4.48)	7.86 (2.9)
duration (ms)	1080 (21)	2460 (30)
	n=18	n=7
HL		
slope	32.71 (13.28)	16.29 (2.98)
duration	1310 (21)	2080 (20)
	n=7	n=7
HLHHLHL		
slope	28.75 (6.69)	
duration	1160 (15)	
	n=7	

Table I
Mean rates of (negative) slope in % per second, and mean durations in milliseconds, both with standard deviations, of the short and long sentences.

Table I shows the mean % per second rate of (negative) slope with standard deviations of the topline slopes, and the mean sentence durations. The different means of the rate of slope for the different sentences show that the slope is affected by both sentence length and tone patterns. The longer sentences have a slower rate of descent of the topline, and paired t-tests show that the differences between short and long sentences are significant ($p<0.001$). Very approximately, doubling the number of syllables will halve the rate of the slope for these data. However, as only two sentence lengths for each tone pattern were studied, a more precise relationship between sentence length and topline slope cannot be determined here. At this stage sentence length will thus have to be part of the specification of the topline. The rate of slope also varies with different tone patterns. The utterances with alternating Highs and Lows have about twice as steep a slope as those with Highs only in utterances of comparable length. Paired t-tests show that the differences in slope between utterances with different tone patterns are significant ($p<0.001$). The rate of slope of the utterance on the HLHHLHL pattern with two adjacent Highs is also significantly less than that of the strictly alternating sentence ($p<0.05$). The downward slope in statements is wellknown as DOWNDRIFT. Downdrift in Hausa and other African tone languages has sometimes been regarded as an intonational phenomenon (e.g. Hombert 1974, Ladefoged 1975, Meyers 1976), sometimes as an effect of local tone assimilations of Highs to preceding Lows in sentences with alternating Highs and Lows (Abraham 1941, Hyman and Schuh 1974). The data here show that both views are partially correct. The downward slope in the sentences on High tones only cannot be an effect of local tone assimilations, but must be considered to be a global statement intonation. The much steeper slope in the sentences with alternating High and Low tones is here interpreted as an effect of combining a global statement intonation that can be seen in sentences with Highs only, which have a

fairly slow rate of slope, with a local tone assimilation rule that lowers Highs following Lows. This view is supported by the fact that in the sentence with two adjacent Highs in the middle ("Maalàm yaa rabà naamàa"), where only two Highs follow Lows, the slope is less than in the strictly alternating "Maalàm yaa àuni lèemoo", where three Highs follow Lows. The more Low-High sequences the steeper the slope gets. In this view then, Hausa has both a global statement intonation with the slope of that on High tones only, and local tone assimilations that will change the slope depending on the number of Low-High sequences in the derivation of the final slope. The final slope can be described as an effect of both an almost universal downward slope in statement intonations, and more language specific modifications of that declination.

The topline thus varies with sentence length and tone patterns. It does not seem to be predictable from any other parameter. It is possible that the slope could depend on the frequency of some point at the beginning or end of the utterance. But in these data there is no significant correlation between the slope on the one hand, and the beginning or endpoints, or the initial or final Highs, on the other. The topline slope in statement intonation is a parameter that functions like a primitive. Barring a more precise knowledge of the relationship between sentence length and slope, it has to be set for each length in generating pitch curves.

The topline has to be anchored at some point, presumably at the beginning or end of a statement. Paired t-tests were used to determine which point differed least for the various types of statements. The first High was the only point that showed nonsignificant differences between most sentences of different tone patterns and different lengths. In the relatively few cases where the first High did differ between sentences, the difference was in the direction of this point being higher for the longer sentences than for the shorter sentences. The first High is also the point with the smallest coefficient of variation for different types of sentences within speakers. The first High will thus be used as a further set value to anchor the topline.

This is a different situation than in Swedish and in English, where the first High is higher, the longer the utterances are (Bruce 1982, Pierrehumbert 1980). In these Germanic languages the endpoint tends to be invariant, so the slope here is in a sense a function of these two points. In Hausa, however, the first High, as well as the starting point (BEG), are points that, within a speaker, show very little variation. It is possible that the differences in the behavior of the first High in different studies could be an effect of the recording situation, types of speakers, language material, or some such variable. This remains an open question.

Within each type of statement there is a great deal of variation in the topline slope between speakers, as evidenced by the fairly large standard deviations in Table I. Although a speaker will vary the topline slope with sentence length and tone patterns, the precise amount of slope appears to be idiosyncratic. The variation between speakers is not correlated with the speakers' mean fundamental frequency, or his range. In generating the grid for a statement, the mean slope will be taken to represent a typical speaker.

In the sentences with alternating Highs and Lows speakers also vary to some extent in their range. This variation probably reflects a non-linguistic factor of the attitude of the speaker. An involved speaker uses a larger range than a detached speaker (Hadding-Koch 1961, Bruce 1982). The range can be expressed as

the mean ratio between Highs and following Lows throughout the utterance. Speakers tend to keep this ratio fairly constant throughout an utterance. The mean ratio between Highs and following Lows for the short "Maalàm yaa àuni lèemoo" is 1.25 (SD=0.05, n=7), and for the longer "Maalàm yaa àuni lèemoo gàban gàrinmùu" this mean ratio is 1.2 (SD=0.06, n=5). The mean ratio is used here in deriving the bottom line from the topline in the grid.

The timing of turning points in the fundamental frequency curve in relation to the segmental structure of the utterance is also a necessary part of the specification. In sentences with alternating Highs and Lows there is a strong tendency for the turning points to occur at, and around the syllable boundaries. In sequences with open syllables, like CVV-CVV, and CV-CVV the turning point coincides the V-C boundary, but in sequences with closed syllables, CVC-CVV, the turning point occurs in the middle of the double consonant. The fundamental frequency curve associated with a High tone syllable will thus describe a movement towards a maximum during the syllable, and it will not reach its maximum until the point of transition into the next syllable. Similarly, the fundamental frequency curve associated with a Low tone syllable will reach its minimum at the border to the next syllable. The turning points in the fundamental frequency curve thus serve as a phonetic correlate to syllable boundaries in Hausa. This pattern in the timing of the turning points is relatively stable throughout an utterance and not much influenced by position in the sentence.

Figure 3 shows how a pitch curve for the High/Low sentence "Maalàm yaa àuni lèemoo" can be generated. The input is a phonetic transcription of the sentence with the lexical tones marked. 1) Set the duration for the given number of syllables (here, the seven syllables take an average of 130 centiseconds). 2) Set the value for the first High (the average speaker's value is 160 Hz). 3) Construct the intonation grid, anchored on an arbitrary (speaker-specific) value for the first High. The basic topline slope for a statement of this length is taken to be the slope for High tones only, about 14% per second. Construct the bottom line as a ratio (here 1.25) of the topline. 4) Insert low start (BEG) and end (END) points. As a convention, every statement begins and ends with the pitch for a Low, even when the first (or last) tone is a High, to (or from) which the pitch must immediately rise (or fall). 5) Insert High and Low tones on the grid lines. The Highs and Lows will constitute turning points in the intonational curve. Align the sentence so that the Highs and Lows occur at the syllable boundaries. 6) Apply local tone assimilation rules. This will result in a steeper slope, the more Low-High sequences the sentence has. For the sample sentence the rate increases from about 14% per second to 33% per second. At this stage optional rules may apply that move specific tones away from the grid lines. These processes are not found for all speakers. 7) Concatenate all the turning points into a smooth curve. The smooth trajectory is generated by interpolating between turning points, using a piecewise application of third order polynomials. Each turning point is specified as t_n, y_n, where y = fundamental frequency at time t. Each pair of points is joined by a curve, subject to the condition that the turning points have zero velocity. A program calculates the constants a_0, a_1, a_2, a_3 in a polynomial of the form

$$y = a_0 + a_1 t + a_2 t^2 + a_3 t^3$$

For these simple sentences the above rules and the interpolation procedure approximate the date well. Figure 4 shows superimposed real and model generated fundamental frequency curves for three statements. The fit between the real and the generated curves is quite good.

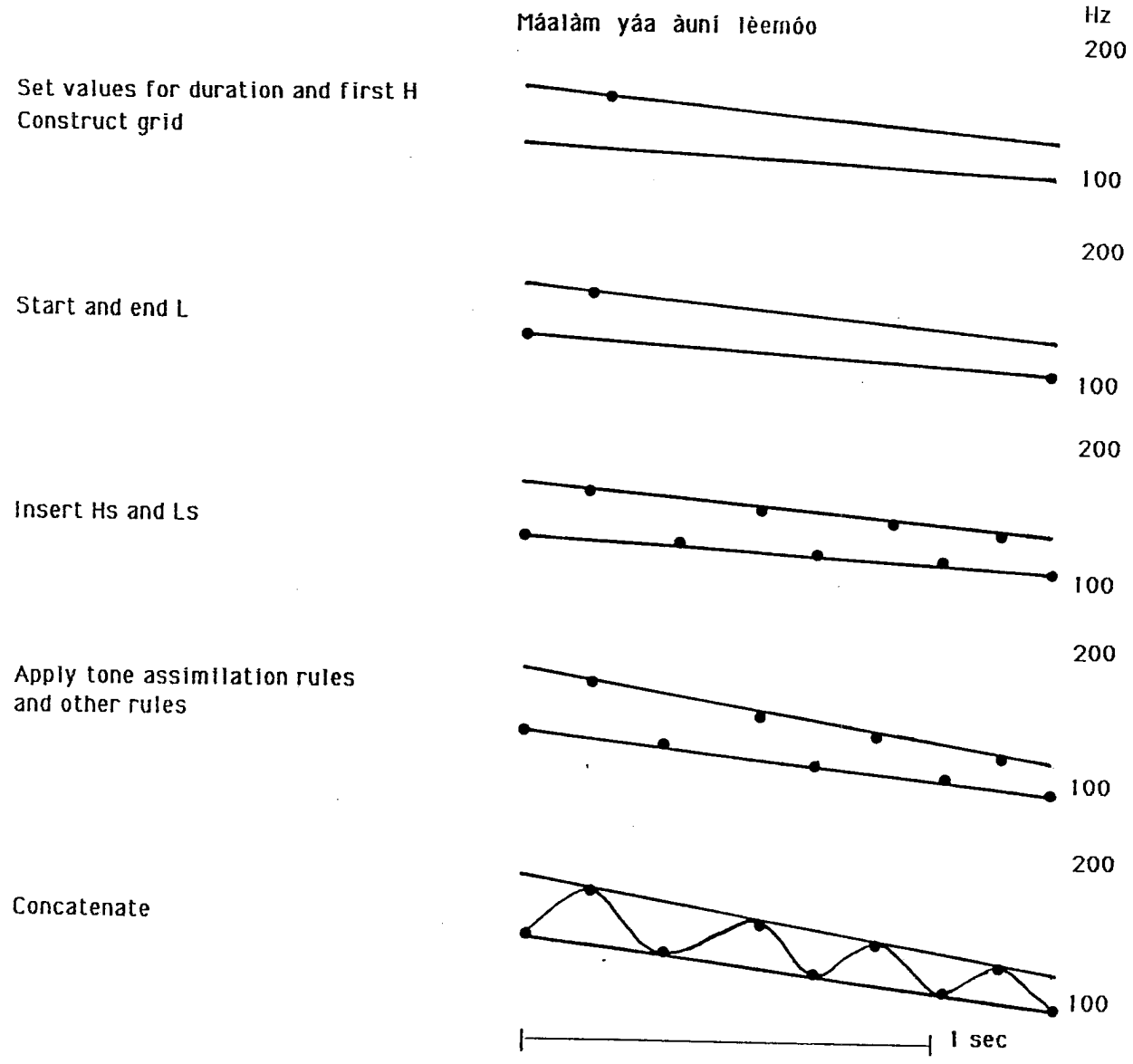


Figure 3. Rules for generating the pitch curve associated with a sentence on alternating High and Low tones.

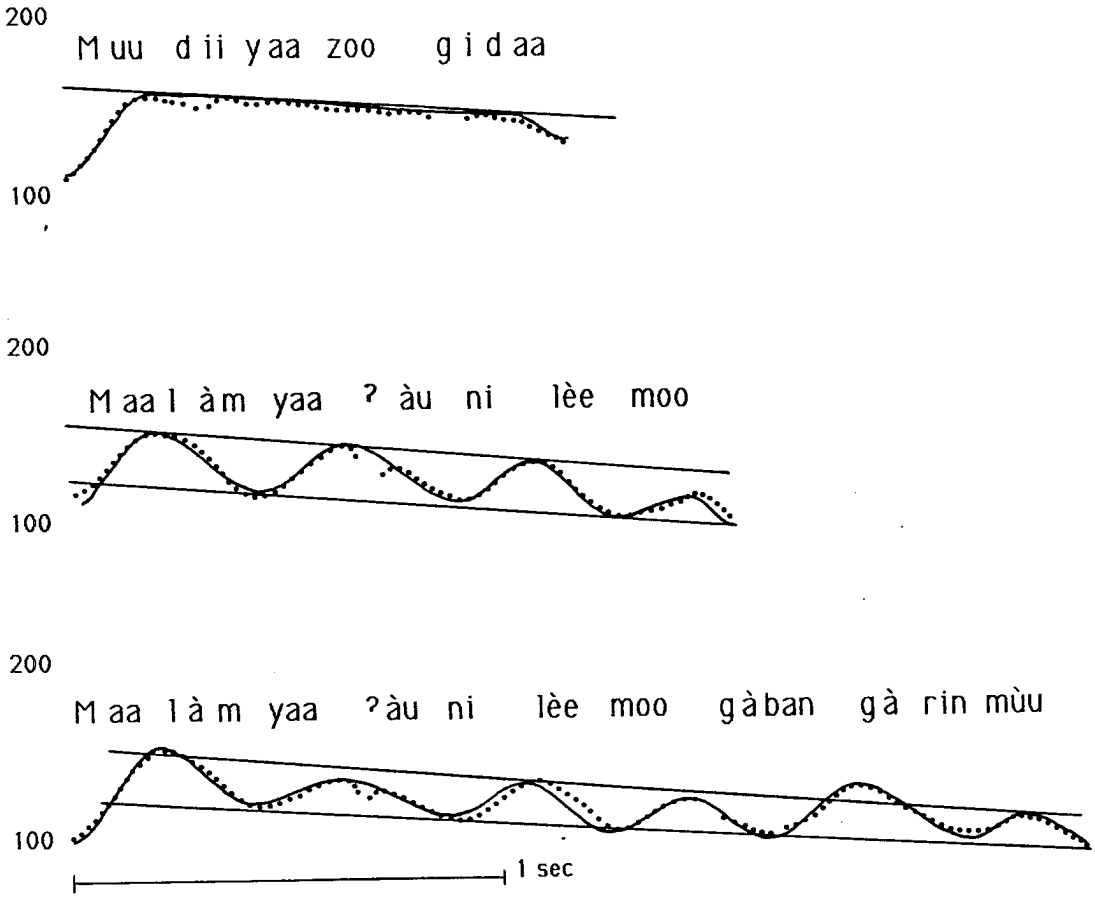


Figure 4. Superimposed real and model generated pitch curves for three statements.

Now the intonation in questions can be considered. In Hausa, yes/no questions are signalled by intonation. Cowan and Schuh (1976), and Miller and Tench (1980) characterize this type of question intonation as similar to intonation in statements, but with less of a slope, and a local rise of the final High. Newman and Newman (1981) posit a separate question morpheme at the end of all questions. This morpheme consists of a Low tone with length. The length part is justified by the fact that short final vowels lengthen in questions. The Low tone is argued for by a claim that final raised Highs are followed by a fall in questions.

Typical fundamental frequency curves of yes/no questions are illustrated in figure 5. The salient features of yes/no questions are a suspension of the statement downward slope to zero slope, and a considerably raised pitch of the last High tone. Yes/no questions on High tones only never exhibit any downward slope, and in the questions with alternating Highs and Lows only two out of nine speakers follow the pattern described in the references with somewhat of a downward slope of the grid before the final raised High. The rest have zero slope. The width of the grid is not significantly different from that in statements. The frequency of the first High tone does not differ significantly between statements and corresponding yes/no questions, so Hausa questions are not characterized by a raised register, as questions may be in other languages, for example in Swedish (Hadding-Koch 1961, Gårding 1979, Bredvad-Jensen 1983).

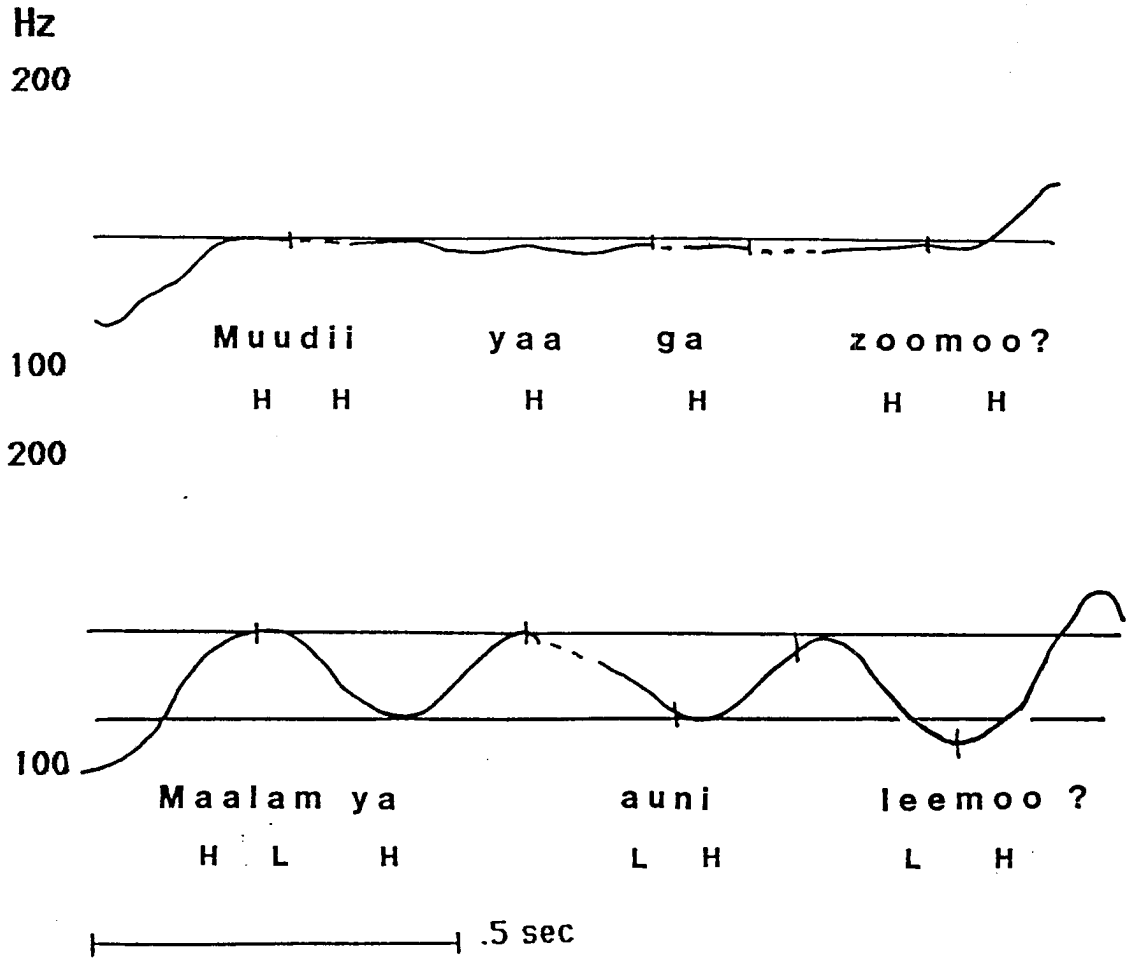


Figure 5. Pitch curves of yes/no questions on different tone patterns of one speaker.

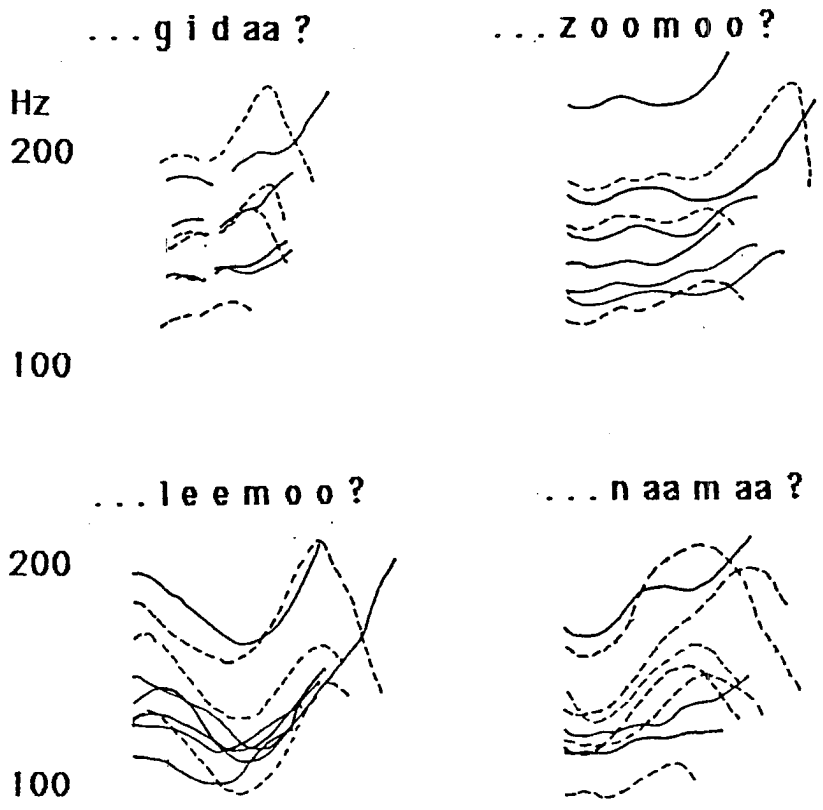


Figure 6. Superimposed pitch curves of the last two syllables of all the speakers for four yes/no questions.

The raised last High is sometimes, but not always followed by a fall. Figure 6 shows the fundamental frequency curves of the last two syllables of all the speakers superimposed for each yes/no question. The unbroken lines indicate a final rise that is not followed by a fall, and the dashed lines indicate a rise that is followed by a fall. Many speakers do not end a yes/no question with a fall. In fact, even when the last syllable is on a Low tone, as in the utterances ending with "...naamàa?", three of the nine speakers do not end in a fall. A final Low tone question morpheme is not part of a general pattern in the yes/no question. The final falls that do occur can be viewed as instances of the general convention of starting and ending low, rather than as a manifestation of an abstract Low tone question morpheme. Apparently, the convention of ending in a final low is cancelled in questions for some speakers, but not for all.

In addition, statements and corresponding yes/no questions differ significantly in total duration. The questions are about 10% shorter than the corresponding statements. The difference is due to an overall shortening of the question as compared to the statement, not to shortening of any particular part of the utterance. This shortening is not found in all languages, for example not in German (Bannert 1983). If it is language specific, this type of overall shortening will have to be part of the rules for question formation in Hausa.

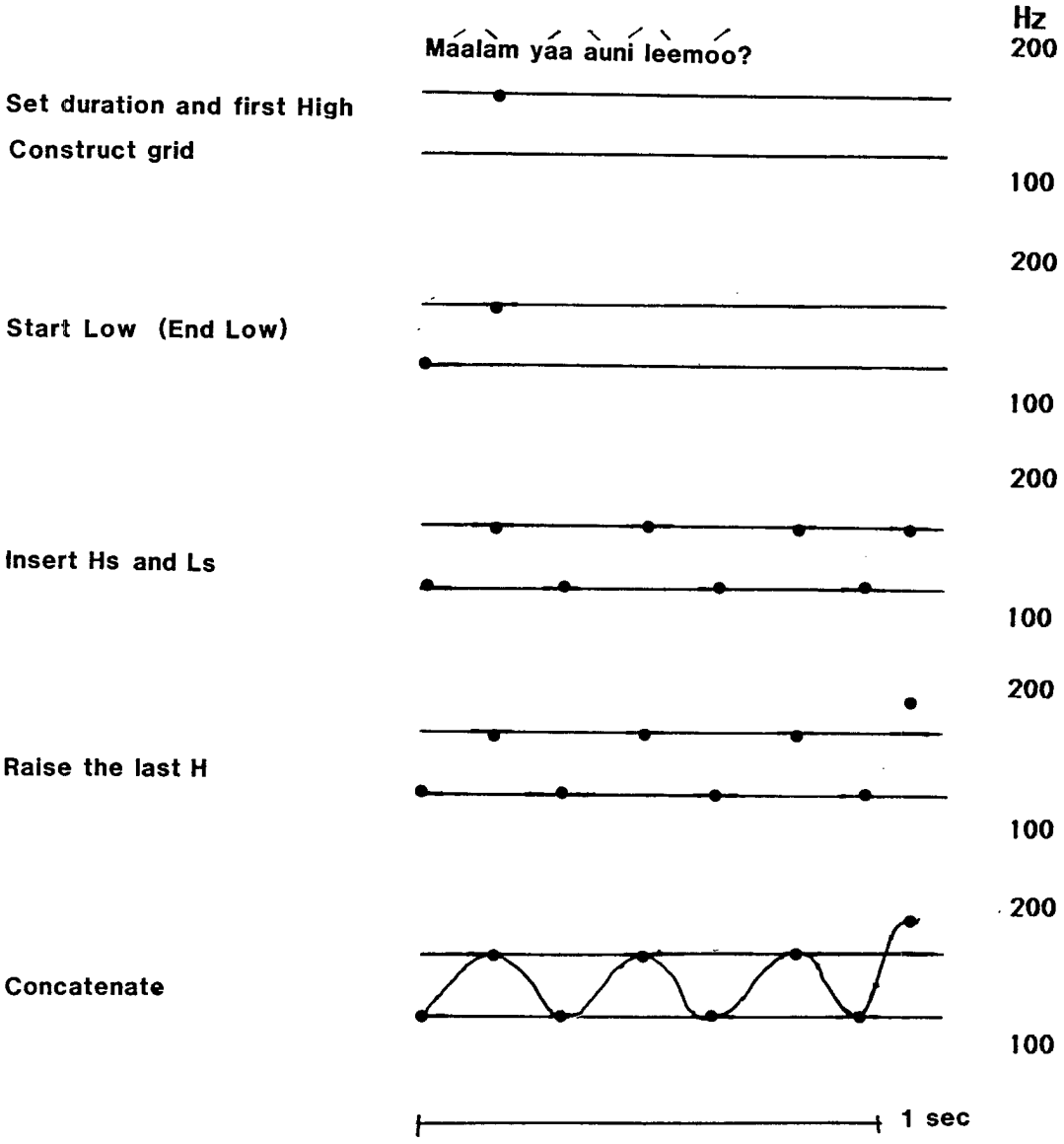


Figure 7. Rules for generating the pitch curve associated with a yes/no question on alternating High and Low tones.

Figure 7 shows how a yes/no question "Maalam yaa auni leemoo?" can be generated. 1) Set the duration of the question at 10% less than the corresponding statement (here = 1170 msecs.). 2) Construct the grid with zero slope, anchored on the first High as for statements. 3) Insert a Low starting point, and optionally, a Low endpoint. 4) Insert the tones on the grid lines. 5) Raise the last High tone. The amount of increase of the last High varies considerably, but it is typically around 5-10%. 6) Lastly, concatenate the turning points into a smooth curve.

Typical pitch patterns of question word questions are illustrated in figure 8. They display characteristics that place them as something between the pitch patterns in statements and those of yes/no questions. Like statements, a question-word question with alternating High and Low tones will have a downward

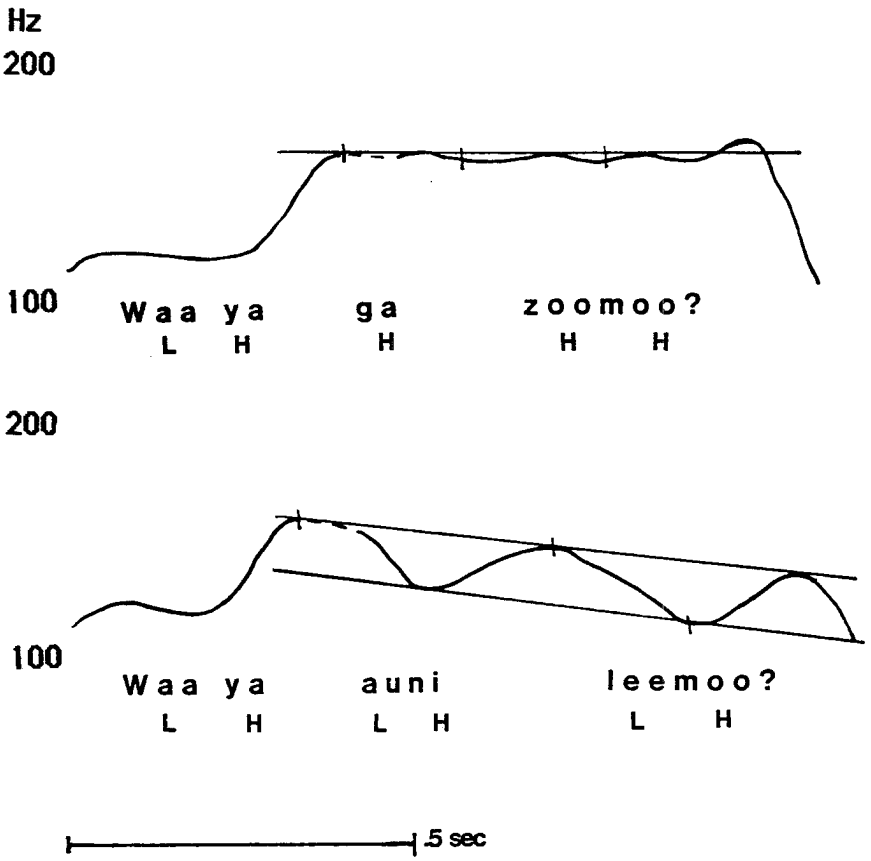


Figure 8. Typical pitch patterns of question-word questions, beginning with "waa?", 'who?'.

slope, but this slope is significantly less than that in statements. The mean slope of "Wàa ya àuni lèemoo?" (Who weighed the oranges?) is 23% pr second (SD=3.1, n=8). But "wàa"-questions of the basically High tone sentences have a zero slope, just like the yes/no questions. Also, like the yes/no questions the last High tone may be raised, but this only happens for two out of the nine speakers. Mostly the last High tone is not raised. In this type of question, there is, however, always a fall after the final High tone. This fall is most simply accounted for by assuming it to be part of the general convention of starting and ending Low.

CONCLUSION

The model described here provides a sufficient basis for generating the sample sentences. The intonation of simple sentences in Hausa can be represented as grids of (near)parallel lines. The rate of slope of the grid is related at least to sentence type (statements and question), sentence length, and tone pattern. Fundamental frequency curves can be generated by rules that refer to the underlying grids. The base form of a grid is a topline from which the bottom line is derived as a ratio. The topline is anchored on the first High tone. The slope of the base intonational grid is that which appears from sentences on High tones only, when there is no influence from tonal assimilations. All questions are specified with zero slope, and statements with a slope that depends on the length of the sentence. In this study only two slopes appear as base toplines, one for short, and one for long sentences, but obviously a function can be worked out

from more data on different sentence lengths. Lexical tones and boundary marks are mapped onto the gridlines. In sentences with alternating High and Low tones assimilatory processes apply that lower Highs following Lows, so that the slope of the grid becomes steeper, thus claiming that downdrift is an effect of both intonation and tonal assimilations. These assimilations apply in statements and question-word questions, but not usually in yes/no questions. Lastly, a smooth curve is generated through the boundaries and tones. At least for the simple sentences dealt with above, the resulting pitch curves approximate the data well.

Acknowledgments: Many thanks to Will Leben, Nicolas Faraclas, and Brian McHugh for recording the speech material in Kano. Eva Gårding provided much inspiration. I am also much indebted to Peter Ladefoged, Ian Maddieson, and the UCLA Phonetics Lab for comments on earlier versions, and to Norma Antoñazonas for the programming. Financial support came from the Swedish Council for Research in the Humanities and Social Sciences, and from the National Science Foundation.

References.

Abraham, R. (1941). Modern Grammar of Spoken Hausa. (Government Printer, Kaduna).

Bannert, R. (1983). "Some phonetic characteristics of a model for German prosody," Working Papers 25 (Department of Linguistics, University of Lund), pp. 1-34.

Bredvad-Jensen, A.-C. (1983). "Perception studies of Swedish interrogative intonation," Paper presented at the 10th International Congress of Phonetic Sciences (Utrecht).

Bruce, G. (1982). "Experiments with the Swedish intonational model," in Preprints of papers for the Working Group on intonation, edited by H. Fujisaki and E. Gårding.(The XIIth Internatinal Congress of Linguists, Tokyo), pp. 36-45.

Cowan, R. and Schuh, R. (1976). Spoken Hausa (Language Services, Inc., Ithaca).

Gårding, E. (1979). "Sentence intonation in Swedish," Phonetica 36, 207-215.

Gårding, E. (1981). "Contrastive prosody: a model and its application," Studia Linguistica 35, 146-165.

Gårding, E. and Bruce, G. (1981). "A presentation of the Lund model of intonation," in Nordic Prosody II, edited by T.Fretheim (Tapir, Trondheim), pp. 33-39.

Hadding-Koch, K. (1961). Acoustico-phonetic studies in the intonation of Southern Swedish. (Gleerups, Lund).

Hombert, J.-M. (1974). "Universals of downdrift: their phonetic basis and significance for a theory of tone," Studies in African Linguistics, Supplement 5, edited by W. Leben, pp. 169-183.

Hyman, L. and R.Schuh (1974). "Universals of tone rules: evidence from West Africa," Linguistic Inquiry 5, 81-115.

Ladefoged, P. (1975). A course in phonetics (Harcourt, Brace, Jovanovich, New York).

Meyers, L. (1976). "Aspects of Hausa tone," Working Papers in Phonetics 32 (University of California, Los Angeles).

Miller, J. and P. Tench (1980). "Aspects of Hausa intonation: utterances in isolation," Journal of the International Phonetic Association 10 (1-2), 45-63.

Newman, P. and R. Ma Newman (1981). "The question morpheme q in Hausa," Afrika und Ubersee 64 (1), 45-63.

Pierrehumbert, J. (1980). The phonology and phonetics of English intonation. Ph.D. dissertation (MIT).

Cross-linguistic differences in diphthongs.

Mona Lindau, Kjell Norlin and Jan-Olof Svantesson

In this paper I will discuss some phonetic differences that are found in diphthongs. The kinds of questions that will be addressed have to do with the specification of timing in speech. Are the vowel elements of diphthongs joined in language-specific ways, or can they be joined by more general principles, such as the assumption of a fairly constant transition duration, so that the transition will be faster if the distance between onset and offset vowels is greater? This kind of principle was proposed by Kent and Moll (1972) for some types of vowel-to-vowel sequences. Or does the duration of the transition lengthen as the distance of the transition increases? If such principles are at work, then their effects would show up acoustically as a tendency towards similar transition durations in different languages for the "same" diphthongs.

Two diphthongs were selected for this study: /ai/ and /au/. These are by far the most frequent diphthongal phonemes in the languages of the world. A survey of the sources of Maddieson's (1984) phonological survey of over 300 languages reveals that diphthongs occur in about a third of the world's languages. In this survey a diphthong was defined as a vowel sequence within a syllable. Diphthongs of the /ai/-type occurs in 75 % of these languages, and an /au/-type in 65% of these languages.

Phonetic properties of /ai/ and /au/ were studied from four languages with different systems of vowels and diphthongs: Arabic, Hausa, Chinese, and English. Figure 1 below shows vowel systems and diphthongs in Arabic, Hausa, and Chinese.

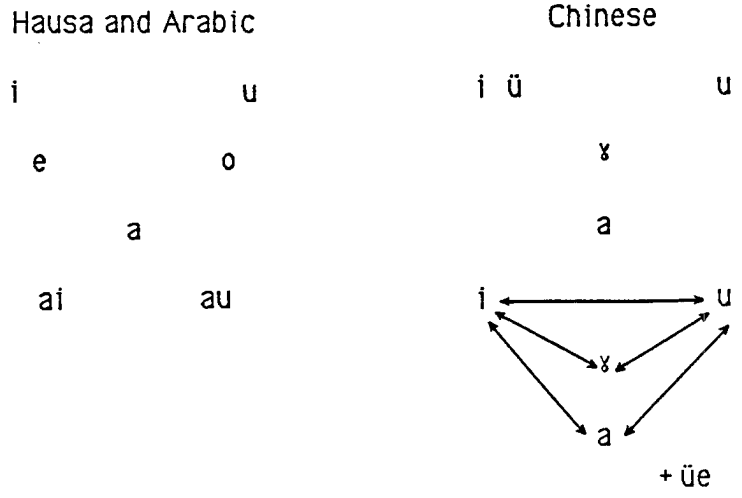


Figure 1. Vowels and diphthongs of Hausa and Arabic, and Chinese.

Arabic and Hausa both have small simple systems with five vowels distributed along the periphery of the vowel space. These languages have been described as having two diphthongs only, /ai/ and /au/. Chinese can also be described with a five vowel system, but of a different shape than in Arabic and Hausa. The 11 Chinese diphthongs can be summarized as on the figure. The

diphthongal system has many mirror image type diphthongs with the vowel elements involved in two directions. English represents another language with several diphthongs, including /ai/ and /au/.

The data consist of tape recordings from 10 speakers of Kano Hausa, 6 speakers of Cairo Arabic, and 4 speakers of Peking Chinese, saying words illustrating the diphthongs in carrier sentences. Diphthongs in Hausa and Chinese were all spoken on high tone. The rate of speech was moderate in all cases, but it was not strictly controlled. Formant frequencies and durations were measured from wide band spectrograms. Figure 2 shows spectrograms of /ai/ diphthongs in Arabic and Chinese. The formant frequencies of the onset vowel and the offset vowel were taken as midfrequencies of steady state portions. The duration of the whole diphthong was measured, as well as the duration of the second formant transition. The F2 transition was taken to start and end at points of intersection between the transition and the steady states, as illustrated in the figure. To normalize for some variation in speech rate, the F2 transition duration was calculated as a ratio between the transition duration and the total duration of each diphthong, and expressed as the percentage of transition duration of the whole diphthong. The American English data were taken from the results of the five speakers used in Thomas Gay's classical paper on English diphthongs (1968).

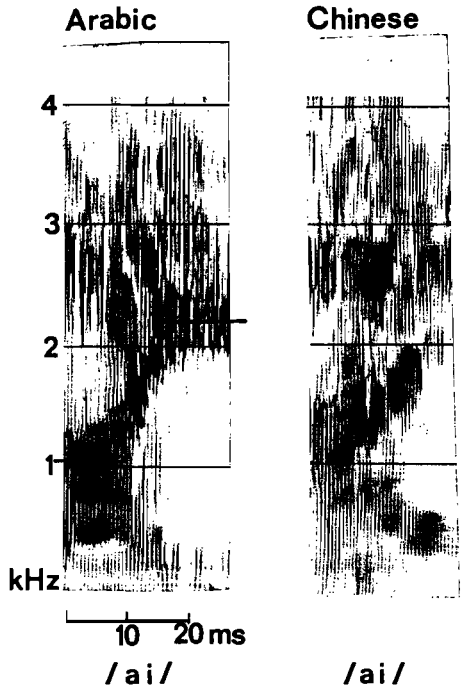


Figure 2. Spectrograms of /ai/ in Arabic and Chinese.

The spectrogram shows that the transition between onset vowel and offset vowel behaves differently in the two languages. The second formant transition of /ai/ is slower in Chinese than in Arabic, and it takes up a larger percentage of the diphthong.

The transition duration percentages of the total duration in the different languages are displayed as bargraphs in figure 3. The transition duration is very short in Arabic for both diphthongs, and equally short for Hausa /au/. For the diphthongs in these languages the transition occupies only 16-20 % of the whole diphthong. In Chinese the transition occupies 40-50 % of the diphthong. In English the transition tends to dominate the diphthong by taking up as much as 73 % of the /au/ and 60% of the /ai/. These transition duration percentages are significantly different between the three groups of languages, Hausa/Arabic, and Chinese, and English. It looks very much as if the transition duration of diphthongs is a language-specific trait.

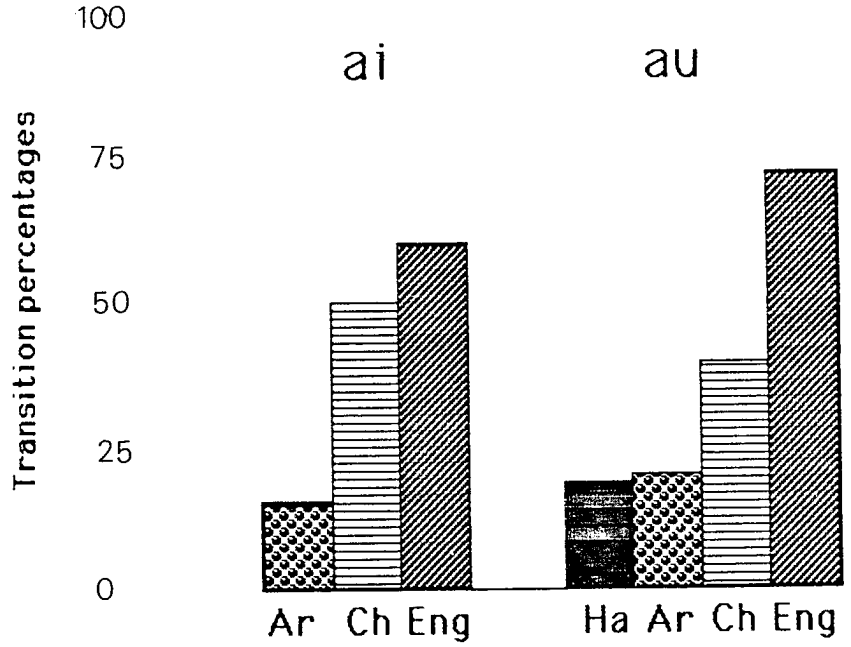


Figure 3. Bargraphs of the transition duration percentages of /ai/ and /au/ in Hausa, Arabic, Chinese, and English. There is no Hausa /ai/.

But the phonetic qualities of the vowel segments involved in the diphthongs also differ to some extent between languages. Thus there is the possibility that the transition duration could be an effect of the distance between onset vowel and offset vowel. Possibly the transition takes longer if it has to travel farther. To test this thought the distances between the onset vowel and the offset vowels for each diphthong in each of the four languages were calculated as the Euclidian distance in the F1/F2 space measured in mel. In figure 4 the mean acoustic distances were plotted against the transition duration percentages. There is no obvious general relationship between the duration of the transition and the acoustic distance it has to travel. However, when the diphthong /ai/ is considered on its own, there is a good correlation between the duration percentage of the transition and the acoustic distance it has to travel. The figure represents mean values. The correlation coefficient for 18 speakers of the three languages was 0.87. This no doubt reflects the tongue body displacement and its rate of movement acting in accordance with a principle such as "the further to go, the longer it takes". But in the case of the /au/ diphthong there is no such correlation. This is probably due to the more complex and language-specific relationships that exist between the tongue movements and the lip movements in moving from /a/ to a rounded /u/.

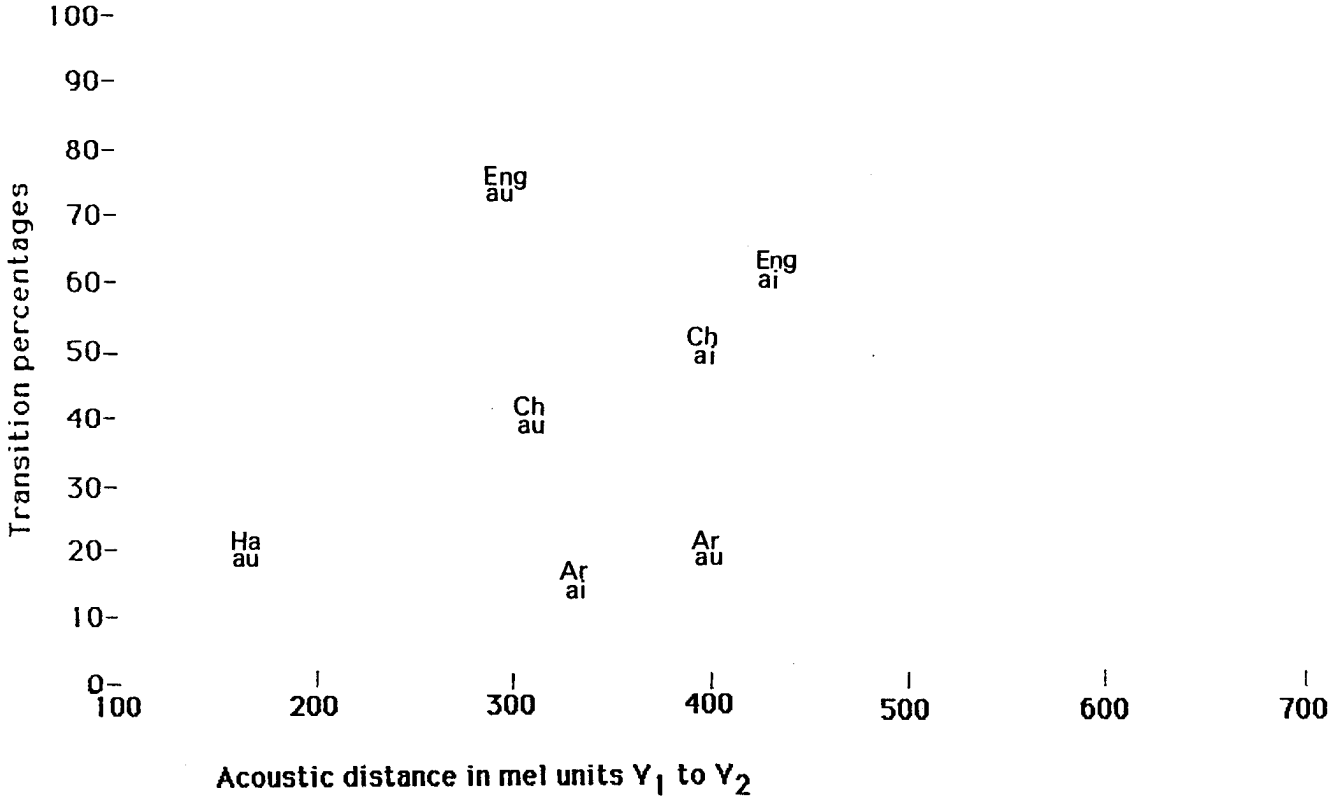


Figure 4. Mean acoustic distance plotted against mean transition duration percentage for /ai/ and /au/ in Hausa, Arabic, Chinese, and English.

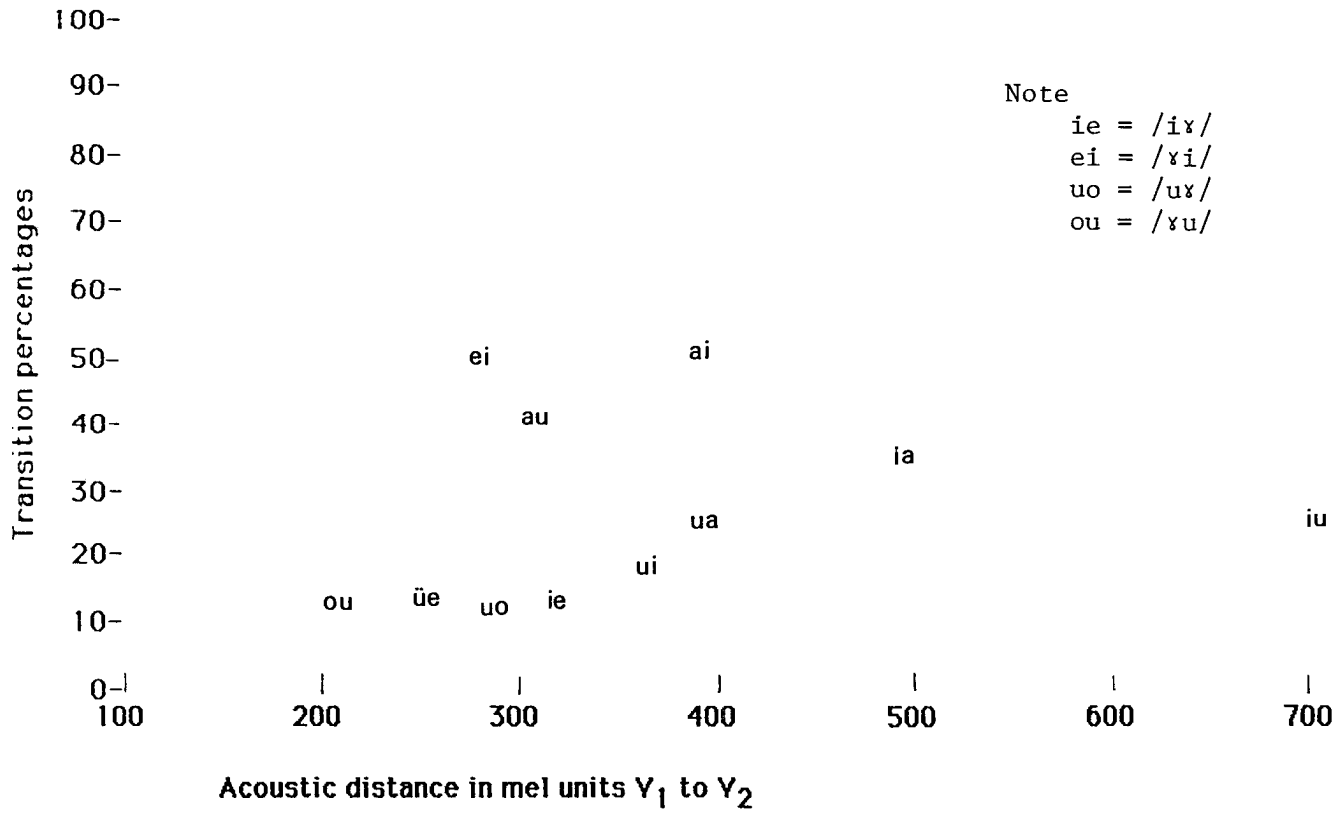


Figure 5. Mean acoustic distance plotted against mean transition duration percentage for all the diphthongs of Chinese (4 speakers).

Not even within a language are the transition durations constant, or always related to the distance it has to travel. In figure 5 the transition duration percentages of all 11 diphthongs in Chinese are plotted against the acoustic distances between the onset vowel and the offset vowel. The diphthongs represent mean values of the four Chinese speakers. The correlation for all 11 diphthongs is nonsignificant. But when only those four diphthongs are considered that move upwards in the usual vowel figure, then the correlation for these diphthongs is good ($r= 0.7$, $n= 14$). These upwards moving diphthongs also behave in accordance with the principle "the further to go, the longer it takes". These results indicate that the transition durations may be not only language-specific, but sometimes also specific for certain types of diphthongs.

It is possible that the relative amount of transition in diphthongs has something to do with the system of vowels and diphthongs. Hausa and Arabic with their very short transitions both have few vowels and few diphthongs, while Chinese and English with their more dominant diphthongal transitions represent languages with quite a lot of diphthongs. But two languages of each type is really not enough for drawing any inferences. We conclude that the timing of the diphthongal transition is not constant for the "same" diphthong in different languages, and that the timing relationships can be predicted from a general principle of duration varying with the distance for some diphthongs, but not all. Thus the transitional rate and duration may be language-specific, and possibly even diphthong-specific as well. The diphthongal timing properties must be specified as part of the phonetic description of diphthongs in different languages.

References

Gay, T. (1968) "Effect of speaking rate on diphthong formant movements." Journal of the Acoustical Society of America", vol.44, no.6:1570-1575.

Kent, R.D., and K.L. Moll (1972) "Tongue body articulation during vowel and diphthong gestures." Folia Phoniatrica. vol. 24:278-300.

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

Acknowledgments: We are grateful for constructive suggestions and discussions with Pat Keating, Peter Ladefoged, Ian Maddieson, and Ren Hong-Mo.

Mona Lindau-Webb is at the Department of Linguistics, University of California, Los Angeles, Los Angeles, Calif. 90034. Kjell Norlin and Jan-Olof Svantesson are at the Phonetics Laboratory, University of Lund, Helgonabacken 12, 223 62 Lund, Sweden.

Voice Discrimination by Two Listener Populations

Jody Kreiman and George Papçun*

[Paper presented at the 109th Meeting of the Acoustical Society of America, April 8-12, 1985, Austin, Texas]

Over the past 2 years we have been studying the cognitive and perceptual processes underlying voice recognition by humans. This report focuses on one aspect of that research: the effect of the age of the listener on ability to discriminate voices. The voices used in this study were those of 10 young male Southern Californians, selected from an original set of 22 who were recorded making telephone survey calls. In order to select the 10 voices to be used, we had a group of listeners rate how easy or hard they felt the voices would be to remember, on a seven point scale. The voices whose ratings as a group most nearly conformed to a normal distribution were selected. Thus we are assured that, in the opinions of the raters at least, some of the voices are easy to remember, some are hard to remember, and some are of average memorability. The final 10 speakers used in the study ranged from 20 - 31 years in age; all were free from vocal pathology.

Speakers made several telephone survey calls at each of two recording sessions separated by at least one week. To avoid recording over telephone lines, a high-quality dynamic microphone was attached to the mouthpiece of the telephone. This made it possible to obtain good-quality recordings while allowing the speaker to carry on a normal telephone conversation; additionally, only the speaker's voice (and not that of the interviewee) was recorded. One call was selected from each recording session for use in our studies. Three sentences were excerpted from each call for each speaker, and a stimulus tape was constructed using both orders of all possible pairs of the 10 voices, plus an additional 30 pairs where the speakers were the same. Speakers within a pair always said the same thing; when speakers were the same, one sentence was taken from each of the two survey calls, so listeners never compared two identical tokens. The design of the stimulus tape was as follows. It contained 90 paired comparison trials in which the voices were different (noise trials) and 30 in which voices were the same (signal trials), making a total of 120 trials per subject.

Forty-four listeners participated in the experiment. They ranged from 21-74 years of age, with a mean age of 43 and a standard deviation of 16.25. They were asked to decide, for each pair of voices, whether the voices were the same or different, and to rate their confidence in their response on a five-point scale. The stimulus tape was played over a loudspeaker; listeners were tested individually or in small groups. Listening sessions lasted about 40 minutes.

For each listener we calculated the percentage of correct "voices same" responses--the hit rate--and the percentage of incorrect "voices same" responses--the false alarm rate. ROC curves were also constructed from each subject's confidence ratings; and the area under each curve, A_z, was calculated using a program by Dorfman and Alf (reprinted in Swets & Pickett, 1982). A_z is a non-parametric measure of discrimination ability which is relatively unbiased by listener response criteria. Mean values for hit rates, false alarm rates, and A_z are given in Table 1; chance equals 0.5 for all three measures.

Table 1. Accuracy measures averaged over all listeners.

	Mean	Range
A_z	0.867	0.54 - 0.97
Hits	0.857	0.57 - 1.00
False Alarms	0.236	0.01 - 0.59

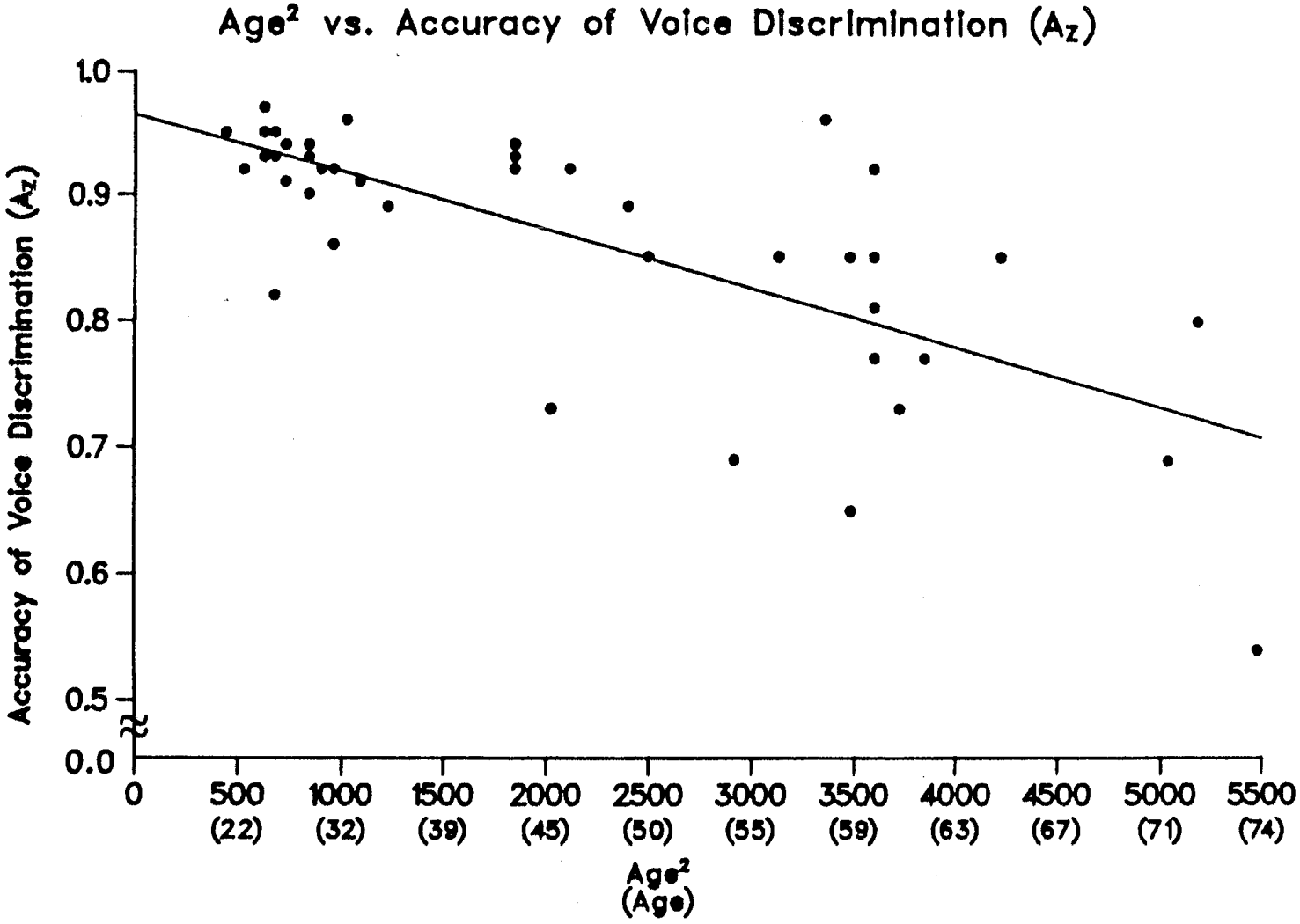
On the average, listeners performed quite well on these measures, although scores range from chance to nearly perfect for all three. However, performance on all measures decreased significantly with listeners' ages. Separate linear regressions were calculated using Age and Age^2 to predict hit rates, false alarm rates, and A_z values; in all cases, Age^2 provided better prediction than did Age.

Figure 1 shows the results of the regression of A_z on Age^2. Age and Age^2 are shown on the x-axis, and discrimination scores are on the y-axis. The regression is highly significant; in addition, $r^2 = 0.53$, indicating that discrimination performance is very well predicted by the square of a listener's age. The older the listener, the worse his or her discrimination performance; and worsening accelerates with aging.

These decrements in performance are reflected in both hit rates and false alarm rates. Figure 2 shows the regression of hit rates on the square of Age. Age and Age^2 are again on the x-axis; the hit rate is given on the y-axis. Given the fact that hit rates and false alarm rates both fluctuate considerably with listeners' response criteria, the increased scatter in the data is not surprising. Despite this extra source of variability, the regression is highly significant: older listeners make reliably fewer correct "same" judgments than do younger listeners. Figure 3 shows the result for false alarm rates; as before, Age^2 is on the x-axis, and the false alarm rate is on the y-axis. Again, despite the added effects of response bias, older listeners made significantly more incorrect "same" judgments than did younger listeners.

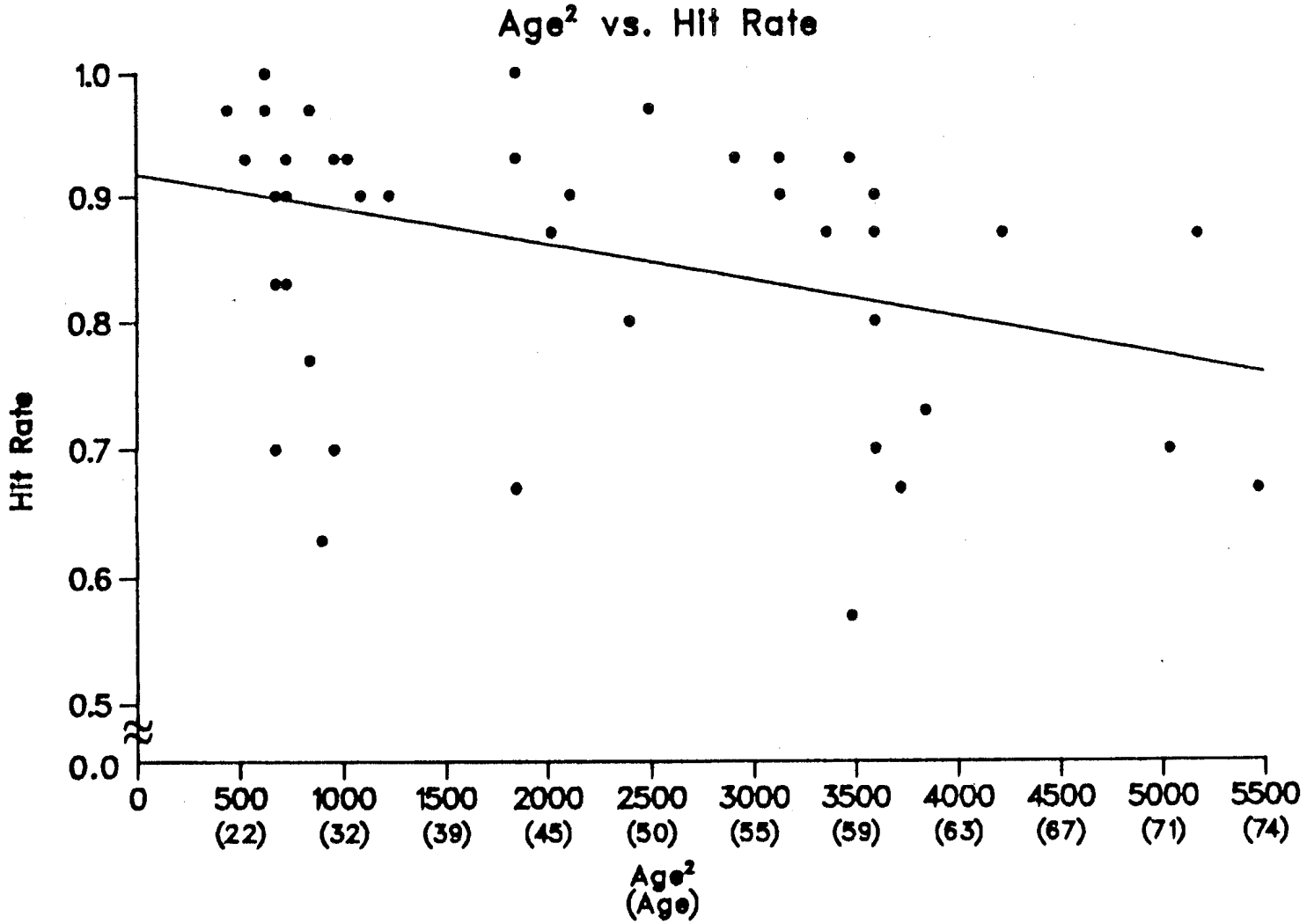
Previous studies have found decreases in visual acuity and in ability to identify odors with age (Corso, 1971; Doty et al., 1984); and a number of investigators have reported age-related drops in speech discrimination scores (e.g., Marston & Goetzinger, 1972; Bergman, 1971; Price & Simon, 1984). It is possible that decreased ability to discriminate among voices is related to a more general decline in sensory and cognitive functions which accompanies aging. An alternative explanation is that older listeners performed poorly because they lack experience with young voices, and thus are less able to distinguish within- from between-speaker variations in voice quality. An informal examination of the data provides some support for this view: older listeners who come into contact with younger speakers at work, or who have college-aged children living at home, did generally perform better on the task than did listeners who know few or no young speakers. These two explanations are of course not mutually exclusive; and an interaction between age- and experience-related factors may well occur.

Studies of aging and speech discrimination have claimed that difficulties in recognizing speech are related to problems with processing fine temporal detail--for example, differences in closure durations. Temporal detail has been implicated in the voice recognition process, in particular in spectrographic studies of voice quality (e.g., Stevens et al., 1968); however, machine recognizers and human listeners also use long-term features of the voice (e.g., La Riviere, 1972; Doherty, 1976; Abberton & Fourcin, 1978). It remains to be determined if losses in voice discrimination ability are similar in nature to those for speech discrimination, or if they stem from different causes.



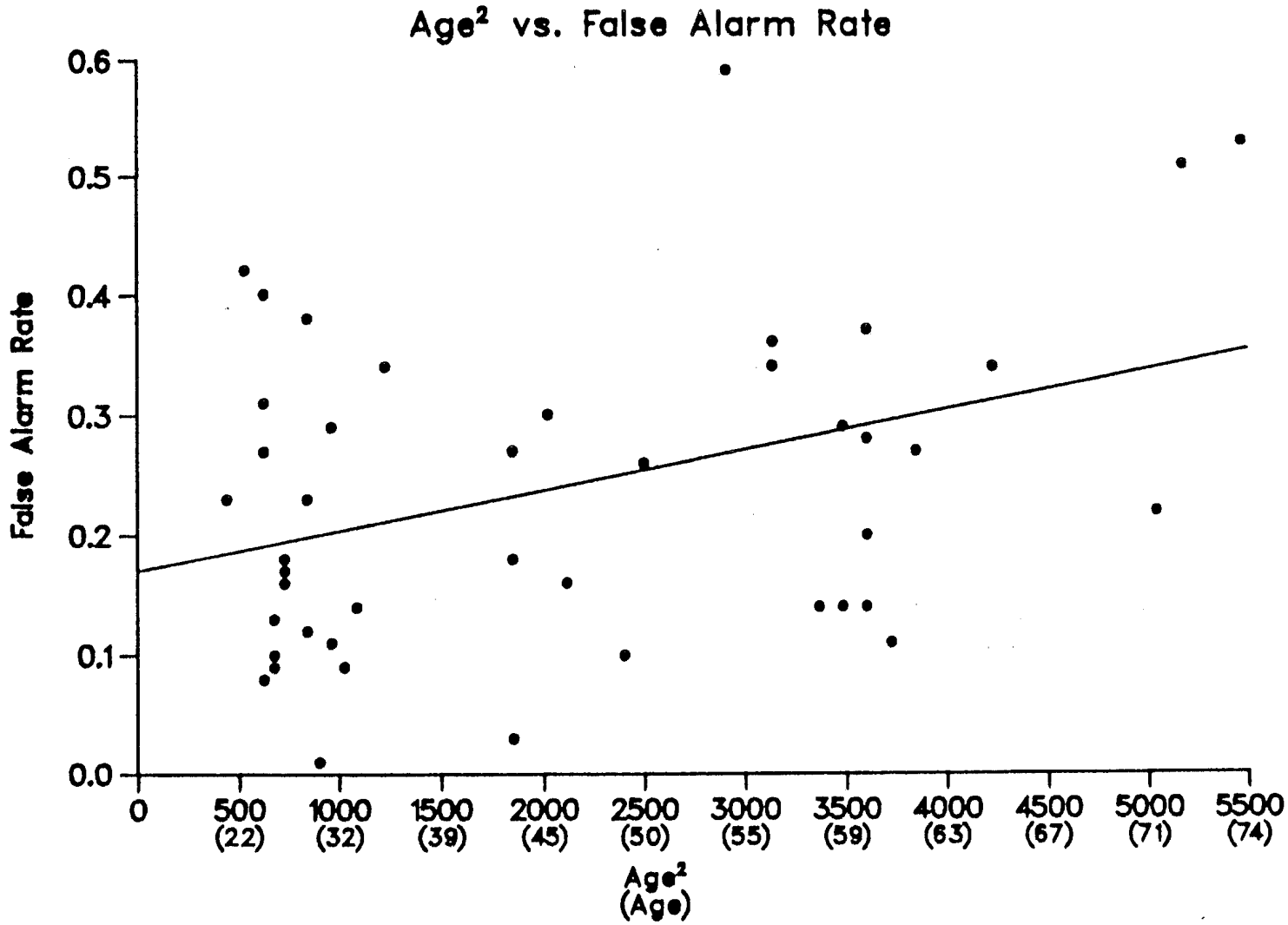
Regression equation: $A_z = 0.471e{-}4\ Age^2 + 0.966$
$F(1,42) = 76.326$ $(P<0.001)$ Multiple $r^2 = 0.525$

FIGURE 1



Regression equation: Hit Rate = 0.301e−4 Age² + 0.920
F(1,42) = 7.446 (P<0.01) Multiple r^2 = 0.15

FIGURE 2



Regression equation: False Alarm Rate=.334e−4 Age² + 0.166
F(1,42) = 6.673 (P<0.02) Multiple r^2 = 0.14

FIGURE 3

*George Papçun's address is: Los Alamos National Laboratory, Computer User Services C-10, Mail Stop B-296, Los Alamos, NM 87545.

References

Abberton, E., and Fourcin, A.J. (1978). "Intonation and speaker identification," Language and Speech 21, 305-318.

Bergman, M. (1971). "Hearing and aging," Audiology 10, 164-171.

Corso, J.F. (1971). "Sensory processes and age effects in normal adults," J. Gerontol. 26, 90-105.

Doherty, E.T. (1976). "An evaluation of selected acoustic parameters for use in speaker identification," J. Phonetics 4, 321-326.

Doty, R.L., Shaman, P., Applebaum, S.L., Giberson, R., Siksorski, L., and Rosenberg, L. (1984). "Smell identification ability: changes with age," Science 226, 1441-1443.

La Riviere, C. (1972). "Some acoustic and perceptual correlates of speaker identification," Proceedings of the 7th International Congress of Phonetic Sciences, 558-564.

Marston, L.E., and Goetzinger, C.P. (1972). "A comparison of sensitized words and sentences for distinguishing nonperipheral auditory changes as a function of aging," Cortex 8, 213-223.

Price, P.J., and Simon, H.J. (1984). "Perception of temporal differences in speech by 'normal-hearing' adults: Effects of age and intensity," J. Acoust. Soc. Am. 76, 405-410.

Stevens, K.N., Williams, C.E., Carbonell, J.R., and Woods, B. (1968). "Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material," J. Acoust. Soc. Am. 44, 1596-1607.

Swets, J.A., and Pickett, R.M. (1982). Evaluation of Diagnostic Systems: Methods From Signal Detection Theory (New York: Academic Press).

Borrowed sounds

Ian Maddieson

1. Introduction and Preliminaries.

When one language is affected by another, one possible result is an addition to the phonemic inventory; new contrastive sounds may enter the affected language via the process of lexical borrowing. Of course, adaptation of loanwords to existing phonemic and phonotactic patterns is more common (Haugen 1950, Deroy 1956), and several linguists have suggested that this process provides a tool for investigating the nature of phonological representations (e.g. Hyman 1970, Kaye and Nykiel 1981). However, it is not unusual for borrowing to bring about some phonological realignment in the recipient language. Such a realignment might be one which affects only phonotactics, for example, introducing previously impermissible clusters (like /ʃm/ in English, introduced from Yiddish). More rarely, it is one which results in an extension to the phoneme inventory by the addition of previously unused phones or by the promotion to phonemic status of an allophonic variant of an existing phoneme. Language contact, including phonemic borrowing, has been discussed as an important factor in limiting the variability of phoneme inventory size (Maddieson 1984). Phonemic borrowing has also been put forward as the explanation for a suggested correlation between paucity of speakers and size of phoneme inventory (Haudricourt 1961). The question of how such borrowing proceeds is therefore of relevance to some quite general questions about language structure.

Linguists seem to hold quite opposed views about which sounds are the ones that are likely to be borrowed. Some predict that the sounds that are the most distant from the phonetic resources of the recipient language are the ones most likely to be added to the inventory as a result of contact and borrowing. According to this view, exotic segments are only likely to be borrowed when they are rather markedly different from the sounds of the recipient language; otherwise, loanwords are adapted in shape so that the nearest matching existing segments in the language are used. When segment borrowing occurs, the motive might well be the prestige that is associated with the exotic pronunciation. Often cited as the prototype for this view is the borrowing of the voiced uvular trill /R/ from French into German (e.g. Bach 1965). However, it is unclear if this is in fact a case where adoption of loanwords led the change (followed by innovation of the newly fashionable pronunciation for the existing German /r/ phoneme), or whether a uvular /R/ may have existed as a dialectal variant in German prior to the strong 17th century influence of French (see Brunt 1983 for a succinct summary of the issues). In discussion of a less ambiguous case, Deroy (1956) suggests a prestige motivation for the borrowing of /h/ and /w/ from Germanic into the Vulgar Latin of France.

A different, but somewhat related hypothesis, has been put forward in a brief paper by David Cline (ms, 1984). Cline suggests that those segments which are more marked on a universal basis (i.e. less frequently found in the world's languages) are the ones more likely to be borrowed. What is common to these views is the opinion that borrowing is not influenced by the phonological structure of the recipient language, apparently also the view of Hagège and Haudricourt (1978).

On the other side are those linguists who believe that the segments most likely to be borrowed are those which are easily fitted into existing oppositions. For example, a language with voiceless stops /p, t, k/ and voiced stops /b, d/ would be expected to easily borrow a voiced velar plosive /g/, since it already has a stop at the velar place of articulation and a voicing contrast in stops at other places. Jakobson (1931 [1964]) expressly proposed as a general principle of historical phonology that "les phonèmes étrangers que la langue s'approprie le plus aisément sont ceux qui s'incorporent dans les corrélations déjà existantes" (p. 323). He exemplified this principle with the case of the voiceless labio-dental fricatives in Russian. Given pre-existing voiceless fricatives /s, s^j, x/ etc, and voiced fricatives /v, v^j, z, z^j/ etc, later nativizations of loanwords containing an original /f/ introduce voiceless /f/ and /f^j/. (Which of these occurs in a given word depends on the vowel context in the donor language.) Obviously, no new features are introduced by the addition of these phonemes to Russian.

Along these lines, Maddieson (1984), assuming that a multi-lingual social setting is the normal situation for human languages, argues that an important factor in maintaining the typical phoneme inventory size of 30 or so segments (plus or minus, say, 5) is the limitation on inventory expansion imposed by the requirement that added segments must be capable of being "generated" by combinations of familiar features already in the recipient language. This process is seen as more explanatory than positing constraints on efficient communication as being repsonsible for the upper and lower limits on the size of a phhoneme inventory.

In order to decide how strongly the phonological structure of the recipient language affects which segments are borrowed, all of the consonant segments marked in the UPSID inventory as loan phonemes were examined in relation to the pattern of contrasts which previously existed in the languages. UPSID (the UCLA Phonological Segment Inventory Database) is a computer file containing the phonemes of 317 languages chosen to represent the genetic diversity of the world's extant languages (Maddieson 1984). It was inspired by the work of Ferguson, Greenberg and their colleagues on the Stanford Phonology Archive. In compiling UPSID, those segments in a language which only have phonemic status as a result of relatively recent borrowing were indicated by a special variable value. For the present paper, every language which has any consonantal segment marked with this value was examined. As is inevitable, some errors in the data file were detected during this process. These were corrected. Also, a few additional loan segments were identified - often those of an older layer of borrowing - and a few cases from other languages outside the UPSID data base were also reviewed. One language, Zulu, was replaced by a close sister, Xhosa, for which more complete documentation on loan forms is available. The result is a set of 184 instances of borrowed segments to examine. Data are drawn from the references in Maddieson (1984) except where a reference is included here. Only consonants were examined because the greater number of parameters involved in consonants makes it easier to devise an intuitive metric for determining the relative distances of loan segments from the inventories they are entering.

For this survey we will rely on a traditional phonetic framework of classification which divides consonant segments into series (voiceless plosives, voiced nasals, voiced fricatives, etc.) which have representatives at different places. Degree of similarity between segments will be assessed mainly by whether they share major manner and place features. Here "features" does not refer to phonological melody features in the technical sense (distinctive features) but

rather to possession of articulatory or auditory phonetic properties. The framework is similar to that used for classifying segments in the UPSID database and described in Chapter 10 of Maddieson (1984). However, some phonetic distinctions will be disregarded. A labio-dental fricative will be considered to be matched in place by a bilabial stop, since bilabial fricatives are relatively rare (and labio-dental stops are impossible for many speakers and are not known to occur distinctively in any language). Palato-alveolar fricatives and affricates are considered matched in place by palatal segments of other types, such as nasals or approximants. Dental and alveolar places are not distinguished. Fricative and affricate series are not divided into sibilant and nonsibilant. Both /h/ and /ʔ/ are treated as sui generis, /h/ being a voiceless vowel with consonantal function but a variable place of articulation, and /ʔ/ a segment to which (most of) the usual series-generating contrasts cannot apply.

On reviewing this data, it appears that the evidence is heavily in favor of the view that segments are most likely to be borrowed when there are already appropriate segments to promote the adoption of the new segment. In order to show that this result is not simply a reflection of the general frequency of segments, we will compare this data with a simulation. We have not examined the question of what social conditions favor segment borrowing, nor the issue of foreign influence on the direction of internal change in a language (e.g. as discussed by Petrovici 1957), and hence have no comments to make on these issues.

2. Classification of borrowed sounds

We will classify the cases we have found in our data into groups ranging from those in which the loan segment seems to be very close to the pre-existing resources of the languages, to those in which it is remote. Cases are sorted into six classes. Note that the form of the loan segment being compared is that which it has in the recipient language, not that of the donor language. We will return to this point briefly later. An appendix is attached in which the segments are listed, together with the segments which are considered to be the closest promotors of the borrowing in the given case.

(a) Close segments

Many loan segments "fill a gap" in the inventory, that is, they add to an existing series at a place of articulation where an existing minimally different segment in another series already occurs. Such cases form class 1. A representative case is the Togo Remnant language Lelemi. This language had a stop inventory consisting of:

	t		k
b	d		g

Loanwords from European colonial languages have added a new voiceless stop segment /p/. This does not create a new segment type, since voiceless plosives already exist in the language; moreover a segment which is minimally different in voicing from /p/ and shares its place of articulation is present, namely, /b/. A similar but in some ways more interesting case is found in Yoruba, which had the stop inventory:

	t	k	kp
b	d	g	gb

Earlier loans containing /p/ were borrowed with /kp/ or /b/ substituted, e.g. /kpotogí/ "Portuguese", /kpɔ̀ũ/ "pound", /kɔ́bɔ̀/ "penny" (< "copper"). More recent loans, mostly from English (now a much more widely spoken language in the

community) retain /p/, e.g. /pílò/ "pillow", /pidʒɔt/ "Peugeot". These more recent loans fill the gap in the voiceless plosive series at the bilabial position. An example involving fricatives is the borrowing of /ʒ/, mostly through loans from French, into German. German previously had three sibilant fricatives:

s S
z

The previously existing voiceless palato-alveolar fricative acquired a voiced partner, parallel to the voiced alveolar fricative /z/, following the adoption of loan forms such as /ʒeni/ "genius" (< French génie). Not all such cases involve simply a difference in voicing from an existing segment. Mazatec, for example, has borrowed /p/ from Spanish. Although the language previously had no bilabial stops, it had a bilabial nasal /m/ and fricative /β/, and voiceless stops /t/ and /k/. Tseshaht (Nootka) has borrowed uvular ejective stops /q'/ and /q^{w}'/ from Makah and/or Nitinat (Jacobsen 1969). Previously the language had uvular plosives and velar ejectives, including labialized variants. Thus both place and manner are represented in existing phonemic segments. Approximately half of the loan phoneme cases surveyed consist of gap-filling moves of this kind (90/184 cases). In the appendix these cases are marked with the number 1.

In class 1 cases, the newly added segment differs only minimally from two (or more) of the existing segments in the language, one with respect to place and one with respect to manner. We have placed in class 2 those cases where the borrowed segment is minimally different from (at least) two existing segments along two different dimensions of contrast, but where the new segment creates a new manner series. A relatively common type of example in class 2 is a language which has borrowed voiced stops when it had pre-existing voiceless stops and voiced nasals. An example is Tarascan, which had:

p t k
m n ng

This language has borrowed /b, d, g/ from Spanish. Other types of borrowings in this class include voiced fricatives taken into languages which already had voiceless fricatives and voiced stops at the same place, such as /z/ taken into Hindi which already had /s/ and /d/, as well as affricates taken into languages where none previously existed, but which had the separate "components" available. An example of the latter is Bashkir, which has /t/, /s/ and /ʃ/ but no original sibilant affricates. Loanwords from Russian have introduced /ts/ and /tʃ/, e.g. in /tsikl/ "cycle" (< Russian /tsikl/), /tʃan/ "tub, vat" (< Russian /tʃan/). There are 19 examples of class 2. In classes 1 and 2 the segments borrowed are close to the existing phonetic resources of the language.

The third class of cases where the borrowed segment seems to be appropriately considered to be close to the existing resources of the recipient language are those where borrowing results in the promotion of a well-defined allophonic variant to phonemic status. A well-known example of this type is the rise to phonemic status of the voiced allophone [v] of the fricative /f/ in English. In Old and early Middle English, voiced fricatives were allophones of the voiceless fricatives in medial position between voiced segments. During the Middle English period the complementary distribution of [f] and [v] was disrupted because loanwords from Old French with initial /v/ came into English with the initial voiced fricative maintained (e.g. vein, vine, veil); also words were incorporated into East Midland dialects from Southern dialects in which all fricatives had become voiced (e.g. vat, vixen). These borrowings predate the loss of final vowels which caused /z/ and /ð/ to become phonemic in pairs like "house (n.), house (v.)" and "breath, breathe". A similar allophony for stops in Central Zoque was disrupted by loanwords from Spanish with voiced stops in initial

positions and in medial clusters (Wonderly 1946), and Russian loanwords in Chuvash have established both voiced stops and voiced fricatives as phonemic.

A complex situation exists in the New Guinea language Telefol (Healey 1964). The phoneme /b/ appears as [b] initially, as [b] or [β] medially and as lightly aspirated [p] finally. Loanwords from Neo-Melanesian with initial or medial [p] are establishing a contrast between /p/ and /b/. Since Telefol has /t/ (which appears as [t] in all environments) the loan segment /p/ can be regarded as belonging in class 1 as well as in class 3 by our definitions. A somewhat similar situation exists with regard to the loan segment /g/. Natively, [g] occurs as the result of coalescence of /k/ and /l/ at morpheme boundary; loans with initial /g/ have introduced a contrast between /k/ and /g/. Since both /k/ and /b/ exist, this case also belongs both to class 3 and class 1. The third loan segment in Telefol is regarded as being /d/. The earlier situation seems to have been complementary distribution of [d], [l] and [r], with [d] initial, [l] final and [r] medial. This phoneme has been treated as /l/ rather than /d/ since the pattern of allophony is quite different from that of /b/, for example, there is no final devoicing. Although different phonemic solutions are possible it is nonetheless true that it is the borrowing of items with /d/ in medial position that leads to the establishment of contrast between /l/ and /d/. Given pre-existing /b/ and /t/, the borrowing of /d/ belongs also to both class 1 and 3.

In Hausa, /r/, and /ʔ/ have become established as phonemes through borrowing from Arabic, Tuareg and elsewhere. The situation with respect to /r/ is complex (Newman 1980), but essentially it seems that the occurrence of trilled /r/ could be predicted from positional and morphological facts. Loans such as / ɓerà/ "mouse" (< Arabic faʔra) have extended the range of /r/ to positions where it has no predictability. As for [ʔ], it originally occurred as a predictable onset to an initial vowel. It could not be regarded as an allophone of any other consonant phoneme. However, it could appear medially as a result of reduplication of a vowel initial item, e.g. from /àikaa/ "to send" the iterative verb /àʔʔàikaa/ is formed, parallel to such forms as /gàmu/ --> /gàggàmu/ "to meet". Hence there was fertile ground for the adoption of /ʔ/ as a phonemic segment, representing both Arabic /ʔ/ and /ʕ/. This case is a little different from the other class 3 examples, but it is included with them since it also involves an existing non-phonemic segment being promoted to phonemic status.

There are 14 examples in class 3 in our data. It seems likely that more might be found if fuller descriptions were available of some of the less well documented languages, or if more exhaustive research was conducted. Note that this would have the effect of adding to the group of examples in which a close segment is added to the recipient language.

(b) Remote sounds

The remaining classes cover those loan segments which are relatively remote from the pre-existing resources of the recipient language. In class 4 are the remaining cases in which the recipient language has added a segment with a previously unrepresented manner but at a place where existing segments occur. In this class, the condition that there are <u>two</u> segments that are only minimally different is not satisfied. The largest number of cases in class 4 involve addition of stops which create new series. Tiwa and Tunica, for example, have both borrowed voiced stops /b, d, g/. These languages already have voiceless stops (aspirated in the case of Tunica) at these three places but they only have

nasals at two of them, lacking a velar nasal. Hence the loan segment /g/ is regarded as more distant than /b/ and /d/, which fall into class 2, whereas /g/ is in class 4. Po-Ai (a Northern Tai dialect from Yunnan) has only one native series of stops, which are voiceless unaspirated. It has borrowed a complete series of aspirated segments /p^h, t^h, $t\int^h$, k^h/ from Chinese. Somewhat similarly, new series of voiceless aspirated and breathy voiced stops have entered Telugu from Hindi and other Indo-Aryan languages or as part of learned Sanskritic vocabulary. Sources differ on the question of whether these two "aspirated" stop series are actually pronounced differently from the plain series in anything other than a rather artificial high style, and different conclusions might be reached about the extent to which they really represent established loan segments. However, they have been included in our counts in class 4.

There is no question that the clicks in Xhosa and other Southern Bantu languages are fully integrated. Standard analyses of Xhosa suggest that there are 12 click phonemes (Davey 1975). These entered the language as part of the loan vocabulary from Khoisan languages which, by some estimates, amounts to some 40 to 45% of the Xhosa lexicon (Louw 1977). Several other fricative and affricate segments that fall into class 1 have also been borrowed. The clicks fall in class 4. Most of the identified loanwords appear to come from Central Khoisan languages, especially Nama and !Ora (Korana), although this appearance may be partly an artifact of the much greater knowledge of these languages than of other possible donor languages (Louw 1974). In Xhosa, there are three types of clicks - dental, alveolar lateral, and palatal - in each of four series - plain voiceless, aspirated voiceless, breathy voiced, and nasalized voiced. Korana has 24 clicks, four types in six series, indicating that Xhosa has substantially reduced the complexity of the segmental inventory it was selecting from. Specifically, it has merged the separate dental and alveolar series (Xhosa has no dental/alveolar contrast) and eliminated distinctive glottalized and velar-affricated click releases, and made several other realignments (Lanham 1962, Davey 1975). This case is important in that it shows that even when contact between languages is extremely intimate (Ownby 1981), there is not a wholesale borrowing of unfamiliar segments.

Several of the other cases in class 4 involve the borrowing of liquids. It is unusual for a language to have more than one lateral approximant or one rhotic segment and these are most commonly found with dental/alveolar place of articulation. Hence such segments are typically borrowed only if the recipient language lacks an original rhotic or lateral. Consequently, after borrowing has occurred, these segments form a series by themselves, although at a place of articulation which is already represented. Among languages which have borrowed a liquid in this fashion are Swahili, Hausa, Mazatec, Tiwa, Moxo, Guarani and Wappo. Wappo has created both plain and laryngealized /r/'s from the loan process. Guarani had no original laterals and has borrowed both /l/ and / ʎ/ from Spanish. There are 54 examples of class 4 in total.

The fifth class of cases are those in which a loan segment introduces a new place of articulation to an existing series where no pre-existing segment occurs at that place of articulation. Loans of this sort are quite rare. Only four cases are observed in the data file. These examples are the voiceless uvular stop and fricative segments /q/ and /χ/ in Hindi-Urdu, a language which had no previous uvular segments, and the palato-alveolar fricative /∫/ borrowed into both Finnish and Ostyak. More will be said below about the Hindi case.

The sixth class of cases are those where the borrowed segment has both a place and a manner which is different from any pre-existing segment in the recipient language. Hindi-Urdu again provides examples of this class, having /ʔ/ and / ʁ/. Note that /ʔ/ and /h/ have been defined as unique segment types, and hence whenever these are borrowed they fall into class 6, unless there are facts similar to those reported for Hausa above. The other language in our sample which has borrowed /ʔ/ is Bashkir, and Georgian and Rumanian have borrowed /h/. In addition, Guarani has borrowed /tʃ/ into a system which had neither affricate nor palatal segments. Thus there are 6 cases in class 6.

3. Summary of segment borrowing patterns.

The pattern in this data is quite clear. There are very many more instances where the segment borrowed fits closely in with the existing resources of the language. The number of cases in classes 1-3 is 123, whereas the number in classes 4-6 is 61. From this, it is appears that segments closer to the existing resources of the recipient language are more likely to be borrowed. Moreover, a relatively high proportion of the segments in classes 4-6 are somewhat dubiously integrated into the languages they have joined. The aspirated stops in Telugu have already been mentioned. Hindi-Urdu as described by Kelkar (1968) has 9 borrowed segments, of which six are limited to particular communal speech styles outside "common core" Hindi (Ohala 1983). These six are precisely the ones which fall into classes 4, 5 and 6.

We have also noted a number of cases where it seems that a borrowed segment is itself modified in a way that brings it closer to existing sounds in the recipient language. The examples include Hausa /ʔ/ corresponding to Arabic /ʕ/, elimination of the dental/alveolar contrast between clicks entering Xhosa, and Sinhalese accepting /Φ/ where the donor languages have /f/. We also note the relative rarity of the borrowing of sui generis segments such as /ʔ/ and /h/. Both of these facts strengthen the impression that languages do not normally borrow segments that are far afield from the phonetic material that is already at hand.

4. Simulation of segment borrowing

Despite the impression that emerges from the data we have surveyed, can we be sure that there is in fact any special selectivity going on in the process of borrowing loan segments? The numbers of loans of the different types reported in this paper might well be a reflection not of preferential borrowing of sounds that fit into the recipient language's phonological pattern, but instead simply the result of the overall frequency across the world's languages of particular sounds and segment inventories. Frequency is necessarily one factor in the borrowing results since no sound can be borrowed from one language to another if the sound is already common to both languages. Segment borrowings can only occur if there are unmatched sounds. Obviously sounds have unequal chances of being borrowed because of this. Similarly, segments which are candidates for borrowing have unequal chances of being close or remote with respect to possible recipient languages because of the differing frequency of various patterns of inventory structure.

It is hard to devise a real-life test that addresses the question of whether the observed patterns of sound borrowings could result from chance. This is

partly because it is difficult to demonstrate that a segment had an opportunity of being borrowed but was in fact rejected. Examining the inventory of languages which frequently figure as donors, such as English, Spanish and Arabic provides some clues in that segments that are less likely to be matched in either place or manner such as the interdental fricatives of English, the palatal lateral of Spanish, and the pharyngeal and pharyngealized consonants of Arabic rarely appear as loan segments in these forms. However, in a very large number of the cases it is likely that the form of the donor language from which borrowings were being made was itself already modified in ways that may have eliminated these rarer types of speech sounds. Moreover, it is hard to be sure that there were appropriate lexical items with the sounds of interest that were good borrowing candidates. Because it is difficult to test the hypothesis that the results of our survey are not due to chance in the field, it was decided to perform a simulation of borrowing. We can show that the actual borrowing patterns are different from those which would be predicted by segment frequency and patterns of inventory structure.

The simulation was done in the following way. First a subsample of all of the consonantal segments in UPSID was constructed by selecting every thirty-fifth entry in the file. This procedure was used in order to increase the likelihood of including in the sample consonants that are members of classes with a distribution that is disproportionately weighted in certain language families. The procedure produced a sample of 207 segments. These segments were then randomly ordered. Next the 317 language names were sorted into a random sequence and the first 200 language names in this sequence were matched one by one with the first 200 segments in the sample of 207. No language was matched with a segment which had originally been drawn from itself. In this way 200 confrontations of "foreign" segments and existing phoneme inventories were simulated and the resulting patterns examined.

Of these 200, 115 were exact matches, that is, the foreign segment (or a close enough relative) already existed in the elected recipient language. This provides us with a rough estimate of the proportion of cases where such matches might occur in real life situations. Three cases where a long consonant was involved in the comparison were counted as matched by their short counterparts.

The remaining 85 cases were possible segment borrowings. In this simulation the greatest number of loan segments fall into class 4, that is, they would add a member of a new series at a place that is already represented in the language, but without two minimally different promoting segments being present. Recall that the most common class in the real data is the gap-filling type (class 1). Although class 1 is the next most frequent in the simulation, it is only about one third as frequent as the class 4 type, and is about equally frequent as the class 6 type, that is, cases where the segment has no close promoters in the pre-existing inventory. This class is one of the smallest in the real language data. No members of class 3 were identified in the simulation, although some undoubtedly exist. Overlooking their existence might distort our conclusions - we will describe a way of correcting for this below. The percentages of cases in each class in the simulation are tabulated in Table 1 below and compared with the percentages in the real data.

Table 1. Comparison of simulated and real loan segment data. Totals and percentages for classes 1 and 3 in the real data include 3 Telefol examples which are included in both classes.

	Simulated data		Real data	
	n	%	n	%
Class 1	16	18.8	91	48.9
Class 2	7	8.2	19	10.3
Class 3	0	0.0	14	7.6
Class 4	44	51.8	54	29.3
Class 5	4	4.7	4	2.2
Class 6	14	16.5	6	3.3

The two sets of percentages were compared using the χ^2 test and found to be significantly different ($p. < .001$). Comparable levels of significance were also obtained for a comparison between real and simulated data grouping together classes 1-3 and 4-6. The comparison was repeated for both the six classes and the grouped data with loan segments in Lake Miwok (Callaghan 1964) included in the real language data. Callaghan suggests that there are 18 segments in Lake Miwok which might have been borrowed from Patwin (Wintun), Pomo or Wappo. Of these, 13 are classed as "remote", making this language comparable to Xhosa in the nature and degree of foreign effect on its inventory. Lake Miwok was not included in the main data because Callaghan seems tentative in making the claim that these segments are in fact all borrowed. With Lake Miwok included, significance levels were still above .001. Finally, the test for the grouped data was repeated with 7.9% added to the total for classes 1-3 and subtracted from the total for classes 4-6 in the simulated data. This represents an estimate of the correction that might be needed to allow for the overlooking of cases which should have been put in class 3 in the simulation. The correction factor is a little higher than the percentage of class 3 examples found in the real data, and assumes all cases would involve reclassification from groups 4-6. Even with this correction there is a significant difference at the .001 level between the simulated and the real data.

5. Summary

The patterns observed in our survey of loan segments and the comparison between the real data and the simulation indicate that languages have a strong tendency to reject as loan segments those segments that are remote from their existing phonetic resources, or to modify them in such a way that they are no longer remote. In other words, languages do behave in the way that is hypothesized in Maddieson (1984) as contributing to maintaining the typical phoneme inventory size.

References

Bach, A. (1965). Geschichte der deutschen Sprache (8th ed). Heidelberg.

Brunt, R.J. (1983). The influence of the French Language on the German Vocabulary. Berlin and New York: De Gruyter.

Callaghan, C.A. (1964). Phonemic borrowing in Lake Miwok. In W. Bright (ed.) Studies in Californian Linguistics, 46-53. Berkeley and Los Angeles: University of California Press.

Cline, D. ms. (1984). What does a sound have to do to get borrowed into a language? Department of Linguistics, University of California, Los Angeles.

Davey, A.S. (1975). Some Aspects of the Phonology of the Noun in Xhosa M. Litt. thesis, University of Edinburgh.

Deroy, L. (1956). L'emprunt linguistique (Bibliothèque de la Faculté de Philosophie et Lettres de L'Université de Liège, Fascicule 141). Paris: Les Belles Lettres.

Hagège, C. and Haudricourt, A-G. (1978). La phonologie panchronique. Paris: Presses Universitaires de France.

Haudricourt, A-G. (1961). Richesse en phonèmes et richesse en locuteurs. L'Homme 1: 5-10.

Haugen, E. (1950). The analysis of linguistic borrowing. Language 26: 210-231.

Healey, A. (1964). Telefol Phonology (= Pacific Linguistics, Series B, 3). Canberra: Australian National University.

Hyman, L.M. (1970). The role of borrowing in the justification of phonological grammars. Studies in African Linguistics 1: 1-48.

Jacobsen, W.H. (1969). Origin of the Nootka pharyngeals. International Journal of American Linguistics 35: 125-153.

Jakobson, R. (1932) [1964]. Prinzipien der historischen Phonologie. Travaux du Cercle Linguistique de Prague 4: 247-267. Translation into French by J. Cantineau printed as an appendix to Principes de Phonologie, N.S. Troubetszkoy, Klincksieck, Paris: 215-336.

Kaye, J.D. and Nykiel, B. (1981). Loan words and abstract phonotactic constraints. Studia Anglica Posnaniensia 13: 21-42.

Kelkar, A.R. (1968). Studies in Hindi-Urdu 1. Introduction and Word Phonology. Pune: Deccan College.

Lanham, L.W. (1962). The proliferation and extension of Bantu phonemic systems influenced by Bushman and Hottentot. Preprints of papers for the Ninth International Congress of Linguists, Cambridge, Mass., 264-270.

Louw, J.A. (1977). The linguistic prehistory of the Xhosa. In W.J.G. Möhlig et al., eds. Zur Sprachgeschichte und Ethnohistorie in Afrika, 127-151. Berlin: Reimer.

Louw, J.A. (1974). The influence of Khoe on the Xhosa language. Limi 2/2: 45-62.

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

Newman, P. (1980). The two R's in Hausa. African Language Studies 17: 77-87.

Ohala, M. (1983). Aspects of Hindi Phonology. New Delhi: Motilal Banarsidass Delhi.

Ownby, C.P. (1981). Early Nguni history: linguistic suggestions. Suid-Afrikaanse Tydskrift vir Afrikatale, Byblad 1981: 60-81.

Petrovici, E. (1957). Kann das Phonemsystem einer Sprache durch fremden Einfluss ungestaltet werden? - Zum slavischen Einfluss auf das rumänische Lautsystem. Mouton, The Hague.

Wonderly, W.L. (1946). Phonemic acculturation in Zoque. International Journal of American Linguistics 12: 92-95.

Appendix

This appendix contains a listing of the 184 loan segments used as the database for this paper. Loan segments are listed by language, with languages grouped by family. Each entry consists of the following: (i) a number from 1 to 6, indicating the class of the borrowing according to the classification given in the text, (ii) the segment borrowed, in the form of the most representative allophone it has in the recipient language, and (iii) the closest segment(s) in the recipient language to the borrowed segment. A dash (-) indicates that no segment is considered to be close. A segment enclosed in [ ] is an allophone; if preceded by ~ it is an allophone of the segment before ~ . Entries for the same language are separated by a semi-colon.

Irish 1 t^h ($\underline{t}^h$, d)

English 3 v (f ~ [v])

German 1 ʒ (z, ʃ)

Lithuanian 1 f (v, s); 1 f^j (v^j, s^j); 1 x (k, s); 1 x^j (k^j, s^j); 1 ɣ (g, z); 1 $ɣ^j$ (g^j, z^j); 1 ts (tʃ, dz); 1 ts^j (tʃ, t^j)

Russian 1 f (v, x); 1 f^j (v^j, s^j)

Romanian 1 z (s, v); 1 ʒ (ʃ, v); 6 h (-)

Hindi-Urdu 5 q (k, -); 6 ʔ (-); 1 f (p, s); 2 z (s, $\underline{d}$); 5 χ (s, -); 6 ʁ (-); 1 ʃ (tʃ, s); 4 ʒ (dʒ, -); 4 $\tilde{\underline{r}}$ ($\underline{n}$, r)

Bengali 1 s (t, ç)

Panjabi 1 f (p, $\underline{s}$); 1 $\underline{z}$ ($\underline{s}$, $\underline{d}$)

Sinhalese 1 Φ ($\underline{\beta}$, $\underline{s}$); 1 ʃ (tʃ, $\underline{s}$); 1 ŋ ($\underline{n}$, g)

E. Armenian 1 f (v, z)

Ostyak 5 ʃ (s, $\underline{ts}$)

Cheremis 4 p^j (p, -); 4 t^j (t, -); 2 ts (t, s); 1 f (p, s); 4 s^j (s, -) 1 x (ɣ, s)

Komi 1 ts (t, tʃ); 1 f (v, s); 1 x (k, ç)

Finnish 1 b (p, d̲); 1 g (k, d̲); 1 f (v, s̲); 5 ʃ (s̲, -)

Hungarian 1 dʒ (tʃ, d̲z̲)

Azerbaijani 1 k^h (k, c^h)

Chuvash 3 b (p ~ [b]); 3 d (t ~ [d]); 3 g (k ~ [g]); 3 z (s ~ [z]); 3 ʒ (ʃ ~ [ʒ]); 1 ts (tʃ, s); 1 f (p, s)

Yakut 1 p (p:, b); 1 f (p:, s̲); 4 v (b, s); 2 z̲ (s̲, d); 1 ʃ (tʃ, s̲); 4 ʒ (dʒ, s̲)

Bashkir 1 f (p, θ); 4 v (b, θ); 2 z (s, d); 4 ʒ (ʃ, d); 2 ts (t, s); 2 tʃ (t, ʃ); 6 ? (-)

Khalaj 1 f (p, s); 1 v (b, z); 4 dʒ (tʃ, d̲); 1 x (k, s)

Mongolian 1 p^h (b, t^h); 1 k^h (g, t^h); 1 Φ (b, s); 4 β (b, s); 4 l̥ (l, s)

Evenki 1 p (b, t); 2 z (s, d)

Lelemi 1 p (b, t)

Yoruba 1 p (b, t)

Swahili 4 (,)

Xhosa 4 ɬ (t'); 4 ʃ (c'); 4 ɮ (ɬ); 4 $ɬ^h$ (t^h); 4 $ʃ^h$ (c^h); 4 $ɮ^h$ (ɬ); 4 $ɬ^ɦ$ (d); 4 ^{ɦ} (ʄ); 4 $ɮ^ɦ$ (l); 4 ɬŋ (t', n); 4 ʃŋ (c', n); 4 ɮŋ (ɬ, n); 1 kx' (k', tʃ'); 1 x (k^h, s); 1 ɣ (g, z); 1 ts' (t', tʃ')

Kanuri 1 x (k, s)

Amharic 1 p (b, t); 1 p' (b, t')

Hebrew 1 tʃ (ts, ʃ); 4 dʒ (d, ʃ); 1 ʒ (ʃ, z); 1 ŋ (n, g)

Hausa 4 ɗ (d); 3 r (ɾ ~ [r]); 3 ? ([?])

Po-Ai 4 p^h (p, -); 4 t^h (t, -); 4 $tʃ^h$ (tʃ, -); 4 k^h (k, -)

Telefol 1,3 p (b ~ [p], t); 1,3 d (t, b, [d]); 1,3 g (k, b, [g])

Sab'an 1 g (k, d)

Zoque 3 b (p ~ [b]); 3 d (t ~ [d]); 3 g (k ~ [g])

Tzeltal 1 d̲ (t̲, b); 1 g (k^h, b)

Otomi 2 tʃ (t, ʃ); 4 l (ɾ)

Mazatec 1 p (t, m); 4 b (^{n}g, β); 1 ð (s, β); 1 ɣ (k, β); 4 r (l)

Tseshaht (Nootka) 1 q' (q, k'); 1 q^{w}' (q^{w}, k^{w}'); 1 χ (q, x); 1 χ^{w} (q^{w}, x^{w})

Quileute 1 g (k, d)

Yacqui 1 d (t, b); 1 g (k, b); 1 f (p, s)

Tiwa 2 b (p, m); 2 d (t, n); 4 g (k, -); 4 r (l)

Tarascan 2 b (p, m); 2 d (t, n); 2 g (k, ŋ); 1 f (p, s); 1 ɲ (ŋ, ʂ); 1 l (l̥, r)

Tunica 2 b (p^{h}, m); 2 d (t^{h}, n); 4 g (k^{h}, -); 1 f (p^{h}, s)

Wappo 2 b (p, m); 2 d̲ (t̲, n); 4 g (k, -); 1 f (p, s); 4 r (l); 4 r̲̃ (l̲̃)

Moxo 2 b (p, m); 2 d (t, n); 1 f (s, β); 1 ŋ (k, n); 4 l (ɾ)

Guarani 6 tʃ (t̲, s̲); 1 f (v, s̲); 1 ð (t̲, v); 1 x^{w} (x, k^{w}); 1 $ɣ^{w}$ (ɣ, k^{w}); 4 ř (ɾ); 4 l̲ (ɾ); 4 ʌ (ɾ)

Ticuna 4 f (p, -); 4 s (t, tʃ)

Barasano 4 s (t, -)

Telugu 4 p^{h} (p, -); 4 $ṭ^{h}$ (ṭ, -); 4 k^{h} (k, -); 4 b̤̤ (b, -); 4 d̤̤ (d̤, -); 4 ḍ̤̤ (ḍ, -); 4 g̤̤ (g, -); 1 tʂ (ts, ʂ)

Georgian 6 h (-)

Kabardian 1 ʕ (ħ, ʁ)

Burushaski 1 f (p, s); 1 χ (q, s)

Macintosh models and plots for phoneticians

Peter Ladefoged, UCLA Phonetics Lab, Los Angeles, CA 90024

A previous paper (Ladefoged 1985) described some of the phonetic symbols available on the Macintosh computer.[1] A number of programs that are of interest to phoneticians are now available.

The program **PlotFormants** graphs F1 and F2 values with the axes and scales arranged as in Ladefoged (1982), so that the resulting plot is similar to traditional descriptions of vowels. The first formant is plotted on the abscissa (the vertical axis), with either the second formant or, the difference between the second formant and the first formant, as an option, on the ordinate (the horizontal axis). The importance of the first formant is emphasized by having the scale on the ordinate expanded to double that on the abscissa. The points for the vowels and the tickmarks and the Hz values on the scales are all displayed using mel scale intervals.

The program will plot up to 100 vowels in up to 15 different groups, using up to 7 different symbols. Another option allows the user to enclose all the points in selected groups within an ellipse, centered on the mean of the group, with axes aligned along the

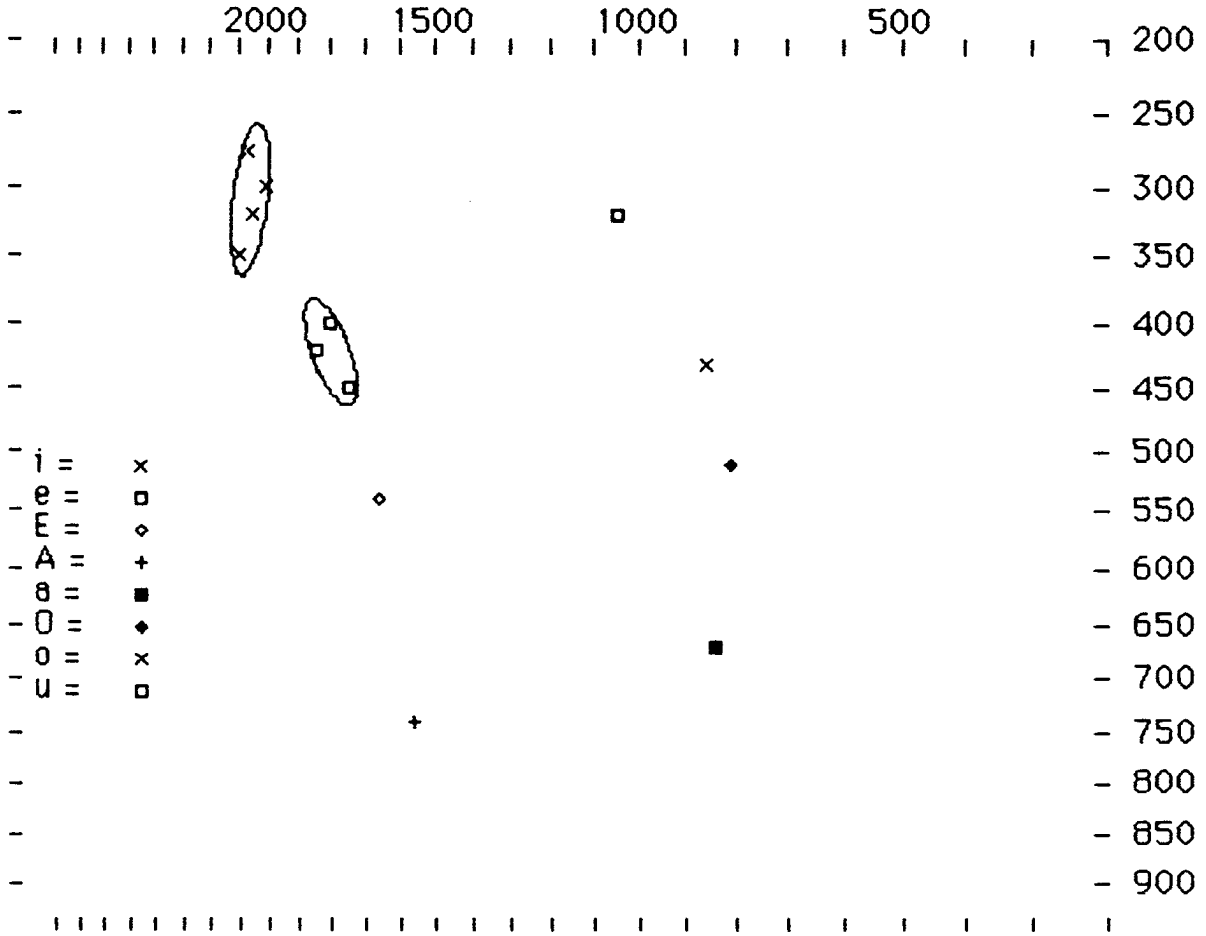


Figure 1. Some vowels of American English as graphed by the program **PlotFormants.**

principal components of the dispersion of the group, and with radii equivalent to two standard deviations. Given appropriate sampling, an ellipse of this kind is likely to enclose 95% of the vowels that could be in this group. Figure 1 is an example of a plot produced in this way. The figures produced by the program are MacPaint documents, and can, of course, be further enhanced by the use of the regular MacPaint techniques, as will be illustrated by one of the later figures in this paper.

The second program, **MouseVocal** allows the user to manipulate the shape of the vocal tract either by using the Macintosh's mouse, or by inputting values of the vocal tract parameters from the keyboard. As discussed in earlier publications (Harshman, Ladefoged & Goldstein, 1977; Ladefoged 1980), the vocal tract is drawn with a fixed part, the roof of the mouth and the back wall of the pharynx, and with moveable lips, tongue, jaw and larynx. The vocal tract is considered to be formed from 18 tubes, one centered on the lips, specified in terms of the parameter lip height, another on the teeth, and thus directly dependent on the parameter specifying the jaw opening, and the

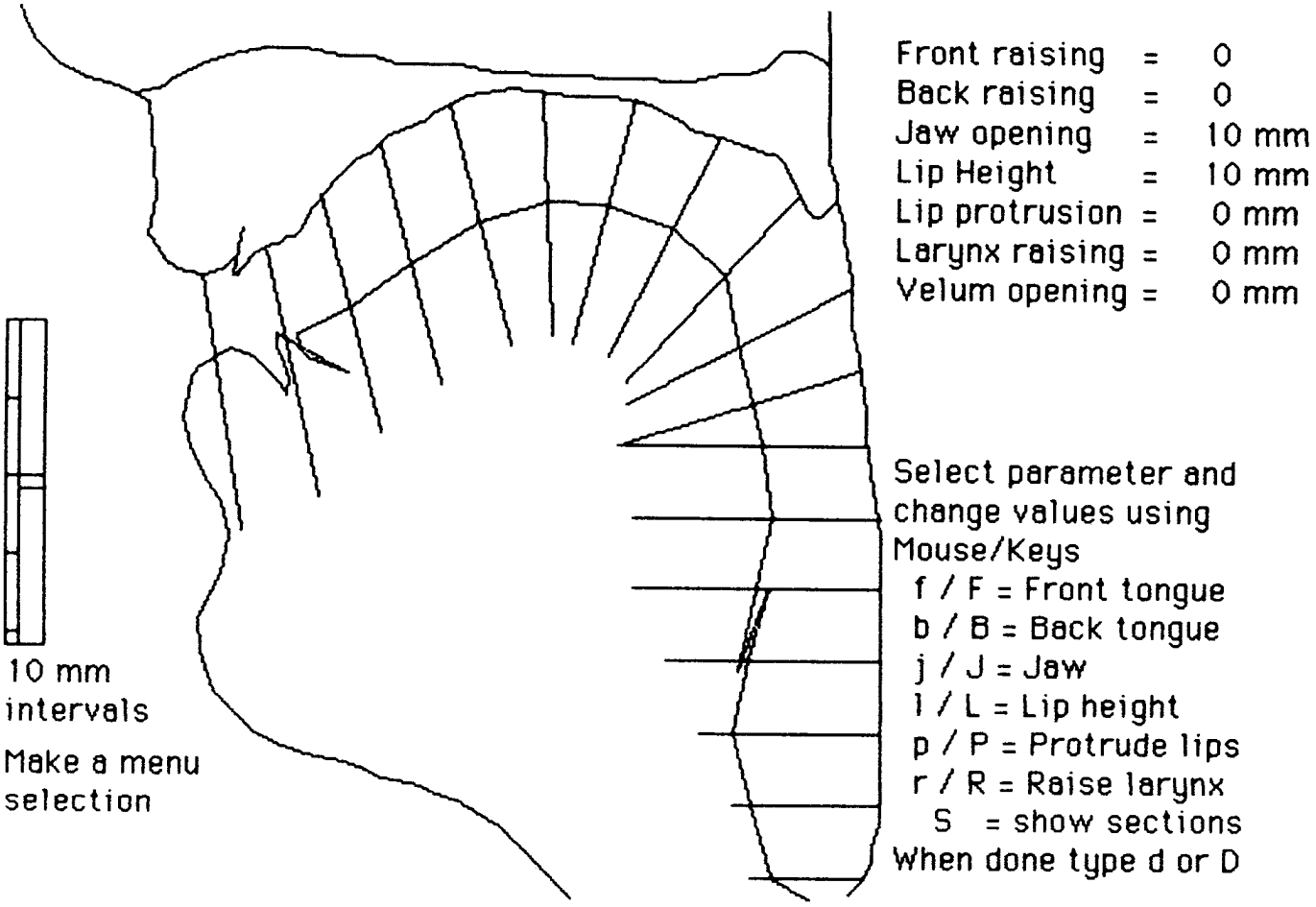


Figure 2. The sections of the vocal tract and the primary menu used in **MouseVocal.**

16 others determined by the positions of points on the tongue. The shape of the tongue is determined by one parameter specifying the degree of front raising of the tongue and another specifying the degree of back raising. Figure 2 shows the centers of the sections of the vocal tract, and also the parameters and the primary menu used by the program. The scale on the left is used for changing parameter values with the mouse. As may be seen, the program allows for both the degree of lip protrusion (with the upper lip being protruded half the distance specified for the lower lip), and the degree of larynx raising (which is distributed over the sections of the vocal tract below the epiglottis). These six parameters are thus a subset of those described in Ladefoged (1980), they are also a subset of those implemented in the more powerful version of this program currently running on the larger computers in the UCLA Phonetics Lab. These programs include parameters specifying the lip width (as well as lip height) and the degree of velum opening (shown on Figure 2, but not currently available on the Macintosh programs).

A secondary menu, which appears when the specification of the required vocal tract has been completed allows the user to calculate the formant frequencies, using a variant of the algorithm described by Liljenkrants and Fant (1975). This algorithm lacks much of the sophistication of later work (eg Wakita and Fant 1978) which is currently in use on the UCLA Phonetics Lab computer. But it provides a reasonable estimate of the formant frequencies. The left hand side of Figure 3 shows the best approximations to three extreme vowels, /i,a,u/. The formant values are quite appropriate for /a/, and fairly good for the other two vowels. In the case of /i/, F1 and F2 are as required, but F3 is too high. In order to produce an appropriate F2 for /u/, a very small lip opening had to be specified; and this resulted in an unnaturally low value for F1. (The right hand side of Figure 3 will be discussed later.)

Less extreme vowels are better determined; and the program is certainly useful for showing the articulatory-acoustic relations involved in a given movement of one or more parameters. The upper part of Figure 4 was produced with **FileVocal** a version of this program in which the parameters are read in from a previously written file. In this case, the effect of varying three different parameters was simulated, in each case starting from the same position (the slightly heavier line in Figure 4). The tongue is in the mid high position, with the lower teeth and lower lip 8 mm from the upper surface of the vocal tract, and with the larynx considerably raised. For the first movement the front of the tongue was lowered in five equal steps, with the back of the tongue naturally becoming closer to the back wall of the pharynx. Next, while retaining the original tongue position, the lips were narrowed and protruded, with the jaw also being raised half as much as the lower lip. Finally, again starting from the same original position, the larynx was lowered considerably (five steps, each of 3 mm). The program also writes out a file of F1 and F2 frequencies, which can later be plotted (as in the lower part of Figure 4) using the **PlotFormants** program. It is interesting to note that

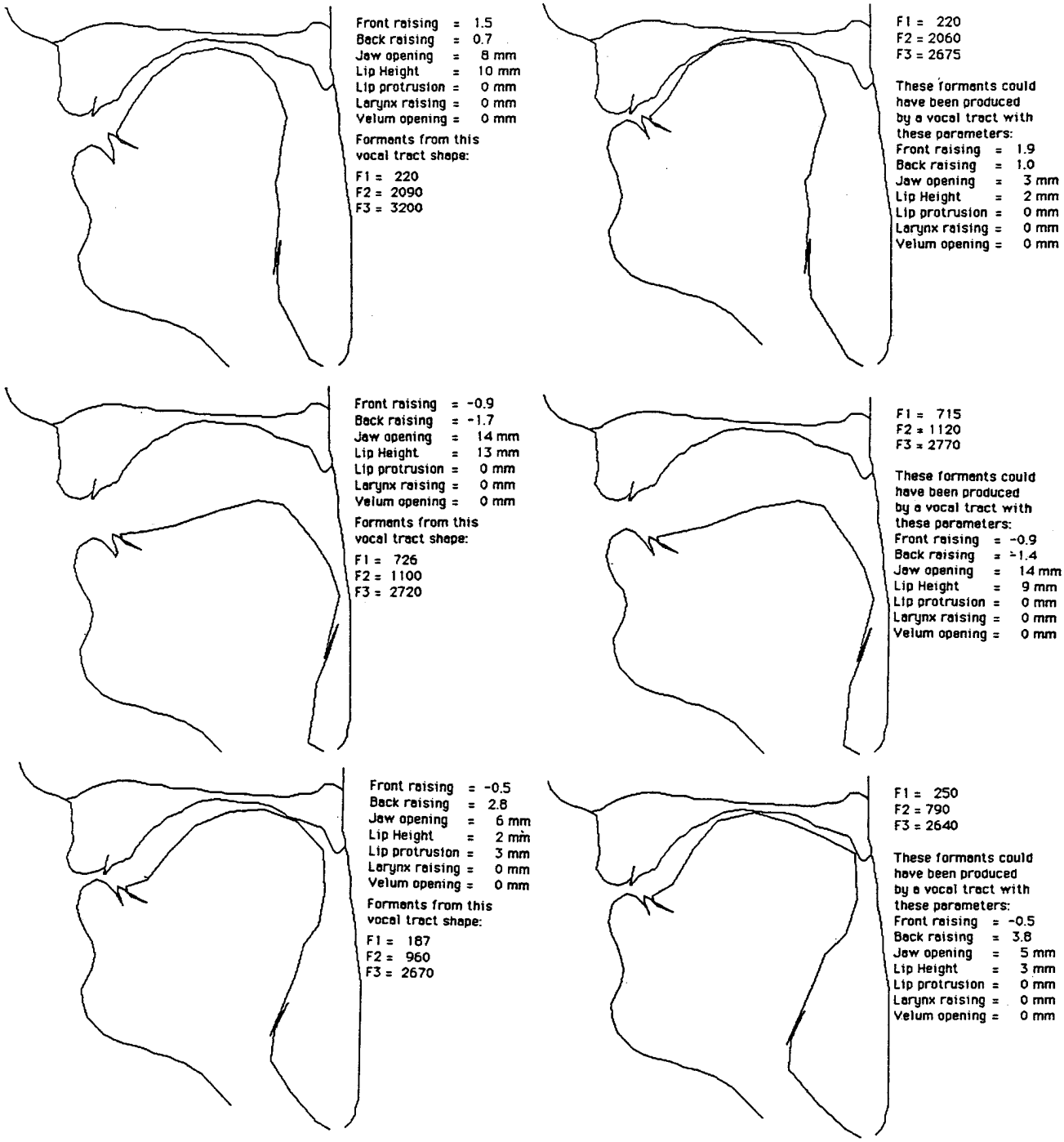


Figure 3. The outputs of the programs **MouseVocal** (left hand side) and **FormantsTracts** (right hand side) for three extreme vowels /i,a,u/.

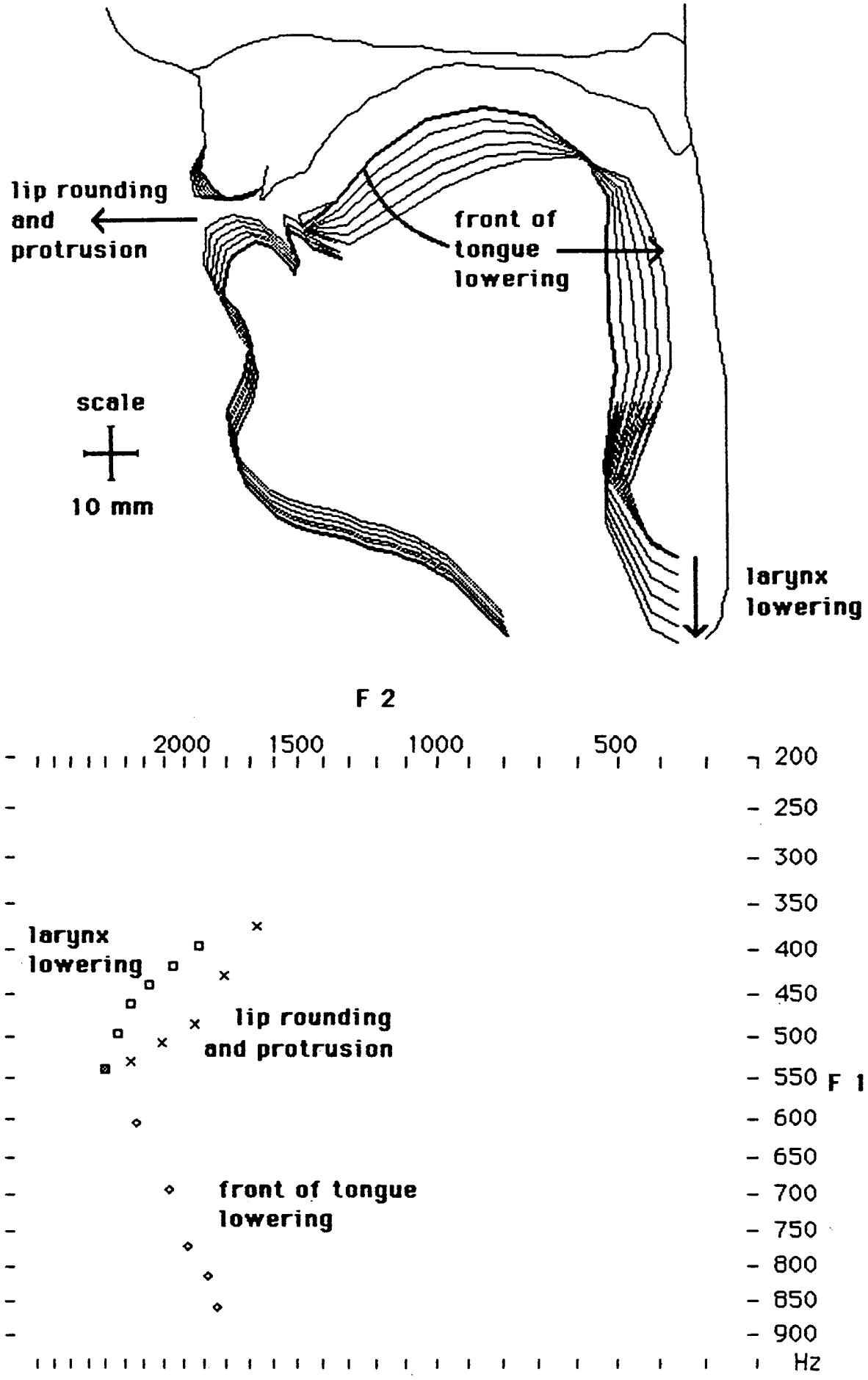


Figure 4 (Upper part.) **FileVocal** representations of three different articulatory movements starting from the same vocal tract position, shown by the heavy line. (Lower part.) The formants generated by this program, as graphed by **PlotFormants** (Labels added to both parts using **MacPaint.**)

the effects of the lip and larynx actions are similar, but far from identical; and the equal steps of tongue movement produce far from equal acoustic changes.

The final program in this set, **FormantsTracts**, takes as input the values of the first three formants, and then produces a diagram of the sagittal section of a vocal tract shape that *could* have produced these frequencies, as shown in the right hand side of Figure 3. This program uses the values of the formant frequencies to generate appropriate values of the factors representing the degree of front raising and back raising of the tongue. These values are then used to generate the sagittal dimensions at 16 points along the surface of the tongue. The vertical distance between the lips is also derived from the formant frequencies; and the position of the jaw is determined from that required for the lips and the tongue. The other vocal tract parameters shown in Figure 2, the amount of lip protrusion, larynx raising, and velic opening, cannot be estimated from formant frequency data.

Diagrams produced in this way have been shown by Ladefoged, Harshman, Goldstein, and Rice (1978) to have a very high correlation (r = .91) with tracings from a cine x-ray film of 5 American English speakers saying words containing 10 different vowels. The side by side comparison in Figure 3 shows that, if appropriate formant frequency values are entered for /i/, the tongue shape is reasonable, but the lip and jaw openings are badly underestimated. The lip opening is also underestimated for /a/. The lip and jaw positions are quite appropriate for /u/, but the mass of the tongue is unnaturally distended. So for these extreme vowels, if the correct shape is entered (as on the left of the figure), the calculated formants will be slightly wrong; and if the correct formants are entered (as on the right of the figure) the derived shapes will be somewhat inaccurate. The **FormantsTracts** algorithms have been further validated by using the published formant frequencies for a set of Russian vowels to recover a series of vocal tract shapes that are very similar to those in the published diagrams based on x-rays of this speaker (Fant 1962). But these algorithms will not produce reliable results for vowels produced with degrees of liprounding that are not comparable with those in American English vowels, or for vowels involving additional features such as r-coloring or pharyngeal constrictions.

These programs provide simple but useful tools for students to study basic articulatory acoustic relations. They are all written in Modula-2, and are available from the UCLA Phonetics Lab. Just send us two empty disks, and we will send you runnable programs; and, if you own Modula-2, we will also send you the sources, so that you can modify them and send back the improved versions to us.

1. Wayne Harbert (Linguistics, Cornell University, Ithaca, NY 14853) has kindly pointed out to me that I am wrong in saying in this first paper that the Macintosh lacks an overstrike capability. He has shown how overstrike characters can be produced using the Resource Editor application, and creating characters in which the right hand boundary mark is moved over so that it *precedes* the left boundary mark.

References

Fant, Gunnar. 1962. ***The Acoustic Theory of Speech Production.*** Mouton: The Hague.

Harshman, Richard, Ladefoged, Peter, and Goldstein, Louis. 1977. "Factor analyses of tongue shapes." ***Journal of the Acoustical Society of America.*** 62.3, 693-707.

Ladefoged, Peter. 1985. "Macintosh usage for linguists" ***UCLA Working Papers in Phonetics.*** 60, 84-86.

Ladefoged, Peter. 1982. ***A Course in Phonetics.*** Harcourt Brace and Jovanovich: New York.

Ladefoged, Peter. 1980. "What are linguistic sounds made of?" ***Language.*** 56.3, 485-502.

Ladefoged, Peter, Harshman, Richard, Goldstein, Louis, and Rice, Lloyd. 1978. "Generating vocal tract shapes from formant frequencies." ***Journal of the Acoustical Society of America.*** 64.4, 1027-1035.

Liljenkrants, Johan and Fant, Gunnar. 1975. "Computer program for VT-resonance frequency calculations." ***Speech Transmission Laboratory, Quarterly Progress and Status Report.*** 4, 15-20.

Wakita, Hisashi, and Fant, Gunnar. 1978. "Toward a better vocal tract model." ***Speech Transmission Laboratory, Quarterly Progress and Status Report.*** 1,9-29.

Measures of spectral tilt

Michel Jackson, Peter Ladefoged, Marie Huffman,
and Norma Antoñanzas-Barroso.

Differences in voice quality have often been described in terms of differences in glottal spectrum. It is now fairly well known that one of the major acoustic differences between breathy voice and modal voice is that breathy voice has proportionately more energy at the fundamental than modal voice does. Fant (1980, 1983a, 1983b), among others (e.g. Flanagan 1958), suggests that various phonation types and degrees of vocal effort can be characterized by a measure of the overall spectral tilt of the glottal flow spectrum. Figure I shows variation in spectral characteristics as Fant's model (1983a) goes from highly symmetrical, breathy voice-like pulses to asymmetrical modal or creaky voice-like pulses.

The aim of this paper is to investigate whether such a spectral tilt is measurable. Answering this question obviously depends on our ability to recover the glottal waveform, from which the glottal spectrum can be determined. We have used inverse filtering to investigate the vowels of a number of different languages that contrast phonation types. We analysed FM recordings of airflow data from three speakers of Hmong, a Sino-Tibetan language that contrasts breathy and modal voice; FM audio recordings (made available by Ian Maddieson) of two speakers of Burmese, a Tibeto-Burman language that contrasts a slightly creaky voice with regular modal voice; and AM audio recordings of 9 speakers of !Xóõ, a Khoisan language that contrasts breathy, modal, and laryngealized phonation.

Since we are interested in spectral characteristics we did not have to worry about possible phase distortion, and could use either FM recordings which preserve phase information, or AM recordings which do not. But we do have to worry about the correct procedures for removing supraglottal influences on the sound wave. The most important of these influences are vocal tract resonances and losses. These are usually described in terms of formants; the losses generally increase formant bandwidths.

Our general procedure for inverse filtering has been described by Javkin, Antoñanzas-Barroso and Maddieson (1985). Inverse filtering requires the specification of both formant frequencies and formant bandwidths. There are no major problems in determining the frequencies of the vocal tract resonances. The resonance locations determined from linear prediction polynomials are well defined, and enable these frequencies to be determined fairly reliably. But it is by no means obvious how one should determine the bandwidths of these resonances. The only thing that is clear is that the observed bandwidths of the formants in the waveform reflect not only supraglottal resonances, but also properties of the glottal source itself. It was important to us in our investigation of spectral slope to ensure that we removed all and only those aspects of the waveform that were due to supraglottal influences. We wanted to be certain that we had removed nothing that could be considered due to the glottal source itself. Accordingly we tried a number of different methods for determining appropriate bandwidths.

Using our Hmong airflow data, drawn from the data described in Huffman (1985), we first inspected the bandwidths derived from LPC analysis. These were considerably larger than normal, perhaps because of damping due to the airflow

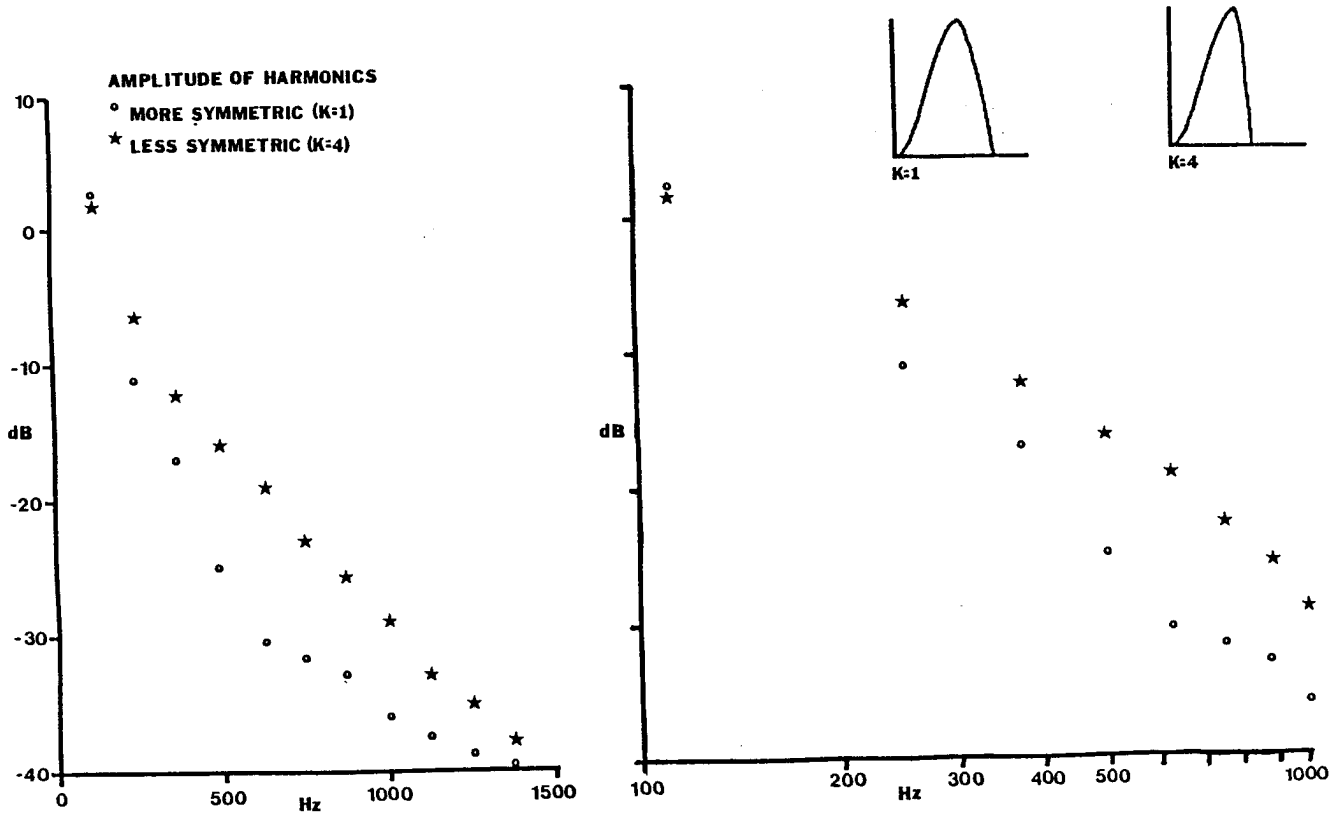


Figure I. Modeled effect of glottal flow pulse variation (after Fant)

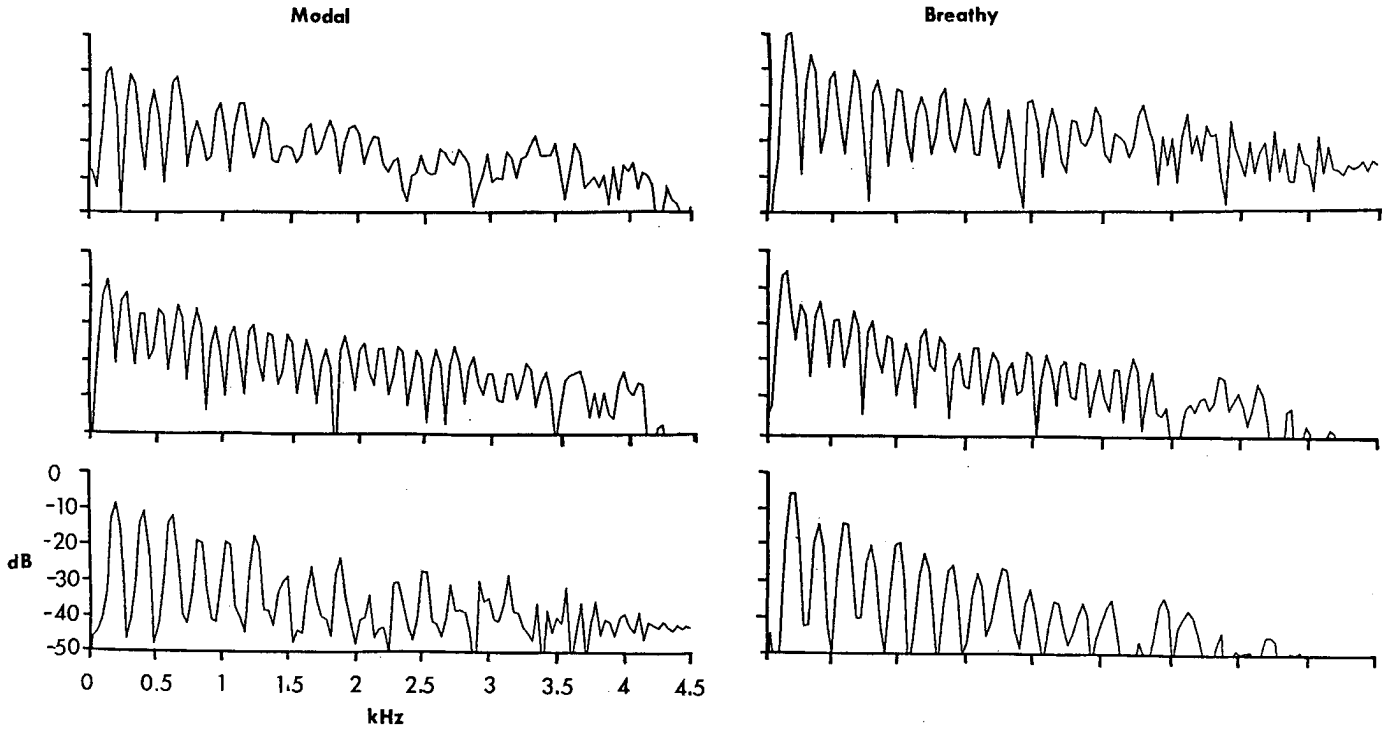


Figure II. Representative spectra of inverse-filtered Hmong modal and breathy vowels

mask. Furthermore, since LPC-derived bandwidths model not only supraglottal resonances, but also the way in which these resonances are excited by the glottal source, using these bandwidths in the input to the inverse filter program would clearly be wrong. After observation of the residual formant oscillations in the inverse filtered glottal flow waveform, we used a trial and error procedure, varying the bandwidths of the resonances in the vocal tract transfer function, and verifying our success (or lack of it) by observation of the filtered waveform. We concluded that we had an appropriate value for the bandwidths when there were no oscillations in the filtered waveform that could be attributed to one of the given formant frequencies. We found that we were able to make observations of this kind most easily in the comparatively flat part of the waveform corresponding to the closed phase of the glottal cycle, when such existed.

This technique is, of course, open to the objection that we are influencing our analysis by our expectation of what the glottal waveform should look like. We felt that by only varying formant bandwidths, and leaving formant frequencies constant, we would still obtain relatively unbiased results. The formant frequencies were determined by LPC analysis, and verified on spectrograms.

Typical spectra of the Hmong tokens are shown in Figure II. Spectral tilts for these tokens were calculated from harmonic peaks estimated by parabolic interpolation.

Our other technique for determining formant bandwidths was to survey the literature to see what they ought to be. This was instantly confusing, as may be seen from Figure III, which shows observed data from Fujimura and Lindqvist (1971) and model predictions by Wakita and Fant (1978) for B1. Fujimura and Lindqvist note that the observed bandwidths were measured with a closed glottis, in such a way as to "exclude the unknown factor of the source spectrum". It seems that, for male speakers, when the first formant has a frequency greater than 400 Hz, the best estimate of the bandwidth is approximately 40 Hz. The situation for the second and higher formants is equally confused. We decided that appropriate values for B2, B3, and B4 would be 50, 60, and 80 Hz, respectively. As there is very little energy in our flow records above 3,000 Hz, the contribution of the higher formants is negligible.

The Burmese data consists of utterances by one male and one female speaker of words contrasting in phonation type, drawn from the data described in Javkin and Maddieson (1983). The female voice typically exhibits only four to six harmonics with significant amplitude. Formant locations from LPC analysis and bandwidths of 50, 50, 100, and 100 Hz for B1 through B4 were used as input to the inverse filtering program. When fitting the peaks of the harmonics to linear spectral envelopes, we only used peaks that were clearly (> 5dB) above the background noise level. Typical spectra of Burmese creaky and modal tokens are shown in Figure IV. Spectral tilts for these tokens were calculated from harmonic peaks located on high-resolution (5 Hz) spectra, obtained using the chirp z-tranform of Rabiner, Schafer, and Rader (1969) from inverse filtered waveforms.

Our !Xóõ data were recorded on tape recorders with unknown phase characteristics. Accordingly we could not rely on the actual wave shape that was produced by inverse filtering, and we were unable to make trial and error estimates of appropriate bandwidths. LPC-derived formant frequencies and fixed bandwidths from the literature review were used for the first four formants.

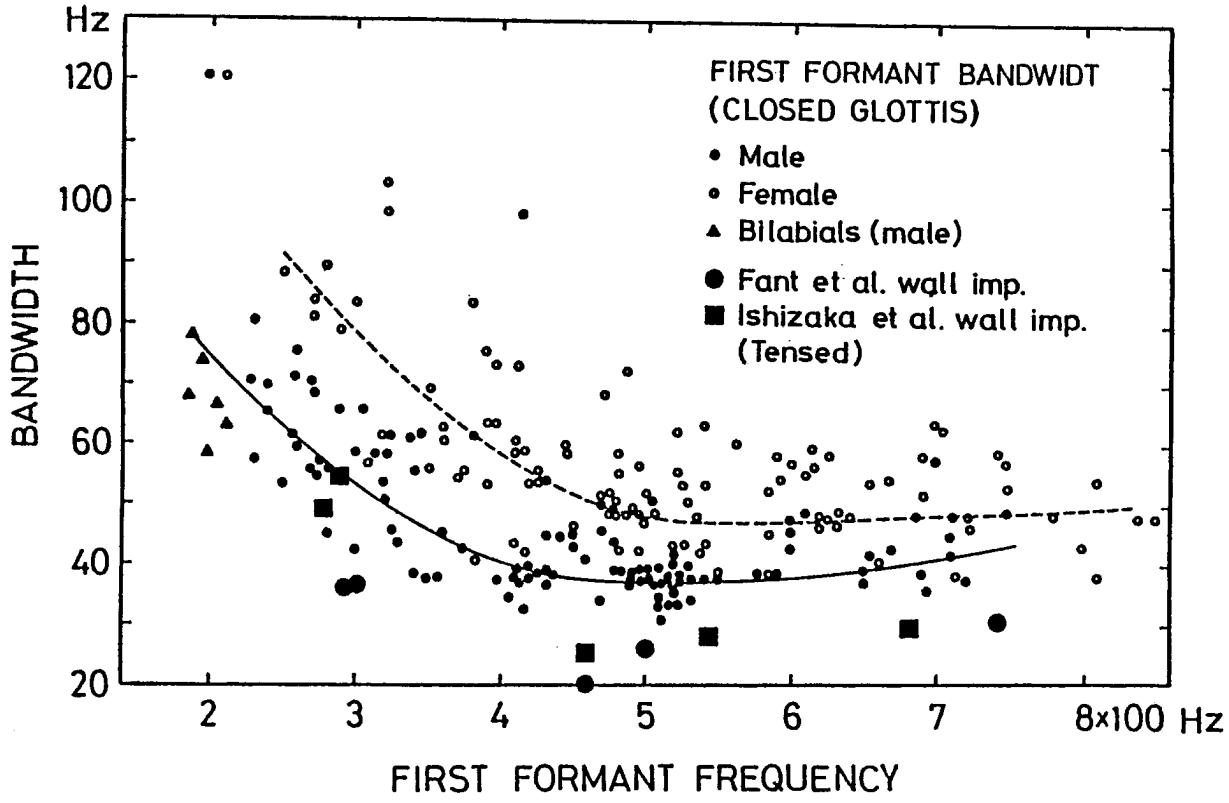


Figure III. Estimates and measurements of first formant bandwidth (B1)

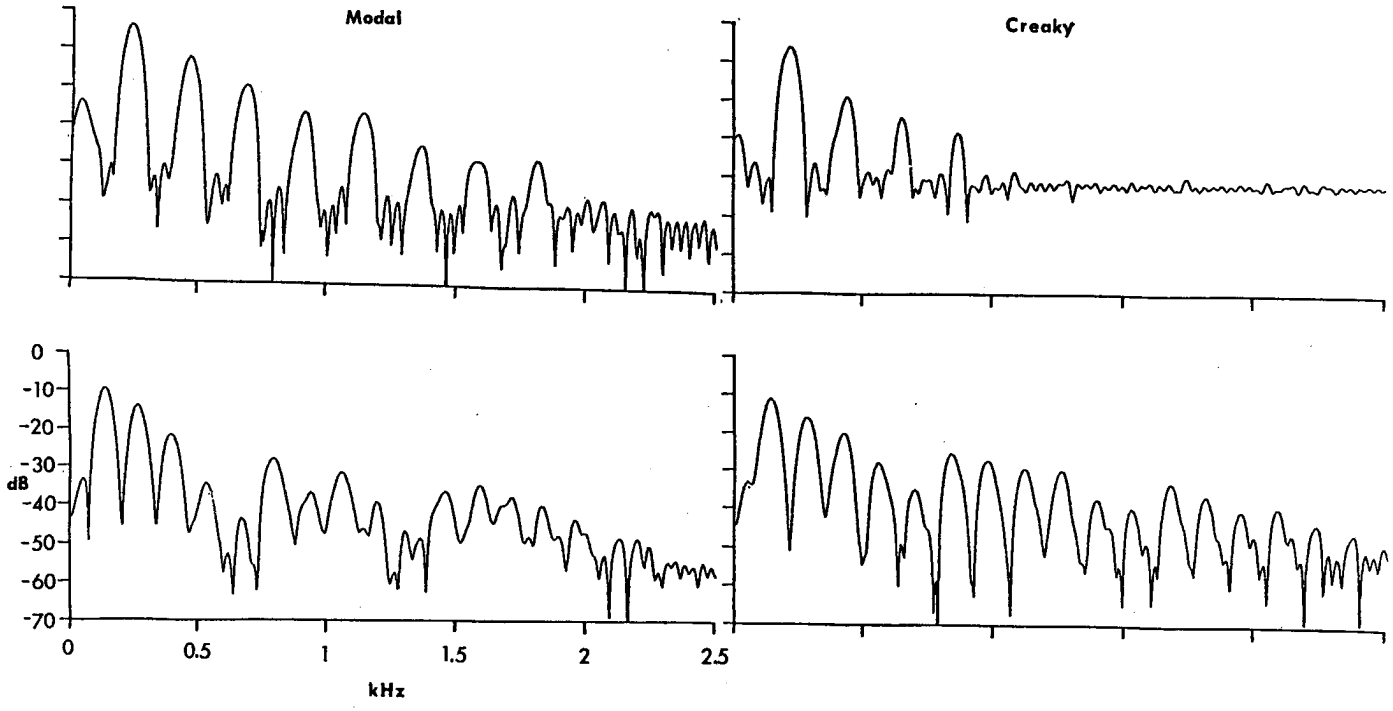


Figure IV. Representative spectra of inverse-filtered Burmese modal and creaky vowels

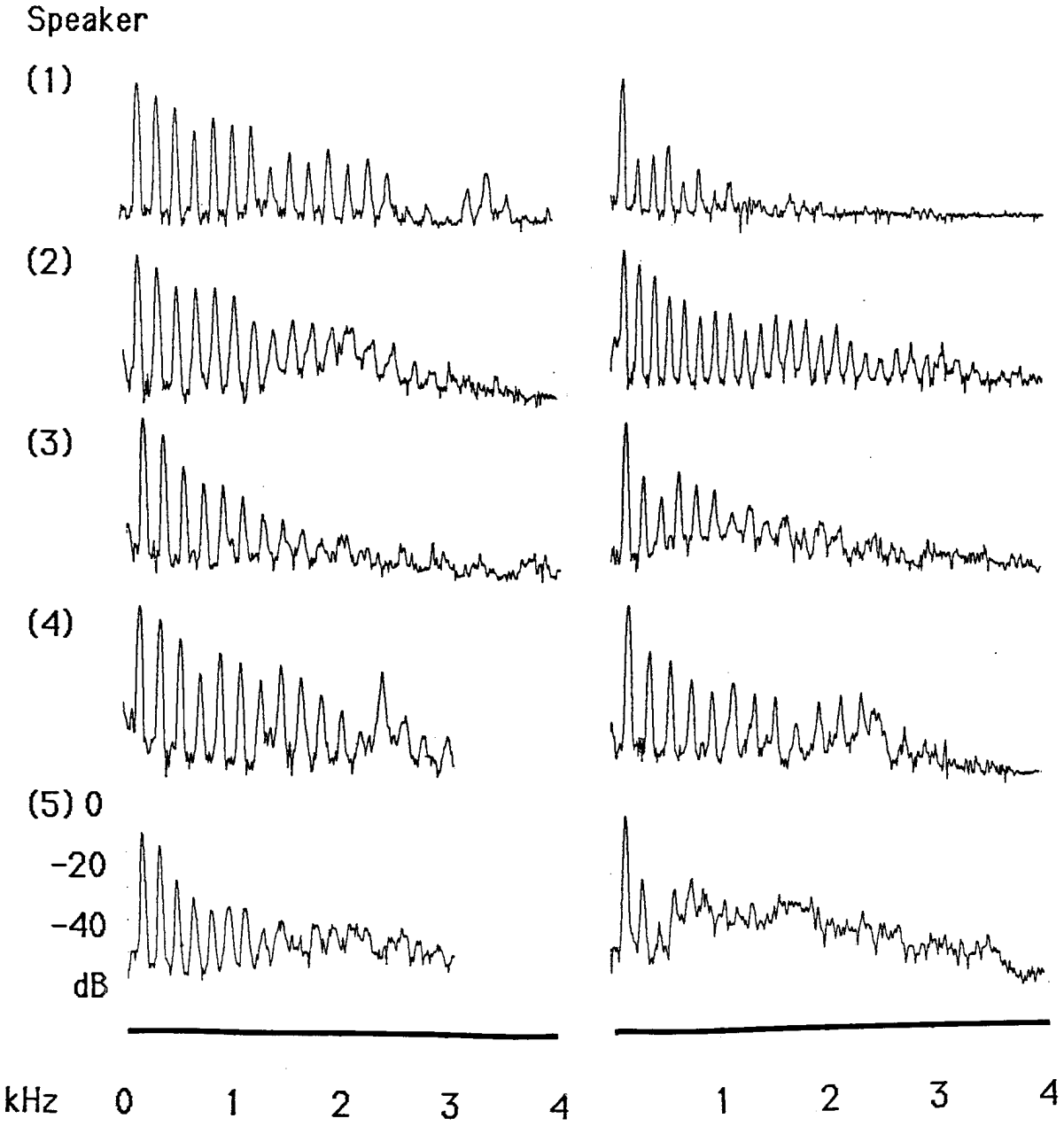


Figure V. Representative spectra of inverse-filtered !Xóõ modal and breathy vowels

Typical high-resolution spectra for a number of different !Xóõ breathy and modal vowels are shown in Figure V. At least some harmonics are clearly visible in each spectrum, but it is not always obvious how many there are.

The slopes of the regression line fitted to the peaks of the harmonics in these and other spectra were determined over several intervals. The statistics corroborated a point that was obvious from a visual inspection of the spectra: in many cases it is impossible to fit a statistically significant straight line (at the $p < 0.01$ level) through even the peaks of the first eight harmonics. In fact we found that we could reliably determine the slope of the glottal spectrum envelope in this region only 79% of the time for our 9 speakers of !Xóõ, 79% for our three Hmong speakers and 25% for our two speakers of Burmese. When we carried out regressions on regions excluding the fundamental, for example, from H5 to H8, or H2 to H9, the figures were even worse.

Simple models of glottal flow such as two-pole approximations (and linear networks in general), lead us to expect that the glottal flow spectrum is linear in log frequency. We find that this prediction is somewhat better verified: regression lines approximating the spectral envelope are statistically significant 72% of the time for !Xóõ, 83% for Hmong, and 100% for Burmese.

We then examined the distribution of significant slopes according to phonation type. For Burmese, with slightly creaky and modal phonation types, all tokens are well-characterized by linear spectral envelopes in log frequency. In Hmong, both breathy and modal flow spectra were linear about 80% of the time. In !Xóõ, all of the modal-voiced tokens are well-characterized by linear spectra envelopes, but only 44% of the breathy tokens are. We conclude that "spectral tilt" in log frequency can often be used to characterize modal phonation, but it cannot be used to characterize breathy phonation.

References

Fant, G. 1980. "Voice source dynamics", Speech Transmission Laboratory, Quarterly Progress and Status Report, 2-3/1980: 17-31.

__. 1983a. "Preliminaries to analysis of the human voice source", Speech Transmission Laboratory, Quarterly Progress and Status Report, 4/1982: 1-27.

__. 1983b. "The voice source - acoustic modeling", Speech Transmission Laboratory, Quarterly Progress and Status Report, 4/1982: 28-48.

Fant, G., L. Nord and P. Branderud. 1976. "A note on the vocal tract wall impedance", Speech Transmission Laboratory, Quarterly Progress and Status Report, 4/1976: 13-20.

Flanagan, J. L. 1958. "Some properties of the glottal sound source", Journal of Speech and Hearing Research vol 1. Republished in Acoustic Phonetics, ed. Fry, D. B. Cambridge University Press, 1976.

Fujimura, O. and J. Lindqvist. 1971. "Sweep-tone measurements of vocal-tract characteristics", Journal of the Acoustical Society of America, 49: 541-558.

Huffman, M. K. 1985. "Measures of phonation type in Hmong", UCLA Working Papers in Phonetics, 61: 1-25.

Ishizaka, K., J. French, and J. Flanagan. 1975. "Direct determination of vocal tract wall impedance", Trans. IEEE Acoustics, Speech, and Signal Processing, vol 23, no 4: 370-373.

Javkin, H., and I. Maddieson. 1983. "An inverse filtering analysis of Burmese creaky voice", UCLA Working Papers in Phonetics, 57: 115-125.

Javkin, H., N. Antoñanzas-Barroso and I. Maddieson. 1985. "Digital inverse filtering for linguistic research", UCLA Working Papers in Phonetics, 60: 87-100.

Rabiner, L. R., R. W. Schafer, and C. M. Rader. 1969. "The Chirp z-Transform Algorithm", IEEE Trans. Audio Electroacoust., vol AU-17, June 1969: 86-92.

Wakita, H. and G. Fant. 1978. "Towards a better vocal tract model", Speech Transmission Laboratory, Quarterly Progress and Status Report, 1/1978: 9-29.

Computer measures of breathy voice quality

Peter Ladefoged and Norma Antoñanzas-Barroso

[Paper presented at the 107th Meeting of the Acoustical Society of America]

Phoneticians often have to describe linguistic contrasts that involve different states of the glottis. These differences cannot be easily specified in quantifiable terms. Linguists often simply describe sounds as being "breathy" or "creaky" without being able to say how breathy or how creaky each sound is. The procedures described here are an attempt to remedy this situation.

Although these procedures were developed as aids in the description of languages, we hope that the results will be of interest not only to linguists, but also to speech pathologists and others who are studying abnormal laryngeal function. A voice quality that is considered pathological when used by speakers of one language may well be necessary for speakers of another, who need it to distinguish words with different meanings. A quantified description of a linguistic contrast may therefore be of great interest to speech pathologists. As Maddieson (1984) notes, if we cannot characterize a stable, phonemic, contrast used by all the speakers of a language, we have little chance of being able to make a valid, reliable, description of similar phonation types that are considered pathological when used by individual speakers of other languages.

The focus of this paper is the distinction between the normal mode of vibration commonly used in languages, which we will call modal voice ("chest voice" for some workers in vocal cord physiology, e.g. Sundberg, 1980), and breathy voice (or murmur) in which the vocal cords do not come together sharply, and may, indeed, not come together at all. Catford (1977) has described one form of breathy voice in which the vocal cords are simply "flapping in the breeze."

There are various acoustic consequences of these two different types of phonation. We may begin by considering those that are readily observable in the spectra of the two types of sounds. In modal voice the vocal cords are drawn together with great rapidity, and slam into each other so that the airflow through the glottis is abruptly terminated. This sharp discontinuity in airflow results in a vocal cord pulse that has a great deal of energy in the higher frequencies. The more lax vibrations of the vocal cords in breathy voice will result in a vocal cord pulse that has neither an abrupt beginning, nor (more importantly) an abrupt closure. The lack of a sharp discontinuity in the airflow through the glottis will result in a glottal pulse that has far less energy in the higher harmonics, and more energy at the fundamental frequency of vibration of the vocal cords.

Another kind of difference between breathy and modal voice arises from the greater rate of airfow that occurs during breathy voice. This greater flow produces more acoustic turbulence both at the vocal cords themselves and elsewhere in the vocal tract. There will thus be a greater random noise component in sounds with breathy voice. In short, in comparison with modal voice, breathy voice is both more breathy -- in the sense of having a more turbulent, irregular, waveform -- and less voiced -- in the sense of having a less well-defined set of harmonics. The algorithms we will describe below allow us to test which of these two characteristics predominates in the distinction between modal and breathy voice.

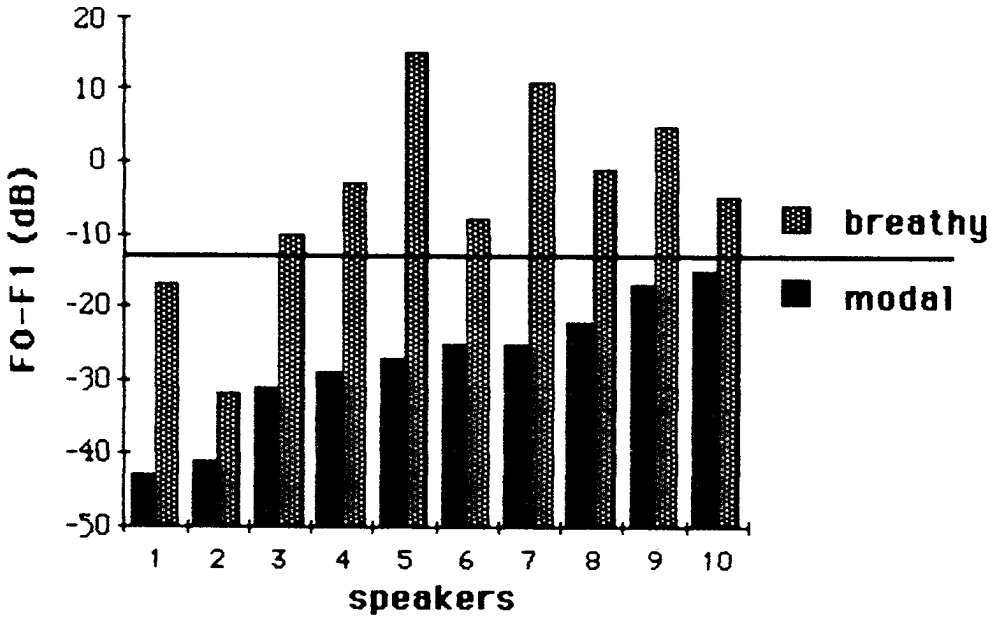


Figure 1. The difference between the amplitude of the fundamental (F0) and that of the harmonic with the highest amplitude in the first formant (F1) for contrasting modal and breathy vowels in !Xóõ.

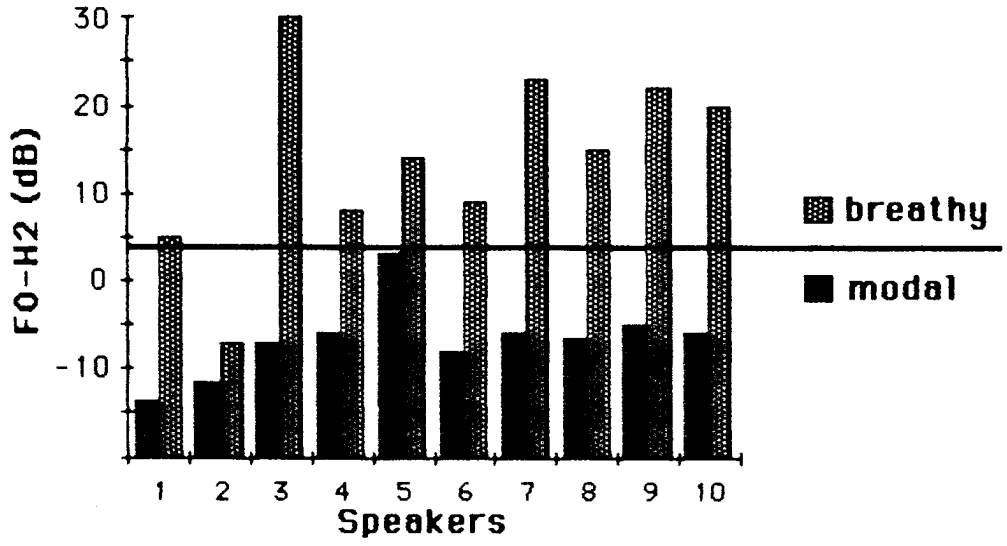


Figure 2. The difference between the amplitude of the fundamental (F0) and that of the second harmonic (H2) for contrasting modal and breathy vowels in !Xóõ.

Linguistic data

The first set of data used to test our algorithms was part of a larger set of data recorded by 10 speakers of !Xóõ, a Khoisan language spoken in the Kalahari desert. The samples illustrating modal voice were taken from the word //áa (a camelthorn tree) and those illustrating breathy voice from the word !ǂǂ (a slope). All the results to be reported are based on the analysis of the first 100 msec after the consonant of each of the two words as spoken by each of the 10 speakers.

Spectral measures

The first spectral measure we used in the current set of analyses is a development of that reported in Ladefoged (1983). The earlier procedure involved measuring the difference in dB between the fundamental frequency and the first formant, as observed in a single FFT spectrum in the middle of a vowel. The data to be reported in this paper used a variant of this procedure in which FFT spectra were produced at 10 msec intervals throughout the relevant section of the vowel. The mean of all these spectra was then calculated before measuring the same factor, the difference in dB between FØ and F1.

Figure 1 shows this measure for a contrasting pair of modal and breathy vowels for each of the 10 speakers of !Xóõ. It may be seen that for each speaker the shaded bar representing the breathy vowel is longer than the solid bar representing the modal vowel. In other words for the breathy vowel, for each speaker, there is a greater spectral tilt in the lower part of the spectrum, such that fundamental frequency has comparatively more energy than it does in the modal vowel. Notice also that the linguistic contrast between breathy voice and modal voice is a relative matter. What is modal for speaker 3 is breathy for speaker 2. As a result, it is not possible to draw a line separating all breathy vowels from all modal vowels. The line indicating the best separation that can be achieved in absolute, non-relative, terms leaves two linguistically breathy vowels among the modal vowels.

This spectral measure reflects to a considerable extent the difference in spectral tilt in the region below 1000 Hz. Its weakness is that is not reliable when comparing vowels with different qualities, and hence different formants that may differ in amplitude as well as frequency. An alternative way of measuring spectral tilt has been described by Bickley (1982) who measured the difference between the amplitude of the fundamental and that of the second harmonic. The results of applying this algorithm are very similar to those just reported, as may be seen by reference to Figure 2. The two classes of vowels are clearly distinguished by each speaker, with the shaded bar for the breathy vowels being longer, indicating that the fundamental frequency in breathy vowels has comparatively more energy than it does in the corresponding modal vowel. With these data, measurements of the difference between FØ and the second harmonic distinguish the two classes of vowels only as well as the measurements of the difference between FØ and F1. It is possible to draw a line separating breathy vowels from modal vowels in absolute terms, with only one breathy vowel being actually wrongly classified, but with another breathy and a modal vowel being only just differentiated.

Waveform measures

Measurements made directly on the waveform provide a different way of quantifying the degree of breathiness in a voice by determining how much of the acoustic energy is due to regular vibrations of the vocal cords and how much is of more random nature, due to turbulent airflow. If there is no turbulent flow and all the acoustic energy is due to the regular sharp pulses from the vocal cords, then one cycle of the sound wave will be exactly the same as the next. The greater the breathiness, the more there will be random variations due to the increased turbulence; hence there will be more difference between one cycle of the wave and the next.

A spectral domain approach of this kind was taken by Yumoto, Gould, and Baer (1982), who measured the harmonics to noise ratio as an index of the degree of hoarseness. However these investigators did not measure just the differences in the waveform that are due to the turbulent airflow. In addition to random components, there are other causes of variation between one cycle of a wave and the next. If the rate of vibration of the vocal cords is changing due to jitter, or to any variations in pitch, there will be variations in the length of the cycle. Simply comparing one period with the next will leave a residue that is due to the change in pitch or jitter. Accordingly, developing a measure of breathiness must use only part of a cycle and compare it with the corresponding part of the next cycle as indicated in Figure 3, which will shortly be explained in more detail.

The first step in our breathiness algorithm is to determine FØ using a cepstral analysis technique. This analysis is used to give a rough estimate of the length of the part of the waveform that will be compared with the corresponding part in the next cycle. To be conservative in the avoidance of FØ variations, the comparisons are of parts that correspond to only 75% of the estimated FØ. As is exemplified in Figure 3, the selected part of the waveform is compared with all other possible lengths within an interval beginning immediately after the current part and ending at twice the current estimated period from the end of the current part. The rms difference between each of these possible comparison lengths and the current part is calculated and used to determine the best match, which is, of course, the corresponding length in the next period. The rms difference between these two best-matched components is noted, and the comparison process is then repeated with another length which starts one msec further on. Each set of comparisons is made independently of the previous set, rather than relying on the previous set of comparisons to give a measure of the period, and hence an indication of the interval between the current length and the best matching comparison length. As we have noted, the breathiness measure should not be affected by jitter or other variations in the period length.

We have found that we get a more stable measure if there is considerable overlap in the lengths of waveform that form the bases for comparison. To give some feeling for the numbers involved, the wave shown in Figure 3 has a fundamental frequency of 100 Hz, and was sampled at 10,000 sps. Each cycle consists of 100 points, so the base lengths each contain 75% of this, namely 75 points. After finding the best match for a given set of 75 points, the start of the base length is moved over 10 points and the algorithm finds the best match for the next set of 75 points. Within 100 msec -- a very short length of vowel -- we will be able to make almost 100 comparisons of this kind. The mean rms difference between each pair of best matching lengths is then calculated, and normalized by dividing it by the mean rms amplitude of that particular section of the vowel.

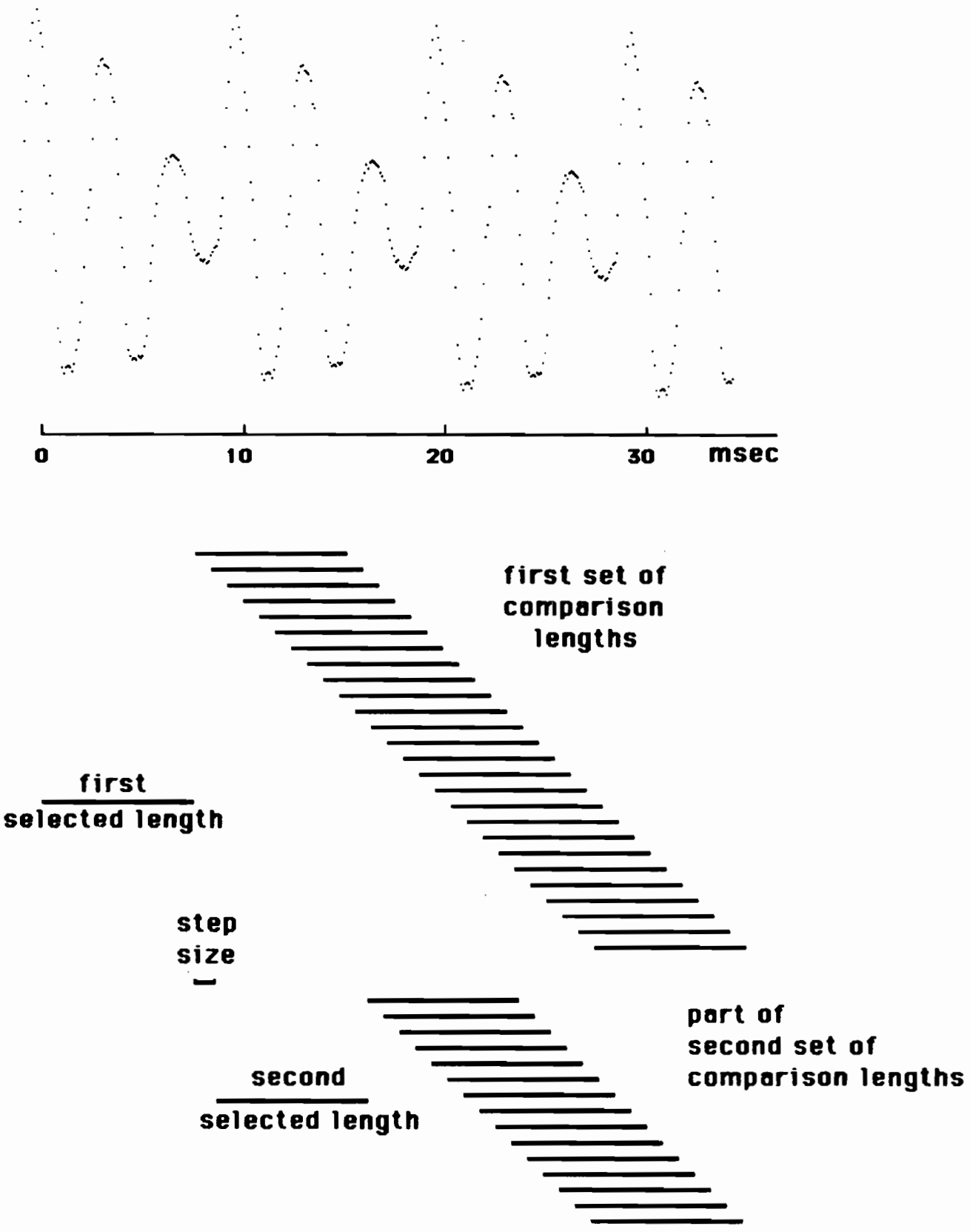


Figure 3. A technique for measuring the degree of randomness in parts of adjacent cycles.

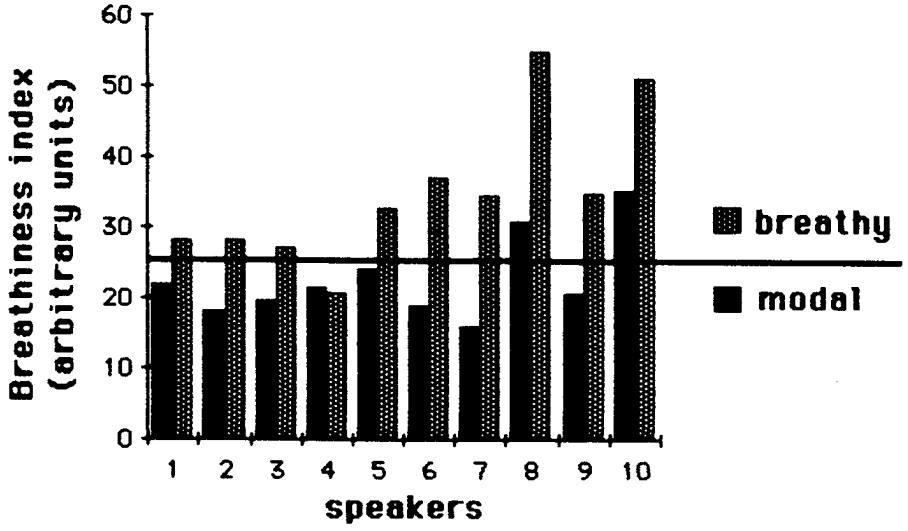


Figure 4. The breathiness index for contrasting modal and breathy vowels in !Xóõ.

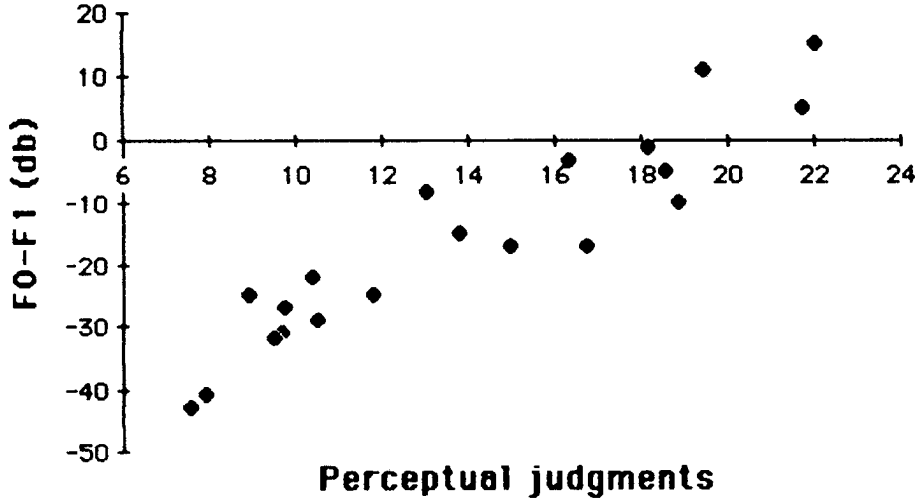


Figure 5. The relation between judgments of perceived breathiness and the difference in amplitude between the fundamental and the first formant.

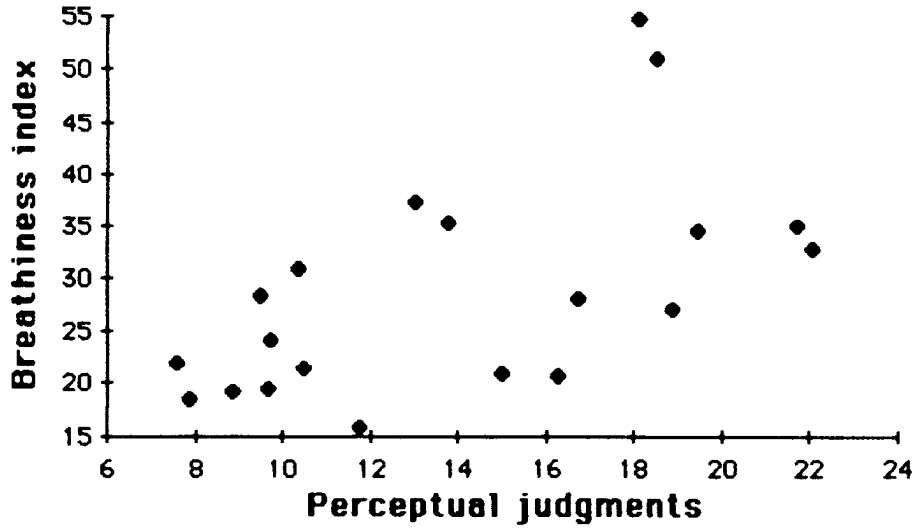


Figure 6. The relation between judgments of perceived breathiness and the breathiness index described in the text.

Figure 4 shows the results of this analysis for the data discussed previously. In every case but one, the shaded bar is longer, indicating that there is more randomness in the breathy vowel. Again, it is not possible to use this measure in an absolute way to separate all breathy vowels from all modal vowels. The best line for this purpose wrongly categorizes one breathy vowel and two modal vowels.

It is interesting to compare the breathiness measures in Figure 4 with the spectral measures in Figures 1 and 2. As can be seen from Figure 4, speakers 1 and 4 make little or no use of the property measured by the breathiness index, but make considerable use of the spectral differences illustrated in the earlier figures. Conversely, speaker 5 makes more use of the breathiness property than of the spectral properties. In other words, some speakers make more use of turbulence for separating linguistically breathy from linguistically modal vowels, whereas others make more use of variations in the spectral tilt produced by a different mode of vibration of the vocal cords. There is virtually no correlation between the measures. Some speakers use both turbulence and spectral tilt, others emphasize one more than the other.

In order to see which of the measures correlated better with differences in voice quality as perceived by American English speakers, we conducted a listening test in which 10 listeners associated with the UCLA Phonetics Lab, and therefore familiar with the terminology, rated the 20 vowels in terms of their perceived breathiness. The vowels were presented in a random order. Listeners neither knew, nor were asked to assess, whether the vowels were linguistically modal or breathy. They made judgments of the degree of breathiness of each vowel without considering its contrastive value.

Figure 5 shows the relation betweeen the FØ - F1 spectral measure and the perceptual judgments. The correlation is high (r=.93), so it seems that the listeners certainly use a factor corresponding to the spectral tilt as measured in this way in making their judgments of breathiness.

On the other hand, as Figure 6 shows, there is a considerably lower correlation between the degree of randomness in the waveform and the perceptual judgments (r=.57). This is somewhat surprising in that it is a good measure for separating, for each individual, the linguistically breathy vowels from the contrasting modal vowels. But as far as American listeners are concerned, the mode of vibration of the vocal cords as reflected in the spectral tilt seems perceptually more important than the degree of turbulence in the waveform. Although one might expect the turbulence to be perceptually dominant, as is suggested by the term "breathiness" it is really the stiffness or slackness of the vocal cord vibrations that matters for our laboratory listeners. We do not know whether this is so or not for the !Xóõ, who are more sophisticated users of breathy voice. We have not yet been able to do any perceptual testing with any of our !Xóõ speakers in the Kalahari desert.

References

Bickley, Corine. 1982. Acoustic analysis and perception of breathy vowels. Working Papers, Speech Communication Group, MIT 1: 71-81.

Catford J.C. 1977. Fundamental Issues in Phonetics. Indiana University Press, Bloomington.

Dart, Sarah. 1984. Testing an aerodynamic model with measured data from Korean. UCLA Working Papers in Phonetics 59: 1-17.

Ladefoged, Peter. 1983. The linguistic use of different phonation types. In Diane Bless and James Abbs (eds) Vocal Fold Physiology; Contemporary Research and Clinical Issues: 351-360. College Hill Press, San Diego.

Sundberg, Johan. 1980. RÖstlÄra: Facta om rösten i tal och sång. Proprius, Stockholm.

Maddieson, Ian. 1984. Investigating the voice quality dimension in Western Nilotic vowel harmony. Journal of the Acoustical Society of America 75.Supp. 1: S9.

Yumoto, Eiji, Gould, Wilbur J., and Baer, Thomas. 1982. Harmonics-to-noise ratio as an index of the degree of hoarseness. Journal of the Acoustical Society of America 71.6: 1544-50.

Cross-linguistic issues in the relationship between semi-vowels and vowels

Ian Maddieson and Karen Emmorey

Introduction

There are known to be reliable cross-language differences between similar vowels in comparable environments (see especially Disner 1983). Differences of the same kind are also found between consonants (e.g. Nartey 1982, Ladefoged and Bhaskararao 1984). These kinds of differences imply that languages may have different acoustic or articulatory "targets" for what is broadly the same segment type. As Ladefoged (1980) has stressed, such differences must be taken into account in any adequate specification of the phonetic properties of languages. We hypothesize that the same kind of cross-language differences are to be found among semi-vowels, for example, that different languages have different targets for the labial-velar and palatal semi-vowels /w/ and /j/. Furthermore, we hypothesize that cross-language differences in semi-vowels will correlate with cross-language differences in the cognate vowels /u/ and /i/. That is, /w/ and /u/ will share some language-specific similarities, as will /j/ and /i/. Experiments addressing these hypotheses, as well as the question of how the semi-vowels differ from the cognate vowels, have been conducted. These experiments produced significant evidence of differences between languages with respect to coarticulation between semi-vowel and vowel; a result which has important methodological as well as linguistic implications.

Methods and Materials

Speakers of several different languages were recorded saying the nonsense sequences /iwi, awa, uwu; iji, aja, uju/ embedded in carrier phrases in their respective native languages. The carrier phrases and speakers are the same ones used by Nartey (1982) in his cross-language studies of fricatives.

The principal data set was obtained from five male speakers of Amharic, five male speakers of Yoruba, and six male speakers of Zuni, who each read the target words three times as part of a longer list of items. The words contain sequences which are phonotactically permissible in these languages. Some readings of particular items were eliminated due to inconsistencies of speech rate, and others were dropped since they were misread with unlike vowels either side of the semi-vowel. The speakers of Zuni seemed to have particular difficulty with the sequence /uju/ (often read as /uja/). The resulting numbers of tokens used in data analysis are given in table 1 below.

Table 1. Languages, number of speakers and tokens in main data set.

	awa	iwi	uwu	aja	iji	uju
Amharic (5)	15	15	15	15	15	14
Yoruba (5)	15	15	15	15	15	15
Zuni (6)	19	18	13	14	16	8

Limited data, consisting of subsets of these sequences, was also obtained from eight speakers of Navaho (five males and three females) and five speakers of Papago (two males and three females). One female Amharic speaker and three female

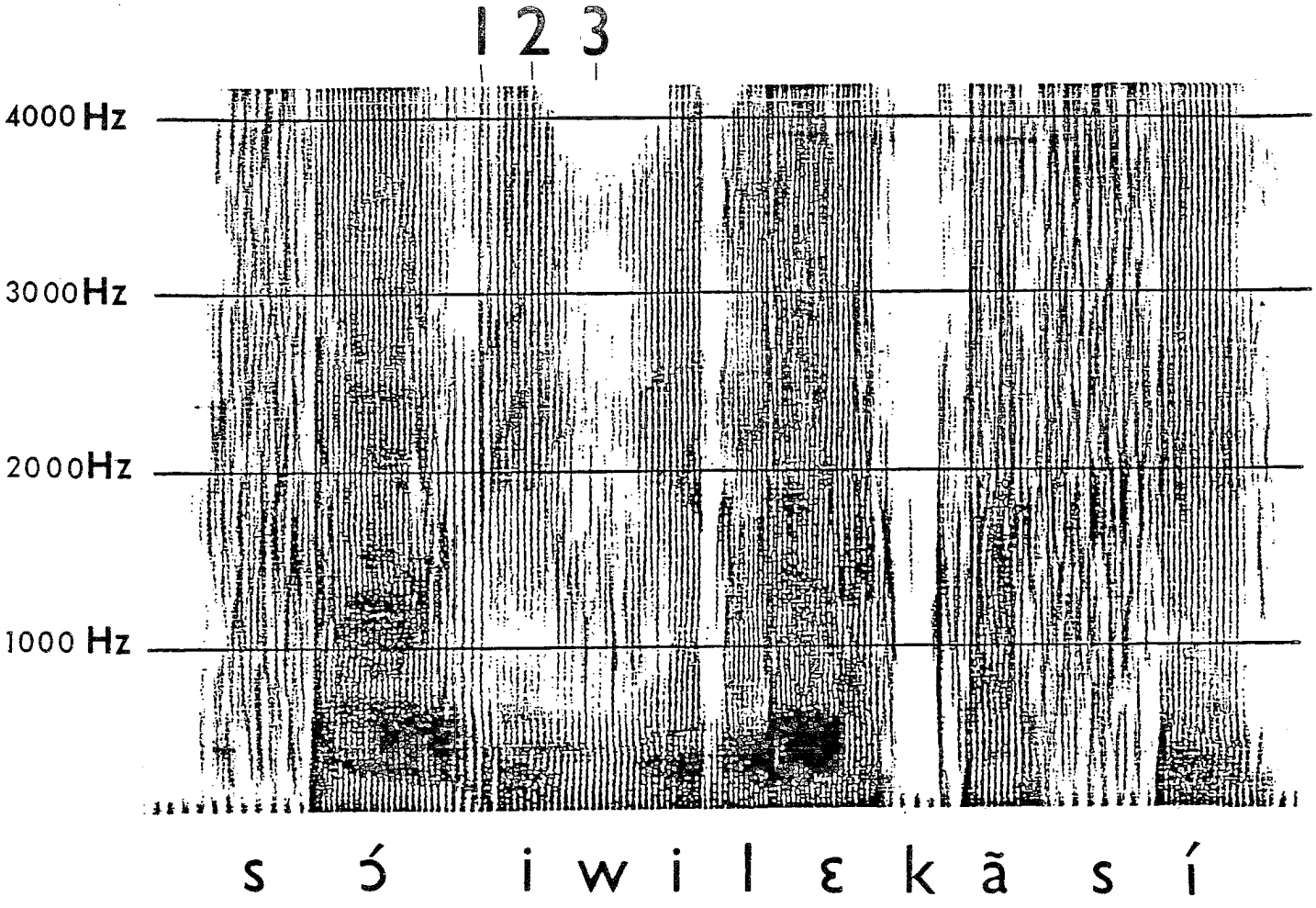


Figure 1. Sample spectrogram illustrating measurement points. The spectrogram is of /iwi/ in the Yoruba carrier phrase /sɔ́ ___ lɛkã sí/. In this token vowel formants were not measured at vowel onset, point 1, but at the later point 2 where influence of the preceding vowel is least evident. Measurements of semi-vowel formants were made at point of minimum amplitude, point 3.

Zuni speakers also provided data. A relatively complete summary of the data is included as an appendix at the end of this paper.

Frequencies of F1, F2 and F3 in both semivowels and vowels were measured from spectrograms, such as that in figure 1, by both authors and a consensus reached. The formants were measured at a point judged to be central for the semi-vowels, and at the onset of the initial vowel in the VCV sequence for the vowels except where values were more extreme at a later point in the vowel, i.e. where there seemed to be coarticulatory influence on the vowel onset from the carrier phrase, as there is in the sample spectrogram in figure 1. Since F3 is often unmeasurable for /w/ and /u/ we will mostly confine our comments to F1 and F2.

Results

a) Mean formant values of semi-vowels and cognate vowels

The overall means of F1 and F2 in the semi-vowels in the principal data set are given in table 2. The values shown are the means of individual speaker means. Analysis of variance showed a significant main effect of language for both F1 and F2 of the semi-vowels in this data. Post hoc analyses of the mean formant frequencies indicate that this is largely because Zuni is different from the other two languages. Pairs of means which differ significantly from each other are indicated in the table by a difference in the number of asterisks following them. Means which are not involved in any significant contrasts are not marked by any asterisks. The significance criterion was at least .01, calculated by Tukey's studentized range test, as implemented in the SAS statistics package. This test for the multiple comparison of means controls for the type 1 experimentwise error rate, that is, for the increasing probability that some pair of means will by chance be significantly different when an increasing number of independent comparisons is made. However, when a large number of comparisons is involved, Tukey's test may lead to false acceptance of the null hypothesis, i.e. significant differences may be overlooked. This caution should be borne in mind in the course of the presentation of these results.

Table 2. Mean values of F1 and F2 of /j/ and /w/, and F3 of /j/, by language, pooled across vowel environments.

	/j/			/w/	
	F1	F2	F3	F1	F2
Amharic	236 *	2118 *	2891 *	288	675 *
Yoruba	239 *	2031	2719	284	639 *
Zuni	296 **	1909 **	2601 **	314	849 **

As Table 2 indicates, there are significant differences distinguishing Zuni from both the other two languages with respect to F1 of /j/ and F2 of /w/, and Zuni from Amharic with respect to F2 and F3 of /j/.

The mean formant values in the corresponding vowels are given in table 3. The pattern of significant differences in the semi-vowel formants is similar to the pattern of differences in formants of the respective cognate vowels, apart from F1 of /i/ which shows no significant cross-language differences. Thus, in Zuni, F2 and F3 are significantly lower in both /j/ and /i/ than in Amharic and F2 is significantly higher in both /w/ and /u/ than it is in the other two languages. F3 of /i/ is significantly lower in Yoruba than in Amharic, whereas there is only a nonsignificant trend for F3 in /j/ to be lower in Yoruba.

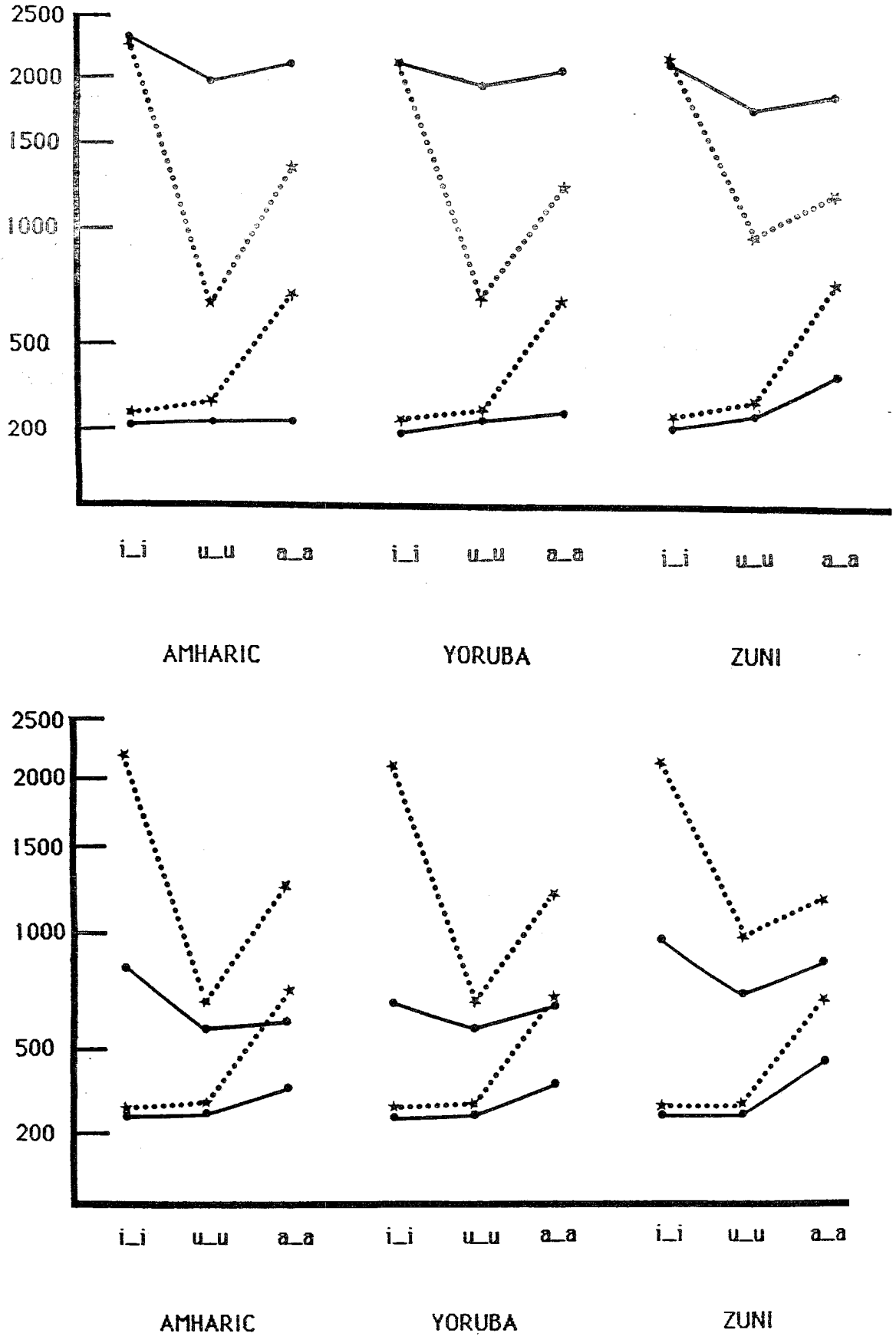


Figure 2. Mean positions of F1 and F2 of /j/ (upper chart) and /w/ (lower chart) in the three different vowel contexts, and of F1 and F2 of the vowel preceding the semi-vowel. Semi-vowel formants are shown by filled circles -- those points belonging to the same language are joined by a solid line. Vowel formants are indicated by a star and points belonging to the same language are joined by a dotted line.

Table 3. Mean values of F1 and F2 of /i/ and /u/, and F3 of /i/, by language, pooled across semi-vowel environments

	/i/			/u/	
	F1	F2	F3	F1	F2
Amharic	274	2213 *	2989*	308	687 *
Yoruba	264	2094	2731**	282	687 *
Zuni	268	2095 **	2487**	295	990 **

b) Coarticulation between semi-vowels and vowels

Showing that languages differ in mean formant values for semi-vowels does not necessarily demonstrate that they have different idealized target values for these segments. Especially, a language which tolerates a high degree of coarticulatory effect of the surrounding vowels on /j/ and /w/ may show a different mean position for these segments than one which permits very limited coarticulation, or which makes coarticulatory adjustments in the vowels rather than in the semi-vowels. (Note that the same point could be made with respect to the vowel data discussed by Disner, or the fricative data of Nartey. Nartey shows that the languages he studied have some differences in the degree of coarticulation between vowel and fricative, but does not discuss the implication of this for the interpretation of the mean cross-language differences that he observed.)

In the analysis of variance there is a significant interaction between language and vowel context for all three measured formants in the semi-vowels. Insofar as the semi-vowel formants seem to be varying across vowel contexts in ways that bring their values closer to the formant values of the adjoining vowels we may assume that this is because there are differences in the coarticulatory relationships between semi-vowel and vowel for these languages. These differences can be seen graphically for F1 and F2 in figure 2 by comparing mean semi-vowel frequencies in each language across the three vowel contexts. Absence of any coarticulatory effect would be indicated by a horizontal solid line connecting semi-vowel formant frequencies in the three vowel contexts. This is approximately what is seen for the first formant of /j/ for Amharic. The more the line connecting semi-vowel formants deviates from the horizontal toward the formant positions of the vowels, the greater the coarticulation with the surrounding vowels. Formant frequency values for the vowel preceding the semi-vowel are connected by dotted lines in the figure.

Table 4 below shows where significant differences were found when comparing semi-vowel formant values across the different vowel environments within each language. This table indicates that, for F1, coarticulation largely concerns how much effect the /a_a/ context has, since all cases where the difference between contexts is significant involve /a_a/. As for F2, the situation is somewhat more complex. Although F2 is always lowest in the /u_u/ context (for both /j/ and /w/), it is not normally significantly lower here than in the /a_a/ context. There is usually a significant difference between /i_i/ and /u_u/ contexts, and sometimes one between /i_i/ and /a_a/ contexts. A case could be made for viewing this pattern as resulting from an upward assimilation of the F2 of the semi-vowels to the high F2 of /i/. This view is obviously more plausible for /w/ than for /j/. Perhaps the least contentious view is to suggest that both /i_i/ and /a_a/ raise the F2 of /w/, and that both /u_u/ and /a_a/ lower the F2 of /j/. F3 in /j/ would also be lowered by both /u_u/ and /a_a/. But this may be to assume the truth of the claim of similarity of /i/ to /j/ and of /u/ to /w/ which is one of the issues that the present experiment was designed to address.

Table 4. Significant differences between semi-vowel formants in different vowel contexts, by language.

	F1			F2		
	/i_i/ vs /u_u/	/i_i/ vs /a_a/	/u_u/ vs /a_a/	/i_i/ vs /u_u/	/i_i/ vs /a_a/	/u_u/ vs /a_a/
for both semi-vowels pooled						
Amharic	no	no	no	yes	yes	no
Yoruba	no	yes	yes	yes	no	no
Zuni	no	yes	yes	yes	yes	yes
for /w/ alone						
Amharic	no	no	no	yes	yes	no
Yoruba	no	yes	yes	no	no	no
Zuni	no	yes	yes	yes	no	no
for /j/ alone						
Amharic	no	no	no	yes	no	no
Yoruba	no	yes	no	no	no	no
Zuni	no	yes	yes	yes	yes	no
					F3	
			Amharic	yes	no	yes
			Yoruba	no	no	no
			Zuni	no	no	no

Both table 4 and figure 2 indicate that, of the three languages, Zuni is the one which shows the greatest effect of vowel context on the semi-vowels. It could be that the same proportional degree of adaptation to vowel context is made in each language but the vowels themselves differ sufficiently to account for the semi-vowel differences. However this is not the case. The F1 values for each vowel type are very similar across the three languages and no vowel shows a significant F1 difference across the languages. Although there are significant differences in vowel F2's, the range of F2 variation in Zuni vowels is smaller, not greater, than in the other two languages, hence greater vowel contrast does not account for the greater coarticulation. (Vowel F2 values are quite similar for Amharic and Yoruba and show no significant differences in the same semi-vowel context.) We must conclude that Zuni simply permits more coarticulation. There is even some evidence that the situation in Zuni is one in which there is greater mutual effect between vowel and semi-vowel. Zuni is the only language of the three which shows a significant difference for mean F2 of the vowels across the two different semivowel contexts.

We are now in a position to re-examine the overall cross-language formant differences reported in table 2. From figure 2 and table 4, we can deduce that the higher overall mean F1 of /j/ in Zuni results to a large extent from the much greater raising of F1 in the /a_a/ context in that language. When each vowel context is considered separately, the F1 of /j/ is significantly higher in Zuni than in Amharic or Yoruba in the /a_a/ context but not in the other vowel contexts. As for the lower F2 of /j/ in Zuni, compared with Amharic, it is lower

in all three vowel contexts -- although the difference only reaches significance by our criterion in the /a_a/ context. In particular, F2 of /j/ is lower in the /u_u/ context in Zuni, even though the /u/ vowel itself has a higher F2 in Zuni than in Amharic. The lower mean F2 of /j/ in Zuni is thus unlikely to be the result of a simple coarticulatory effect. Somewhat similarly, F3 in /j/ is higher in each vowel context in Amharic than in Zuni, although the difference is statistically significant only in the /a_a/ context. Since F3 is always lower in the adjacent vowels than in /j/ in Amharic, the high mean F3 cannot be explained as due to a raising effect of any particular vowel context. Although Amharic does show the greatest effect of vowel context on F3 of /j/, this seems to be independent of its characteristically high position. Thus the different mean positions of F2 and F3 of /j/ in Amharic and Zuni may be interpreted as reflecting different targets for this segment in these languages. (It is conceivable that more complex coarticulatory patterns could explain these overall differences.)

The significantly higher overall F2 of /w/ in Zuni emerges as a result of it being higher in each of the three vowel contexts than in the other two languages. This result is therefore not the consequence of the effect of any particular vowel context, although it could be regarded as the result of a greater overall tendency for coarticulation in Zuni, since the F2 of the vowels is in each case higher than the F2 of /w/ in that context.

Of the differences reported in table 2 above, we may therefore tentatively identify the differences in F2 and F3 as likely to be due to language-specific target values for the semi-vowels. Note that these are the differences which are matched by differences between the languages in the cognate vowels. On the other hand, the difference in F1 of /j/ results from cross-language differences in coarticulation, and is not matched by a corresponding F1 difference in the cognate vowel.

c) Differences between semi-vowels and cognate vowels

There has long been disagreement among both phoneticians and phonologists over whether segments like /j/ and /w/ are actually consonantal in their articulation. Many have argued that the semi-vowels /j/ and /w/ are identical to the vowels /i/ and /u/ apart from their role in the syllable. If this is correct, then /j/ would be expected to have the same formant frequencies as /i/, and /w/ the same formant frequencies as /u/. The differences between vowel and semi-vowel are then regarded as being in amplitude or 'prominence' (Jones 1956), syllabic function (Pike 1943), and timing, e.g. in different rates of transition of the formants (Liberman et al. 1956, O'Connor et al. 1957, Catford 1977). Others have claimed that /j/ and /w/ are made with a more constricted vocal tract than /i/ and /u/. This was apparently the view of Henry Sweet (1877, 1908). It is adopted by many authors in the speech and audiology field (e.g. Zemlin 1981) and incorporated in the phonetic feature framework of Chomsky and Halle (1978) as well as assumed in recent acoustic modeling studies by Bickley and Stevens (1985). If this view is correct, then a difference between the formants of the semi-vowels and the vowels would be anticipated. Specifically, a lower F1 for /j/ and /w/ than for /i/ and /u/ would be expected to result from a closer articulation between the tongue and the roof of the mouth for the semi-vowels, and, in addition a lower F2 for [w] than for [u] would be anticipated if there is also a more constricted lip aperture in [w]. Expectations of the difference in F2 between /j/ and /i/ will vary depending on how far forward the constriction is taken to be (Fant 1960, Ladefoged and Bladon 1982), but a higher F3 might be anticipated in /j/ than /i/.

Because of the effects of the coarticulation discussed above and its language-specific nature, we cannot address the question of the difference between vowel and semi-vowel using our complete data set, even though the vocalic environments of the semi-vowels are matched across the three languages. We can effectively only examine the difference between /i/ and /j/ in the word /iji/ and between /u/ and /w/ in the word /uwu/. Table 5 shows where significant differences between these semi-vowels and vowels were found in the data as a whole. In this case, significance means that the mean within-token differences of the semi-vowel and vowel formants is significantly different from zero, to a level of at least .01.

Table 5. Significant differences between vowel and semivowel formants

	F1	F2	F3
between /i/ and /j/ in /iji/	yes	no	yes
between /u/ and /w/ in /uwu/	yes	yes	---

As predicted by the "consonantal" hypothesis of semi-vowel production, the F1 of the semi-vowel is lower than that of the cognate vowel it occurs with for both /j/ and /w/. In addition, F2 is lower in /w/ than in /u/ and F3 is higher in /j/ than /i/. In other words, we have evidence that /j/ and /w/ are produced with a narrower constriction than /i/ and /u/ and that /w/ is produced perhaps also with greater lip rounding than /u/.

d) dynamic patterns for semi-vowels

Apart from being characterized by particular target values, it is possible that semi-vowels are also characterized by a particular dynamic pattern of formant movements. Lehiste (1964), in particular, suggests that this is true for English. She writes:

> "The transition [of F3] from an initial /j/ to the following vowel consists not of a smooth movement from target to target, but rather a downward movement of F3 to a specific value (near 2000 Hz) and then a rapid movement to the target of F3 for the particular vowel, if that target is different from 2000 Hz."

and

> "the phonetic character of /w/ is not expressed completely by describing the formant structure of the steady state. The other essential cue is the gliding movement of the formants....manifested chiefly by a rise of the second formant. In the word woo, for example, the first formant remains almost steady, but the second formant rises by approximately 250 Hz at the end of /w/ and then falls again to arrive at the target position of /u/."

The present data set, including the supplementary data from female speakers, was examined qualitatively for the kinds of formant movements described by Lehiste. Movement of F3 in the transition from /j/ to a following vowel was examined in the /iji/ sequence, where, because of the high F3 of the vowel, a downward movement of F3 before the vowel can be expected to be clearly differentiable from a simple smooth transition. In no case is there a fall of F3 to a value close to 2000 Hz. The predominant pattern in all three languages in the main data set is for a smooth movement of F3 from the value in the semi-vowel down to a slightly lower value in the vowel. There is one token in Amharic and one in Zuni in which there is a small but definite dip in the F3 frequency at the transition from /j/ to /i/. But this is only one of a number of unusual patterns: one Amharic speaker has a peak of F3 at the /j/ to /i/ transition in two of his

three tokens of /iji/, one Zuni token has a flat F3, one Yoruba token shows a rising F3.

A rising-falling F2 pattern for the transition from /w/ to /u/, similar to that suggested for English by Lehiste, is not found in a single token in the main data set (nor in any tokens from the Papago speakers in the supplementary data set). The great majority of tokens show a smooth rise of F2 from /w/ to a slightly higher position at the transition to the following /u/; often F2 continues to rise during the vowel. A few tokens show no rise at the transition or in the following vowel. In general, therefore, no special dynamic patterns of formant movement are seen for the semi-vowels in the current data. Perhaps they only seem to be present in English because the English /i/ and /u/ vowels typically have some narrow diphthongal movement, and this movement has been misassociated with the semi-vowels.

Discussion

The present paper has demonstrated cross-language differences between segments of an additional class. However, cross-language differences between palatal and labial-velar semi-vowels are paralleled by differences in the vowels to which they are most similar. This indicates that there is a tendency for the phonetic individuality of languages to be subject to limitation, at least in some small ways. Although attempts to capture this limitation by concepts such as "base of articulation" have little success in general (see Disner 1983 for discussion), the present result lends some support to an idea of this overall type.

Cross-language differences also extend to the pattern of coarticulation between semi-vowels and the adjacent vowels. Although semi-vowels are consonants we assume that they are produced using, in Öhman's (1966) terms, the same control channel as the vowels. Öhman showed that apical and velar consonants in VCV sequences with unlike vowels either side are affected by vowel context more in Swedish and English than in Russian. He suggests that this is due to differences in the phonological systems of the languages: the Russian consonants have little freedom to coarticulate with the surrounding vowels because they are distinctively palatalized or velarized, i.e. they must be specified for values for vowel-like features in the vowel control channel. The coarticulatory differences between Amharic, Yoruba and Zuni demonstrated in the present paper cannot be explained by a similar line of reasoning both because there do not seem to be appropriate phonological differences between these languages and because a single control channel is involved. Cross-language coarticulatory differences can occur simply because one language permits adjacent segments to affect, or overlap with, each other more than another.

This last point needs to be carefully considered as a factor in experimental design in studies which involve cross-language comparisons. A comparison of segments in only one context cannot distinguish between differences in underlying target positions for those segments and differences in coarticulatory propensity.

Acknowledgments

This work was supported by grants from the National Science Foundation to the UCLA Phonetics Laboratory. Data collection was also supported in part by funds provided by the UCLA American Indian Studies Center. We acknowledge with gratitude Jonas Nartey's assistance in obtaining recordings for this study.

References

Bickley, C.A. and Stevens, K.N. 1985. Modeling study of influences of vocal-tract configurations on glottal behaviour. Paper presented at the 109th Meeting of the Acoustical Society of America, Austin, Texas, April, 1985.

Catford, J.C. 1977. Fundamental Problems in Phonetics. Indiana University Press, Bloomington.

Chomsky, N. and Halle, M. 1978. The Sound Pattern of English. Harper and Row, New York.

Disner, S.F. 1983. Vowel Quality: The Relation Between Universal and Language-Specific Factors (UCLA Working Papers in Phonetics 58). Phonetics Laboratory, University of California, Los Angeles.

Fant, G. 1960. The Acoustic Theory of Speech Production. Mouton, The Hague.

Jones, D. 1956. An Outline of English Phonetics. Heffer, Cambridge.

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

Ladefoged, P. and Bladon, A. 1982. Attempts by human speakers to reproduce Fant's nomograms. UCLA Working Papers in Phonetics 54: 40-56 (Postscript p. 109).

Ladefoged, P. and Bhaskararao, P. 1984. Non-quantal aspects of consonant production: a study of retroflex consonants. Journal of Phonetics 11: 291-302.

Lehiste, I. 1964. Acoustical Characteristics of Selected English Consonants. Indiana University, Bloomington.

Liberman, A.M., Delattre, P.C., Gerstman, L.J., and Cooper, F.S. 1956. Tempo of frequency change as a cue for distinguishing classes of speech sounds. Journal of Experimental Psychology 52: 127-137.

Nartey, J.N.A. 1982. On Fricative Phones and Phonemes (UCLA Working Papers in Phonetics 55). Phonetics Laboratory, University of California, Los Angeles.

O'Connor, J.D., Gerstman, L.J., Liberman, A.M., Delattre, P.C., and F.S. Cooper. 1957. Acoustic cues for the perception of initial /w, j, r, l/ in English. Word 13: 24-43.

Öhman, S.E.G. 1966. Coarticulation in VCV utterances: spectrographic measurements. Journal of the Acoustical Society of America 39: 151-168.

Pike, K. 1943. Phonetics. University of Michigan Press, Ann Arbor.

Sweet, H. 1877. A Handbook of Phonetics. Clarendon Press, Oxford.

Sweet, H. 1908. The Sounds of English. Clarendon Press, Oxford.

Zemlin, W.R. 1981. Speech and Hearing Science (2nd ed). Prentice-Hall, Englewood Cliffs.

Appendix

The carrier phrases used for the respective languages were as follows (from Nartey 1982).

Amharic /antə ___ bɛl/ "Say ___ "
you ___ say

Yoruba /sɔ ___ lɛkã̀ sí/ "Say ___ again"
say ___ once repeat

Zuni /aɬna ___ lɛskwa/ "Say ___ again"
again ___ say

Navaho /díí saad ___ beeádíní/ "Say the word ___ "
this word ___ say you

Papago /hɨmu ___ kaidʒ/ "Say ___ now"
now ___ say

In the tabulation below, means of F1 and F2 for both male and female speakers separately are given for each semi-vowel/vowel combination in each language. The values given are the means of the individual means for all speakers who provided measurable tokens of the particular combination.

word	n	F1		F2	
		semi-vowel	vowel	semi-vowel	vowel
Male speakers					
Amharic (5 speakers)					
iji	15	230	266	2263	2252
uju	14	240	310	1984	694
aja	15	238	708	2106	1367
iwi	15	259	282	851	2173
uwu	15	269	305	564	679
awa	15	335	722	611	1280
Yoruba (5 speakers)					
iji	15	211	246	2106	2103
uju	15	239	277	1951	698
aja	15	268	684	2037	1279
iwi	15	243	282	677	2085
uwu	15	249	287	574	676
awa	15	361	679	663	1207
Zuni (6 speakers)					
iji	16	224	256	2093	2111
uju	8	260	301	1783	994
aja	14	403	761	1851	1218

iwi	18	251	279	961	2078
uwu	13	255	288	722	986
awa	19	437	692	865	1156

Papago (2 speakers)

iwi	6	293	318	929	2029
uwu	6	300	354	674	828
awa	6	427	586	740	1058

Navaho (5 speakers)

iwi	15	228	248	888	2146
awa	13	332	649	596	1125

Female speakers

Amharic (1 speaker)

uju	3	307	308	2383	850
aja	3	317	688	2410	1550
iwi	3	333	303	907	2257
uwu	3	327	337	625	780
awa	3	292	642	478	1417

Zuni (3 speakers)

iji	9	354	362	2588	2584
uju	1	450	450	2070	1160
aja	7	405	999	2300	1537
iwi	9	374	396	2588	2584
uwu	7	418	420	968	1039
awa	9	428	848	1007	1391

Papago (3 speakers)

iwi	9	379	390	1184	2370
uwu	9	414	422	748	808
awa	9	481	778	945	1238

Navaho (3 speakers)

iji	3	347	397	2703	2670
aja	2	420	750	2300	1673
iwi	8	301	349	900	2482
awa	9	342	681	701	1342