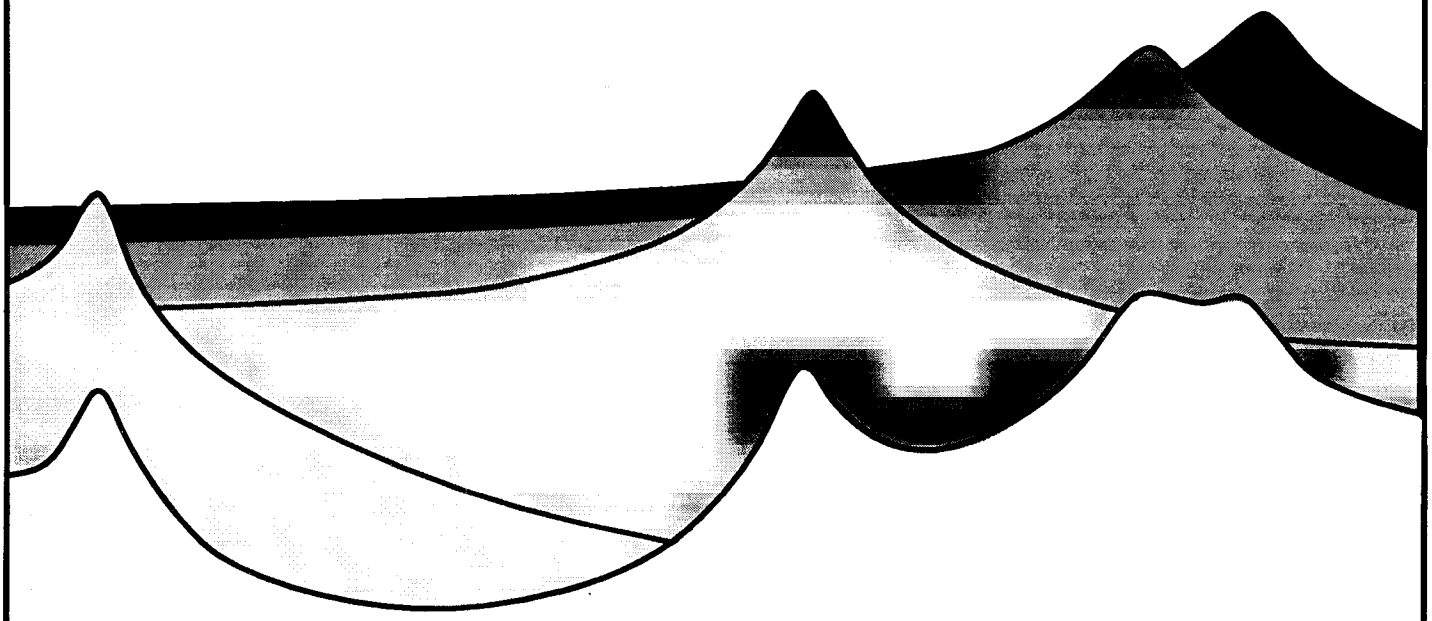


UCLA Working Papers in Phonetics

Number 88

September 1994

RH94

The UCLA Phonetics Laboratory Group

Victoria Anderson
Barbara Blankenship
Dani Byrd
Court Crowther
Sandy Disner
Minna Dokko
Beatriz Dukes
Edward Flemming
Vicki Fromkin
Susan Meyers Fosnot
Cécile Fougeron
Matt Gordon
Robert Hagiwara
Bruce Hayes
Susan Hess
Sue Banner Inouye
Sun-Ah Jun
Pat Keating
Paul Kirk
Jenny Ladefoged
Peter Ladefoged
Ian Maddieson
Peggy McEachern
Benjamin Munson
Bonny Sands
Cheng Cheng Saw
Aaron Shryock
Dan Silverman
Caroline L. Smith
Siniša Spajić
Donca Steriade
Henry Teheranizadeh
Kimberly Thomas
Richard Wright

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:
USHHS grant 5 T32 DC00029
NSF grant SBR 9107004
NSF grant DBS 9213604
American Speech-Language-Hearing Association Foundation Award
and the University of California

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

Phonetics Laboratory
Department of Linguistics
UCLA
Los Angeles, CA 90024-1543

The General Editor of UCLA Working Papers in Phonetics is Ian Maddieson. This issue was edited by Court Crowther, Associate Editor.

UCLA Working Papers In Phonetics 88

September, 1994

Table of Contents

A phonation type synthesizer for use in the field

Peter Ladefoged
Phonetics Laboratory, UCLA, Los Angeles, CA 80024-1543

This paper reports a somewhat eccentric endeavor: to design, construct and test a method of synthesizing differences in voice quality in vowels that uses only acoustic parameters and works in real time on a portable computer. The eccentricity of this notion may not be immediately apparent, as the utility of a system of this kind is clear. It would obviously be useful if there were a tool that enabled clinicians or linguists to go off into hospitals and schools, or remote areas such as the Kalahari Desert, and study voice qualities of patients or native speakers of little known languages, noting how they compare with a standard set of voice qualities as produced on a computer. The eccentricity comes when we require this to be done using strictly acoustic parameters in real time on a portable computer.

We are so used to describing speech in terms of the acoustic theory of speech production (Fant 1960), that we tend to forget that this is a description that does not rely on acoustic notions alone, as it is in partly physiological terms. It separates the properties of the vocal fold source from the properties of the vocal tract transfer function. Virtually all speech synthesizers do this in some way, in that they consider the source function not as an impulse but as having particular acoustic properties. For example, the Klatt synthesizer, which has been used in an excellent description of differences in phonation type (Klatt and Klatt 1990) has a glottal pulse source with a particular shape. More importantly from the point of view of considering the variables in descriptions of different voice qualities, one of the controllable parameters is the open quotient of the glottal pulse source.

In many respects the synthesizers that will be described in this paper will be the same as the Klatt synthesizer, although, in order to meet the constraints of being able to operate on a small computer, they will be somewhat simpler. They will be based on a purely acoustic description of speech sounds. This will have the disadvantage that the description will be less informative about presumed activity of the vocal folds. But it will have the advantage that it makes no assumptions about what belongs in which part of the description; and it will be in terms of measurable physical units. As against this, however, we must note a further major disadvantage: at the moment we have not yet succeeded in making an adequate description of different voice qualities in this form.

How can we describe a sound in terms of its acoustic parameters? Again this is not as straightforward as it might appear, even if we are considering only a vowel with normal phonation, say [a], spoken as a steady state quality but with a rise and then a fall in pitch. Just to make our problem more definite, we will assume that we are concerned with frequencies up to 5,000 Hz, and that the vowel has a duration of 200 ms. Figure 1 shows all that we can see of this soundwave.

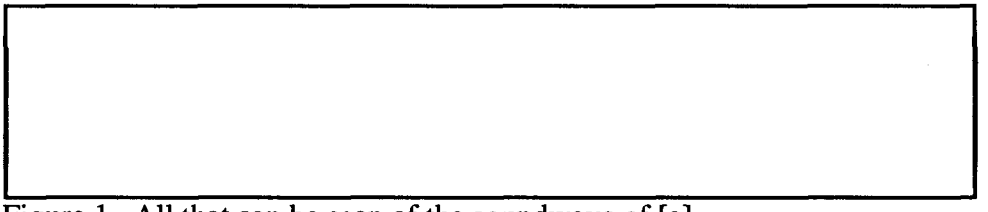


Figure 1. All that can be seen of the soundwave of [a]

This figure is a salutary reminder that we have to make a transformation of some sort before we can make a description of any kind. The simplest possible transformation is to consider the sound pressure wave, which can be represented graphically as shown in Figure 2.

Figure 2. A graphical representation of the sound pressure wave of [a]

We now have to make a parametric description of this soundwave. The next transformation we might make is an analog to digital conversion, sampling the sound wave at regular intervals. We could then represent the soundwave as shown in part in Figure 3. The digital representation *is* a parametric description, but in terms of a large and unspecified number of parameters. Accordingly we must consider how to reduce this representation by expressing it in terms of a smaller number of parameters. The obvious technique — one which we will in fact adopt later — is to do this in the spectral domain. But this involves a number of hidden assumptions about what parts of the soundwave are, and what are not, the same. We can see this more easily if we continue trying to parameterize the values of the samples in the time domain.

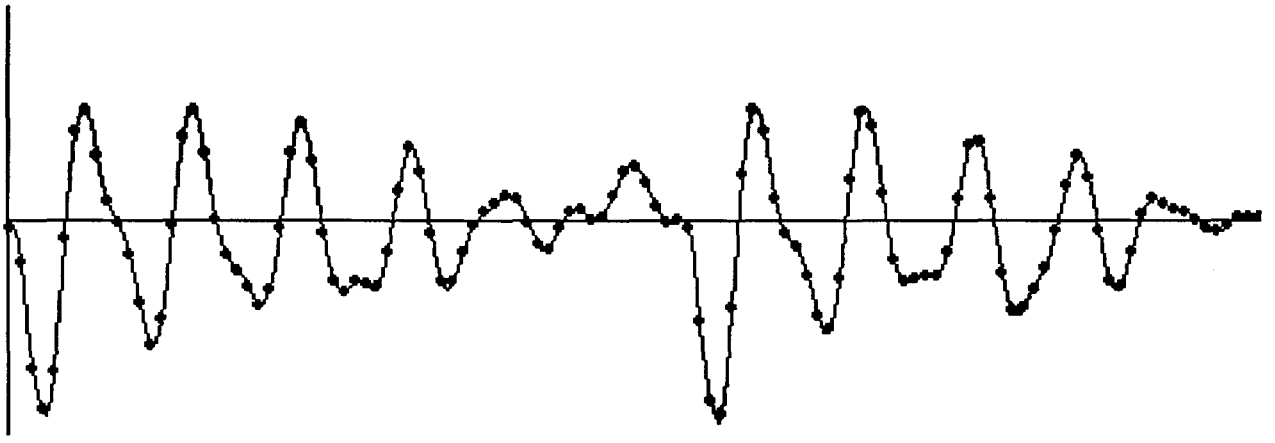


Figure 3. Sampled points in part of the soundwave in Figure 2.

One way of giving a parametric description in the time domain is to use the LPC algorithm. This algorithm calculates a set of N coefficients that can be used as weights that enable any set of N sample points in the waveform to determine the value of the next sample point with the least possible error. This means that we have described the soundwave in terms of N parameters — but with a certain amount of error. The coefficients can be regarded as the time domain specification of a filter with $\frac{N}{2}$ poles corresponding largely to the resonances of the vocal tract. The error signal, the so-called LPC residual, is correlated with differences in phonation type. But it is not at all clear how to parameterize the LPC residual so that it can be used in an acoustic synthesizer.

Now consider a spectral domain description of the waveform in Figure 3. If we literally make a DFT (discrete fourier transform) of this entire signal, which has frequencies up to 5,000 Hz and a duration of 200 ms, we will be specifying the intensities of 1,000 sine and cosine components. Again, this could be called a parametric description of the signal, albeit in terms of a large number of parameters. We could reduce this to a more manageable number by describing the spectrum in terms of the poles. Nevertheless this set of parameters is fatally flawed as a

description that could be used to synthesize an equivalent sound, as the DFT of the 200 ms waveform does not reflect changes in pitch within that 200 ms.

The usual approach to this problem is to analyze a series of windows throughout the duration of the sound. As we are dealing with a steady state vowel, the poles in the spectrum will be the same in each of these windows. If we can somehow characterize the pitch, we could make each window a single pitch interval, and characterize the sound in terms of the poles in a pitch synchronous DFT. These considerations suggest how we might consider the signal so that we can make a purely acoustic description of it in parametric terms. The solution chosen here is to specify the sound in terms of its spectrum *considered as the impulse response of a filter.* We will not be making a traditional source–filter specification, as the source will be considered to be either random noise or a series of impulses. All other specifications of the sound will be made properties of the filter.

Synthesizer overview

Let us now see how we can design a synthesizer that will operate in this way. We will begin by considering how many parameters we will require for the specification of the input to the filter. For a standard modal vowel we will regard the impulse amplitude as fixed, and not requiring parametric specification. We will then need one parameter to specify the percentage relative amplitude of the random noise with reference to the impulse amplitude. We might expect this parameter to have a value of zero in a modal vowel with no random noise, and a value of 100 in something like a very whispery vowel in which there are no vibrations of the vocal folds.

Other parameters will be concerned with the intervals between the impulses in the time domain. As we are considering only vowels with a pre-determined rise and then a fall in pitch, one variable, specifying the starting pitch (or, to be more precise, the interval between the first two pulses) will be sufficient for vowels with a modal voice. From this one pitch parameter a simple rise fall intonation will be derived.

Other parameters will be needed to describe vowels with irregular pulse intervals. In the synthesizer describe by Klatt and Klatt (1990) there are two such parameters, the degree of regular diplophonic double-pulsing, and the quasi-random period to period fluctuations of the pulse frequency which Klatt and Klatt call flutter. We will refer to these variables as regular jitter and random jitter, and vary them in much the same way as in the Klatt and Klatt synthesizer.

The Klatt and Klatt synthesizer does not provide for the parametric specification of another possibility, irregularities in the pulse amplitudes, sometimes referred to as shimmer. As with irregularities in the pulse intervals, these can be of two kinds, regular alternating of high and low amplitude pulses, and quasi-random fluctuations. As the final parameters associated with the input to the filter, we will specify these two possibilities. The other Klatt and Klatt parameter, the tilt of the voicing source, is a property of the filter in our model.

These variations in the impulses are summarized in Figure 4. The top line shows what we may take to be the standard series of impulses. The next line shows the effect of varying the pitch control — the interval between impulses. The third line shows regular jitter, in which impulses are controlled in pairs. The regular jitter parameter sets the interval between the first member of a pair and the second; if there is no regular jitter, this interval will be the same as that determined by the pitch parameter. The fourth line shows random jitter. The mean interval between pulses is set by the pitch parameter, and the random jitter parameter varies this interval by a larger or smaller random amount (which can, of course, be zero, as in modal voice). The final pair of lines show shimmer variations. In regular shimmer the amplitude of alternate pulses is varied by a fixed amount set by the regular shimmer parameter. In random shimmer the amplitude of each pulse is increased or decreased by a random amount whose magnitude is controlled by this parameter.

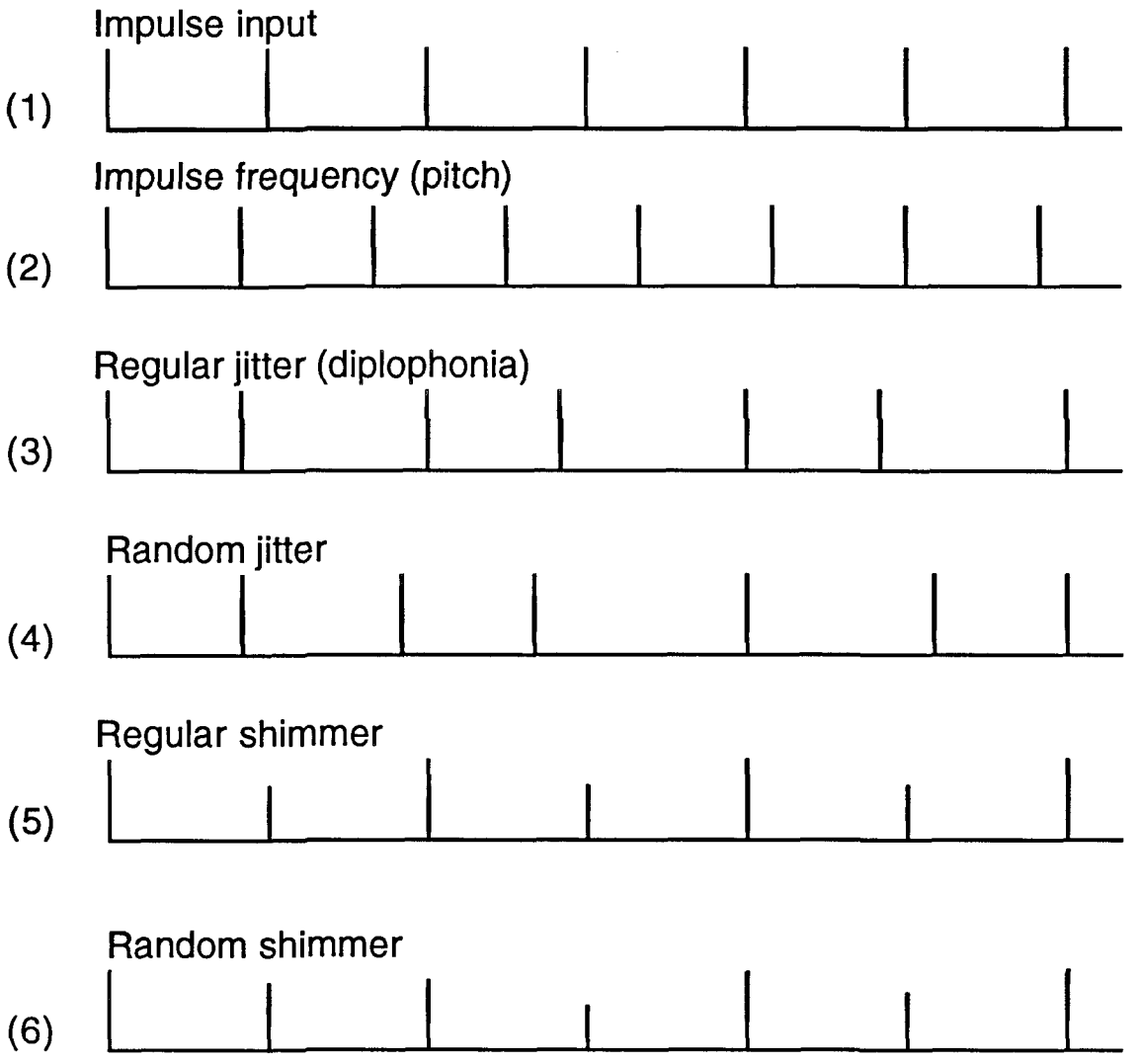


Figure 4. Parameters controlling the specification of the impulses input to the filter.

Next, how do we specify the filter? We could use a time domain approach, and specify, say, 12 coefficients of an FIR filter. These coefficients are, however, notoriously difficult to manipulate. The individual coefficients cannot be interpreted in a meaningful way. Filters are usually described in the spectral domain, with the poles being specified in terms of their frequencies and bandwidths, which is the approach we will adopt here. However, although some of the poles of the filter can be given a physiological explanation in that they correspond to resonances of the vocal tract and the effects of lip radiation, they will be regarded here as simply as determinants of the spectrum.

The general form of such filters is shown in Figure 5. The upper part of the figure shows the four poles for an [a] like vowel and the lower part those for an [i] like vowel. In both spectra there is an additional fixed pole representing the higher formants. As can be seen in the summation of the curves in these two sets of graphs, given the particular function chosen for this higher pole, the formant frequencies above 2,680 Hz are increased, and those below are decreased.

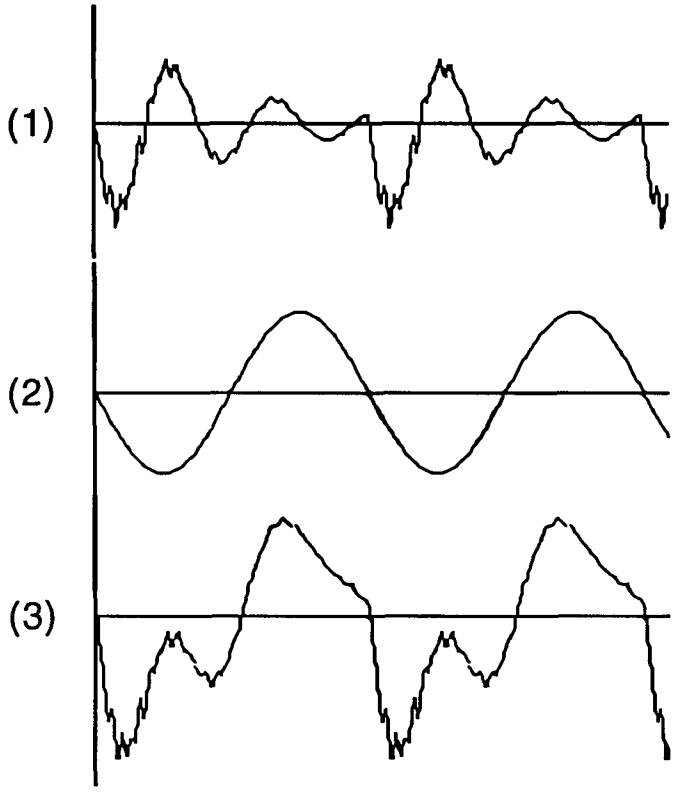


Figure 6. Summation of waves to simulate a breathy voiced wave for the vowel [i].

Synthesizer implementation

The implementation of a synthesizer using these parameters on a small computer is not a simple task, especially if we add the constraint that it should operate in something like real time. We have so far been unable to implement a synthesizer which can produce different vowels with a full range of phonation types using all the parameters described above without a noticeable delay. We have, however, constructed a synthesizer that will produce a wide range of vowels with a number of different voice qualities in almost real time. We have also constructed a variant of this synthesizer using all the parameters described above which will produce a smaller set of vowels with more varied phonation types, but with a small delay.

In both these synthesizers, a set of possible formants is determined, and a single cycle of a damped wave is calculated and stored for each of them. In the synthesizer that will produce a wide range of vowels the lowest possible formant frequency is taken to be 230 Hz. This frequency is then increased by 5% steps up to 3366 Hz, resulting in 56 possible values for the first three formants. Within this range, the values that are used for F1 are between 230 and 859 Hz; the values for F2 are between 742 and 2392 Hz; and those for F3 between 1785 and 3366 Hz. A fixed fourth formant at 3500 Hz is also included. These waves are put into a data array whenever the program is started. Each of these waves can be considered as the impulse response to a single pole in a filter such as that shown in Figure 4.

The waveform for a single cycle of a vowel is constructed by summing three selected damped waves, and combining them with the fourth fixed frequency damped wave. The waveform for a whole vowel is constructed by repeating this waveform for a single cycle, the start of a new repetition being determined by the parameters that control the intervals between impulses. The default is for the intervals to correspond to a pitch curve that starts at 110 Hz, rises to 120 Hz

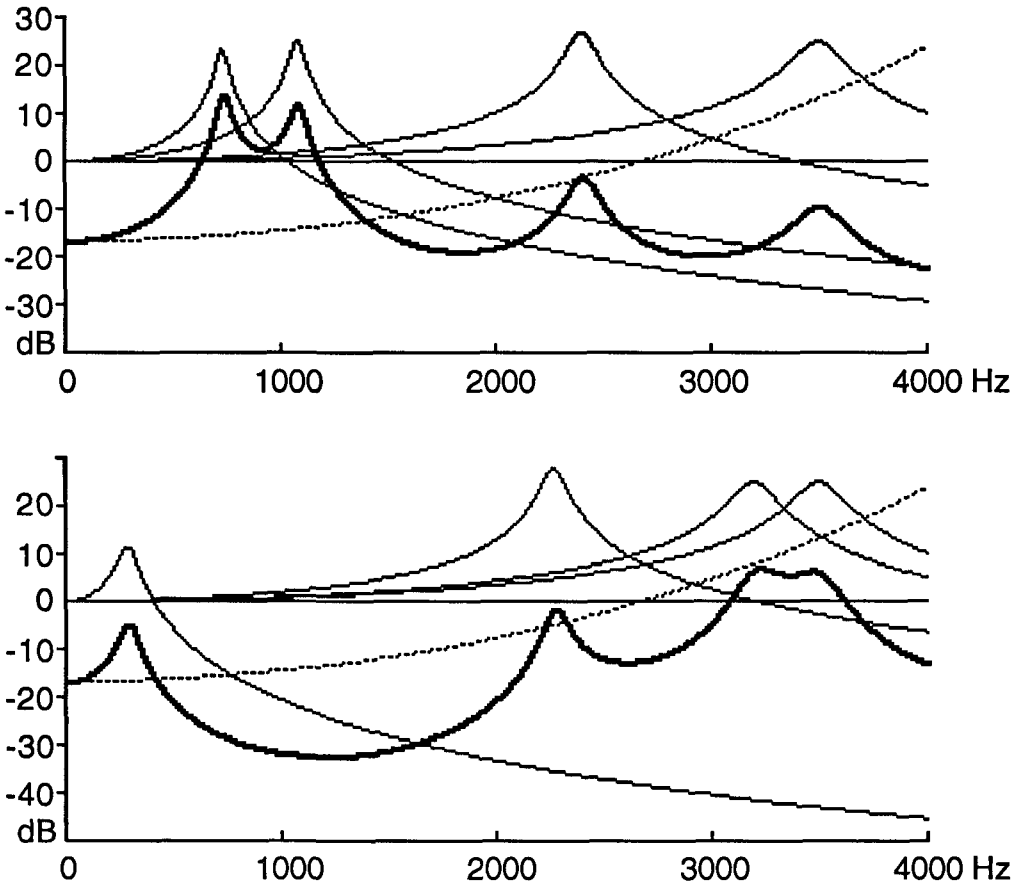


Figure 5. Poles representing the default vowels [a] and [i]. The sum of these five curves is shown by the heavy line.

There is a further problem that has to be considered in the design of a phonation type synthesizer of this type. In a sense we need a further low frequency pole. A great deal of research has shown that one of the best ways of quantifying differences in phonation is in terms of the relative amplitude of the fundamental frequency and some measure of the energy in the rest of the spectrum (Ladefoged, Maddieson, and Jackson, 1987).). Sounds with breathy voice have more energy in the fundamental frequency, and those with a creaky voice have less. But the fundamental frequency is varying, and therefore cannot be considered as a property of the filter determining the spectrum of the entire 200 ms soundwave. The Klatt and Klatt synthesizer achieves this effect by varying the open quotient of the voice source. In our eccentric attempt to use an acoustic specification that does not rely on a particular glottal pulse shape, this variation in the output wave can be achieved by simply adding in a 'sinusoidal' wave, each period of which corresponds to the interval between consecutive impulses into the filter, as shown in Figure 6. The first line in this figure is the calculated wave for two periods of the vowel [i]. The second line is a sinusoidal wave with the same frequency, and the third line is the sum of the first two lines. The combined wave is very similar to that in breathy voiced vowels.

during the first third of the vowel, and falls to 100 Hz during the last two thirds of the vowel. The starting pitch is a user specifiable parameter.

There are three variables to be considered for each damped wave, its frequency, bandwidth and amplitude. A standard equation is:

$$v_n(t) = \sum_{n=1}^{\infty} A_n e^{-\pi B_n} \cos(2\pi F_n t + \phi_n) \qquad (1)$$

where A_n is the amplitude of the damped wave, B_n is the bandwidth, and F_n is its frequency. A consistent value of $\pi/2$ for the phase changes the cosine into a sine function. Sine waves are used in this implementation as it is necessary for the first point of each damped wave to be zero. In this way, when one wave is added at the end of another (where the amplitude is near zero), there is a smoother transition.

If single cycles of each damped wave are to be stored, the bandwidths must be pre-determined. Accordingly, for the faster of the two synthesizers, the bandwidths were calculated by means of a function suggested by Fant (1972):

$$B_n(F) = 15\,(500/F_n)^2 + 20\,(F_n/500)^{0.5} + 2.8\,(F_n/500)^2 \qquad (2)$$

In the synthesizer which allows for a greater range of voice qualities the bandwidths of the poles in the filter are user specifiable parameters. This entails recalculating the damped waves whenever the bandwidth is changed, which causes some of the increased delay in this synthesizer.

The relative amplitudes of the formants are determined by a number of factors. The first is the frequencies and bandwidths of the formants themselves. When two formants are closer together in frequency, their resonance curves will have a greater overlap, so that their amplitudes will be increased. This effect will be somewhat less if the bandwidths of the formants are large. Another factor is the effect of higher formants, as shown in Figure 5. Analytic expressions developed by Fant (1979) provide a basis for modifying the initial amplitudes of the damped waves in order represent the effects of these factors. For a four formant system we can account for the interactions of the formants (including the higher formants) by means of the following expressions:

$$A_1 = hPoles(F1) / ((1 - F1^2/F2^2)(1 - F1^2/F3^2)(1 - F1^2/F4^2))$$
$$A_2 = hPoles(F2) / ((1 - F2^2/F1^2)(1 - F2^2/F3^2)(1 - F2^2/F4^2))$$
$$A_3 = hPoles(F3) / ((1 - F3^2/F1^2)(1 - F3^2/F2^2)(1 - F3^2/F4^2))$$
$$A_4 = hPoles(F4) / ((1 - F4^2/F1^2)(1 - F4^2/F2^2)(1 - F4^2/F3^2))$$

where $hPoles(F) = 20 \log_{10}(0.54 * (F / 500)^2 + 0.00143 * (F / 500)^4)$, a curve which is illustrated in Figure 5.

The parameters that are available to the user of the faster synthesizer can be seen in Figure 7, which shows the screen of the Macintosh computer when the program is running. On the left there is a block of 20 squares, representing major differences in vowel quality as determined by the first two formants. Clicking in any square will produce a vowel with the indicated values of the first two formants. If a narrower specification of vowel quality is required, clicking in the "Detail" button brings up a narrower grid, allowing more precise determination of the formant frequencies. The third formant can be determined automatically, in a way that is appropriate for English on the basis of the first two formants (Broad and Wakita, 1977), or can be selected by the user. There is

also for a scale for varying the pitch, and selection boxes that can be used to vary the vowel length. Actual numerical values of all these variables are displayed in a separate window when a vowel is calculated. The other buttons at the bottom of the screen are for controlling the experimental protocol, allowing the subject to sign in, hear models of the sounds to be matched, see graphic displays, clear the screen, and so on.

Some of the controls are parenthesized. At the moment this synthesizer has no way of simulating nasal vowels (which would require specifying an additional zero in the filter, and recalculating the amplitudes of the damped waves). There is also no provision for breathy vowels. It is, however, possible to simulate creaky voice by varying the impulse rate. This control is the same as that in the second synthesizer which we will now describe.

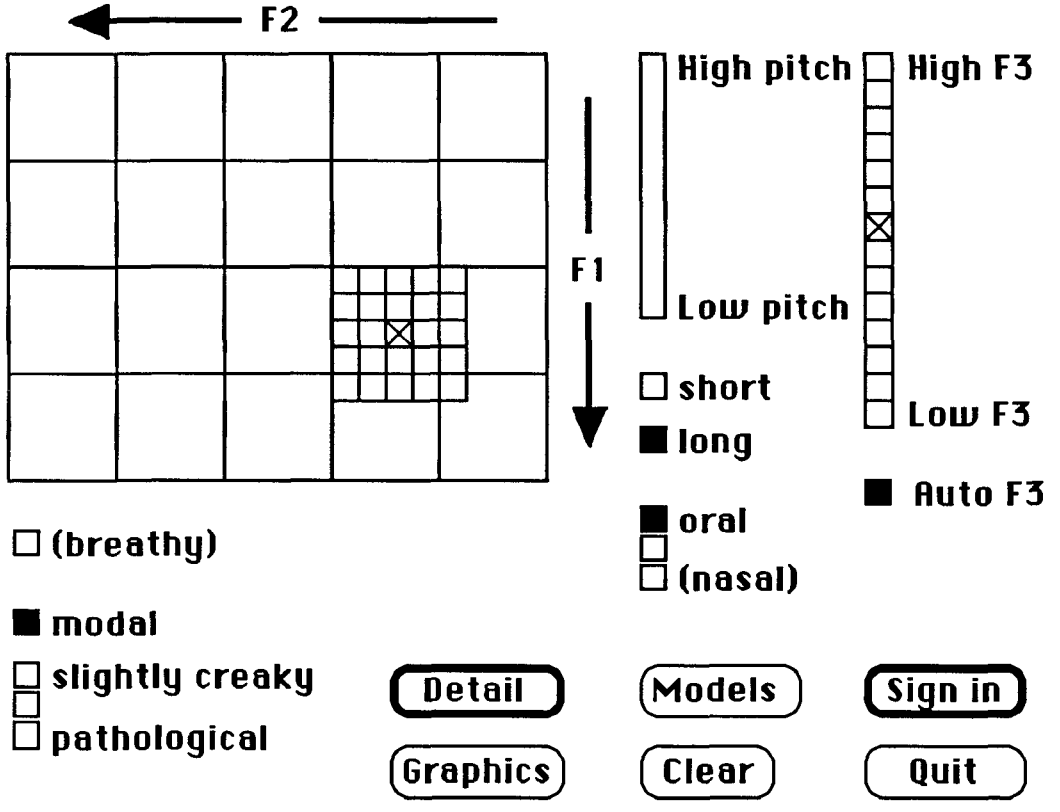


Figure 7. The parameters available to a user in a vowel synthesizer with limited phonation type possibilities.

The 13 controls available in the second synthesizer are shown in Table 1. We will consider each of these parameters in turn. (1) instead of 16 possible values of F1, 20 of F2, and 14 of F3, only two sets of formant frequencies are possible, one for an [i] like vowel, and one for a vowel of the [a] type. (2) The starting pitch can be varied so that it is possible to simulate both male and female voices. (3) Regular (diplophonic) jitter can be produced by varying the interval between the first and second pulses in a pair as outlined above. If the second pulse follows at the interval determined by the intonation curve there is zero jitter; if it follows at half that interval there is maximum regular jitter. (4) The degree of randomness in the irregularity of the pulse intervals can be varied by adding (or subtracting) a random number from the inter-pulse interval equivalent to up to half the inter-pulse interval. (5) and (6) The relative amplitudes of adjacent periods can be altered in two ways. Regular shimmer occurs when the amplitude of one period is a fixed percentage of the amplitude of the adjacent period, random shimmer when it is a random percentage of the

adjacent period. (7) Random noise is produced by adding a random number to each point in each damped wave. (8) The relative intensity of the fundamental frequency is varied by adding the required amplitude of a sinusdoidal wave. Whenever the starting pitch is changed, this waveform has to be recalculated, which is one of the factors causing this synthesizer to run more slowly. (9) The spectral slope can be varied by altering the relative amplitudes of the poles. (10) - (13) The bandwidths of the poles can be independently controlled; each time one of the bandwidths is changed, the corresponding damped wave has to be recalculated, another factor that makes this synthesizer run more slowly. All 13 parameters can be controlled by the user through a Macintosh interface similar to that in Figure 6.

Table 1. User controlled parameters in a system designed to synthesize a number of different phonation types

	PARAMETER	RANGE
1	Vowel quality	[a] or [i]
2	Start pitch	100 to 250 Hz
3	Regular jitter	0 to 0.5 * F_0 Hz
4	Random jitter	0 to random of (0.5 * F_0)
5	Regular shimmer	0 to 100 % alternate periods
6	Random shimmer	0 to random of (100) % alternate periods
7	Random noise	0 to 100 % of maximum pulse amplitude
8	Amplitude added F0	0 to 100 % of maximum pulse amplitude
9	Spectral slope	+3 to -24 dB/octave
10 – 13	Bandwidths 1 - 4	0.5 to 4.0 * default bandwidth

Perceptual testing

So far, only informal testing of the phonation type synthesizer can be reported. This is largely because one of the main objectives of the project has not been achieved. It was hoped that the phonation synthesis could be attempted in the field, in much the same way as vowel matching has been achieved, using a synthesizer as described above (see Johnson, Wright and Flemming, 1992). Unfortunately the phonation type synthesizer has too many variables for subjects to be able to use it to imitate linguistically significant voice quality differences in their own speech. A number of findings can, however, be noted.

The vowel quality and basic pitch controls are satisfactory, although there is some unnaturalness in the female voice, as the formant frequencies chosen were selected for a male voice. There appears to be little perceptual difference between regular jitter (diplophonic voice) and random jitter when only small amounts of either are present. Greater quantities of these variables produce differences that are perceptually apparent, but are not easy to associate with known voice qualities. Shimmer on alternate periods and random shimmer are also similar. The addition of random noise by itself produces an inhuman voice quality. Combining this with an increase in the relative intensity of the fundamental frequency, and a decrease in the bandwidths of the formants, produces a more breathy voice, but it is far from satisfactory as an imitation of linguistically significant contrasts in breathy voice as observed by, for example, Kirk, Ladefoged, and Ladefoged (1993). More work needs to be done in selecting the best combinations of these many variables, and in providing controls that will allow them to be altered in conjunction with one another. It is hoped that these possibilities will be explored in forthcoming linguistic phonetic fieldwork.

Acknowledgments

Thanks are due to Gunnar Fant for his help in the early development of these synthesizers.

Discussion of this paper

Janet Pierrehumbert: I wonder how well the spectra people select match the spectra that they produce, or are there systematic mismatches?

Peter Ladefoged: I haven't done enough to be able to answer that. My impression is that people are pretty random in their behavior. Each person sets off on one track and meanders around trying different things.

Kenneth Stevens: I guess one of the problems for these different voice qualities is you've got too many parameters. What's the next step?

P.L.: My next step is to turn to some group like this and say: "How do we simplify all these parameters? What can we do in order to turn it into something more useful?

K.S.: Why not go to the Klatt type of synthesizer or use an LF model or something like that for the glottal source?

P.L.: This synthesizer is very similar to the Klatt synthesizer when you get down to it. The one control that's (deliberately) missing in my synthesizer is the duty cycle or open quotient. My separation between regular and random jitter is slightly different from Klatt's; and I have also added control of shimmer.

K.S.: But it might be possible to link those parameters, which is I guess what you're saying.

Richard McGowan: It seems that you don't get more articulatory control.

P.L.: I don't get articulatory control because one of my notions was that we should deliberately move away from articulatory controls and see what we could learn. And the answer is not much. I think that you're right, a more articulatory approach would, in fact, be a better way of building a synthesizer like this.

T. Ananthapadmanabha: You do not have a voice source anywhere in the system. You're using impulse excitations and trying to achieve the balance by the amplitude manipulation, so you do miss some information in the low frequency region contributed by the voice source. You don't have the representation for voice source, the glottal source.

P.L.: Yes, deliberately. I thought there were good reasons for avoiding that representation. Whenever you make a source filter distinction, you make certain assumptions about what lives in the source and what lives in the filter, and that leads to problems. However, those turn out to be smaller problems than I have with my model

Osamu Fujimura: I wonder if you could separate out different uses or situations where voice quality changes occur, in particular, idiosyncratic (personal) voice characteristics as opposed to the controlled use of change of voice quality, of whatever kind, for phonetic or pragmatic purposes. Also, you have these features listed in the menu, but the list is probably not complete: for example, you mentioned dynamic breathy voice, but you don't mention dynamic spectral slope, which I'll be discussing in my paper. We clearly need some kind of concept for sorting out different measures associated with different functions.

Margarita Mazo: I wonder if it is possible to hear the qualities you list here. For example, I am interested very much in dynamic breathy voice and the other one, the difference between creaky and pathological voice.

P.L.: It is certainly possible. It's a simple 120K program. Anybody who has a Macintosh disk is welcome to take a copy home and play with these programs for themselves and try them out.

Ingo Titze: Following up on what Ken said, I see only two paths you can take to simplify a synthesizer in terms of parameters. You either build more physics into it , where parameter A varies with parameter B in an actual physical sense or you go to a rule system where you look at enough people doing it and you connect parameter A with parameter B in that fashion. I don't know which path you're taking. It seems like the other paths are the paths people have been taking for years.

P.L.: Absolutely. This was an attempt to break out from the paths that other people were following, and frankly it didn't work. If there's a lesson to be learned it's that one wants to go back to what other people have been doing in a more articulatory way or a more physical way. I got inspired by a paper that came out in *Language* by my colleagues, Johnson, Flemming and Wright (1993) on matching vowel qualities. This type of synthesizer works nicely on vowels, and we have been doing very well getting people to match their vowel quality. I hoped I could do similar work on different voice qualities. But vowels are simpler than voice qualities, and I doubt these matching techniques will be successful. This type of synthesizer has its merits in that it's available, and allows you to control a great number of parameters, using any Macintosh computer; but it is not the wave of the future.

References

Broad, D. J. and H. Wakita (1977). "Piecewise- planar representation of vowel formant frequencies." Journal of the Acoustical Society of America **62**(6): 1467-1473.

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

Fant, G. (1972). "Vocal tract wall effects, losses and resonance bandwidths." Quarterly Progress and Status Report, Speech Transmission Laboratory, Royal Institute of Technology **2-3**: 28-52.

Fant, G. (1979). "Vocal source analysis - a progress report." Quarterly Progress and Status Report, Speech Transmission Laboratory, Royal Institute of Technology **3-4**(1979): 31-54.

Johnson, K., R. Wright, and E. Flemming. (1992). "Using the method of adjustment to study vowel spaces." Journal of the Acoustical Society of America **91**(4): 2387.

Kirk, P. L., J. Ladefoged, and P. Ladefoged. (1993). Quantifying acoustic properties of modal, breathy and creaky vowels in Jalapa Mazatec. American Indian Linguistics and Ethnography. In honor of Laurence C. Thompson.

Klatt, D.H., and L.C. Klatt (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. Journal of the Acoustical Society of America, **87**, 820-857.

Ladefoged, P., I. Maddieson, and M. Jackson. (1988). Investigating phonation types in different languages. Vocal Physiology: Voice Production, Mechanisms and Functions. New York, Raven. 297-317.

The links between theory and fieldwork in phonetics

Peter Ladefoged
Phonetics laboratory, UCLA, Los Angeles, CA 90024-1543, USA.

The relation between phonetic fieldwork and phonological theory provides an interesting chicken-and-egg problem concerning which comes first. Does the phonetic account of the sounds that occur in the language, precede or follow the phonological account of the sounds that contrast? About 50 years ago there was a clear answer to this — one that we would now consider to be wrong. The process of describing a language was considered to be one in which the fieldworker went out and made a detailed record of all the sounds that were present, and then sorted out those that appeared in contrasting contexts and called them phonemes, considering the rest, those that were in complementary distribution, to be allophones. At least, that was what happened in theory. In practice fieldworkers behaved as now, observing what their consultants considered to be different words, and noting the phonetic realities underlying the distinctions. Only rarely did a fieldworker first note the occurrence of two different sounds and then later observe that they were in complementary distribution and therefore members of the same phoneme.

The position taken here is that a complete phonetic description of a language can be made only when the phonology of that language has been determined. The facts that have to be described are phonological facts, so phonetics follows phonology rather than preceding it. Nevertheless, before the phonology is fully known, a partial phonetic description can help determine the most appropriate account of the phonology, so the two often progress together.

Linguists want to explain why language is the way it is. We want to know what constitutes a possible speech sound in a human language. But before we can answer this we need a good phonetic account of all the known phonological contrasts in a wide range of languages. We also want to compare languages, and account for their possible common origins. Much of comparative linguistics rests on the tacit notion that it is possible to identify a speech sound in one language as being in some sense the same as another speech sound in another language. This is tantamount to saying that there is a finite set of speech sounds, and any speech sound in a language can be identified as being equivalent to a particular member of this set. In theory it is possible to enumerate all the possible speech sounds. However, in practice it is impossible to do this, because it is not easy to say whether two sounds occurring in different languages are the same or not. For example, language teachers point out that the vowels in French 'bête' and English 'bet' are not the same. But are they linguistically two distinct sounds? Two sounds are definitely different if they distinguish words within a language, but if they occur in different languages this test cannot be used. This is the basis of what we might call the IPA problem. The International Phonetic Association tries to provide a way of symbolizing every distinct speech sound, but it cannot tell whether two sounds are potentially distinct when they have been observed only in different languages.

An alternative way of comparing sounds in different languages is to say that there is a set of phonological features that have known phonetic attributes. Any speech sound can then be described in terms of these features. It might seem that the entire set of speech sounds could be specified by listing the possible combinations of features that can occur, but a featural approach does not avoid the IPA problem. To continue with the same example, the vowels in French 'bête' and English 'bet' might be given the same feature specification, although they do not sound the same. In all probability, no language can or could use the difference between these two sounds to distinguish between words. But only by doing extensive fieldwork on a large number of languages

can we even estimate the magnitude of the difference in quality that is required for two slightly different vowels to be capable of distinguishing words in a language.

The data from recent phonetic studies (Ladefoged and Maddieson, in press) indicate that it is preferable to use a slightly different approach to the problem of comparing the phonetic structures of different languages. One must start from an account of the phonology, but rather than using a featural approach to characterize the phonological items, it is preferable to use general phonetic parameters that are measurable, such as the frequencies of the formants, and the length of the voice onset time. Many of these parameters are analogous to features, but in each of them there is a continuous range of possible values rather than a binary (or unary or n-ary) choice.

The number of parameters along which sounds can vary is fairly small. There are limitless ways in which the sounds of one language can differ from those of another, but in most cases the variation between similar sounds in different languages is the result of the use of a different value of one of the parameters, rather than the use of some novel parameter that has not been observed in other languages. A parametric approach avoids the IPA problem by taking no decision as to what is a distinct sound. It allows comparison between two languages on a phonetic level, but it leaves it up to the observer to decide in each case whether a given sound in one language can be considered to be the same as a sound in another language or not. A parametric approach merely notes that there is a consistent difference between the values of a parameter for a sound in one language as opposed to another. It does not say whether this difference is potentially capable of distinguishing between words in some third language.

Fieldwork procedures for describing phonetic structures

We will now consider how we can obtain phonetic data for comparing languages. We will take as an example the procedures that are being used in a project sponsored by the National Science Foundation for describing the phonetic structures of endangered languages (Ladefoged and Maddieson, forthcoming). When conducting fieldwork that aims to document the phonetic structures of endangered languages, the first step takes place well before leaving home: the phonology of the language must be studied. As was emphasized above, a complete phonetic description of a language can be best made only if the phonology is known. This restricts the choice of languages that can be studied. We have found that a very good way for us to achieve our goals is when fieldwork can be undertaken together with a linguist who has already analyzed the language. Accordingly we spend time preparing to work with local linguists and their consultants. Before going into the field we read everything that we can on the languages in the area. If there is insufficient material, or no local linguist available, we look for another language.

Another restriction on this type of fieldwork is that there must be a sufficient number of fluent speakers. We need to analyze languages that are not yet moribund. As we will see, phonetic analyses should be based on recordings of at least half a dozen speakers; preferably there should be half a dozen speakers of each sex. In addition we cannot study languages that are spoken only by a small number of elderly speakers, who are themselves fluent in another language. The pronunciation of these speakers is too apt to be influenced by other languages; as a result, they do not provide reliable phonetic data. In addition we need speakers who are willing (and, we hope, even eager) to assist us. The most productive way to work is to go where we have the enthusiastic support of the local community, even though this may mean that we do not undertake studies of some well known dying languages.

The major task is to record the right material from a sufficient number of speakers, in a technically satisfactory way. The procedures that we use have been described in detail by Ladefoged (1993). Illustrating the phonological events in a language entails recording the complete

set of vowel and consonant contrasts as well as the suprasegmental patterns. It would be nice if it were possible to record all the desired items in naturally occurring speech, but this is impossible. There is no way in which one can sit around waiting for half a dozen speakers to use each of a set of words such as "heed," "hid," "head," "had," etc. in spontaneous speech. Accordingly it is necessary to prepare word lists. Tone and stress contrasts are usually fairly easy to illustrate in citation forms, but phrases or sentences illustrating features such as tonal assimilations and stress shifts must be included.

It cannot be emphasized too strongly that the preparation of good word lists is the key to linguistic phonetic fieldwork. These word lists represent the links between previous theories of the sound system of the language being investigated and our own fieldwork. Before going out into the field we make consonant, vowel and suprasegmental charts, and list words illustrating each item (much as the "Illustrations" of languages in the *Journal of the International Phonetic Association*). When the time comes to work with speakers of the language, it often turns out that they do not know some of the words, or pronounce them differently. But it pays to have thought the whole process through before beginning work with native speakers.

Findings

We will now consider some of the findings of this research, and show how they have contributed to theoretical issues in phonetics and phonology. We will begin by examining data from Tsou, an Austronesian language spoken in Taiwan, and seeing what light it sheds on the issue of permissible consonant clusters.

Table 1. Tsou consonant clusters attested in the data in Wright and Ladefoged (1994)

x = clusters that appear in word initial and word internal position

(x) = clusters that appear only in word internal position

	C2 →	p	f	v	ɓ	m	t	ɗ	ts	s	z	n	k	ŋ	ʔ	h
C1 ↓	p						x		x	x	(x)	x	x	x	x	x
	f						x		x			x	x	x	x	
	v								x			(x)			(x)	x
	ɓ											x	(x)			
	m	x	x		(x)		(x)		x	x	x	x			x	x
	t	x	x	x	(x)	x						x	x	x	x	x
	ɗ															
	ts	x	x	x	(x)	x						x	x	x	x	x
	s	x		x	x	x						x	x	x	x	
	z														(x)	
	n	(x)				x	x		(x)	x	(x)					(x)
	k								(x)	x		x			(x)	
	ŋ			x		(x)	(x)		(x)	(x)	(x)		(x)			x
	ʔ	x	(x)	x		x	x		x	x		(x)	(x)			(x)
	h	x		x		x	x		x		(x)	x	(x)	(x)	x	

Tsou has been studied by a number of fieldworkers (Tung 1964, Starosta 1974, Tsuchida 1976, Ho 1976, Li 1979). From their work it is apparent that Tsou has a wide array of consonant clusters, but the phonetic characteristics of these clusters have not been described. Tsou has a

(C)CV(V) syllable structure, so these clusters are permissible in word initial position, and medially as the onsets of syllables within a word. Table 1, from Wright and Ladefoged (1994), illustrates the combinations of consonants found in data we collected in recent fieldwork.

A Tsou consonant cluster is made up of two and only two consonants, and may consist of almost any two of the Tsou single consonants in either order. There are a few exceptions to this generalization. Homorganic fricative-stop or stop-fricative clusters do not occur, and nasals only precede homorganic stops or fricatives. The implosive consonants **ɓ** and **ɗ** are comparatively rare and have not been observed in some possible clusters. Not all of the consonant clusters that are possible have been documented in word initial position.

Many of these clusters are not at all common in the world's languages, presumably because, if such a cluster were to develop in a language, one of the two components would not be sufficiently salient, and would be lost. The interesting question is how does Tsou avoid this loss? The answer is by reinforcing the characteristics of a consonant that might otherwise not be perceptually salient. For example, both members of a stop-stop cluster are fully released. There is an audible burst between a voiceless stop and a following voiceless segment, as illustrated in Figure 1. As a result the two components of the cluster are both clearly evident.

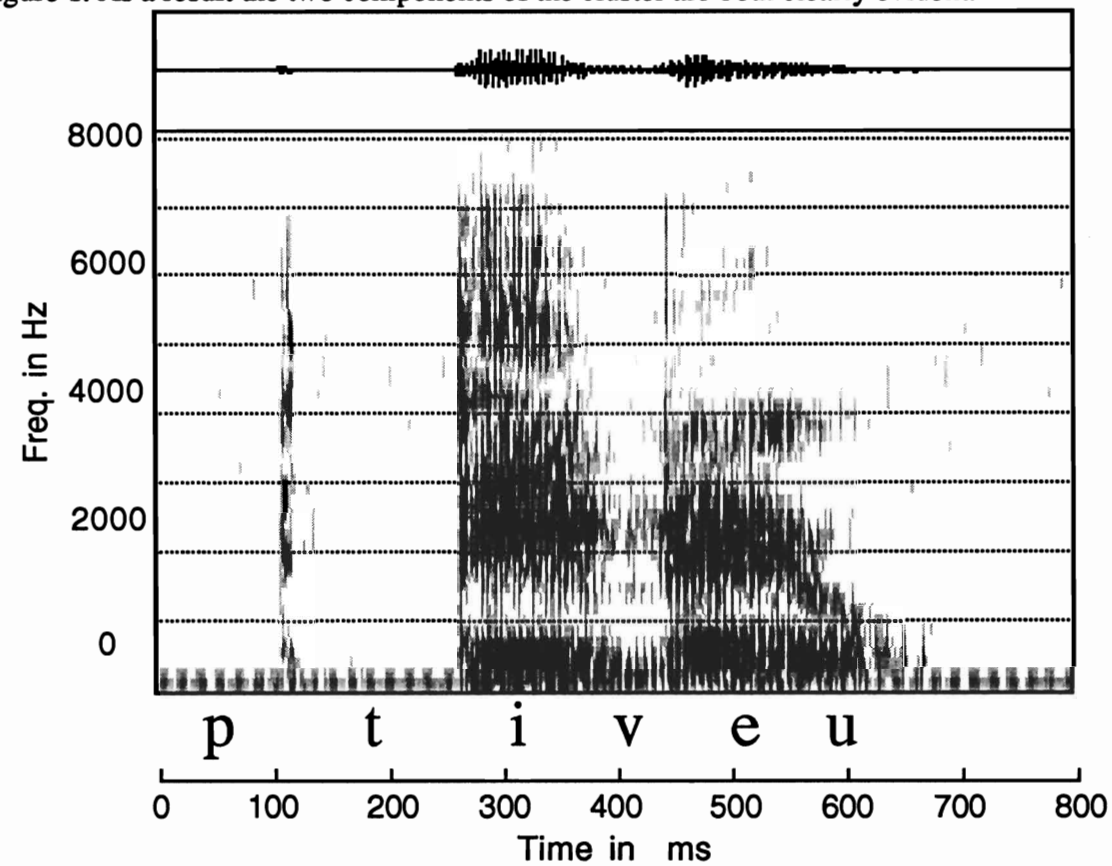


Figure 1. Example spectrogram showing the characteristic voiceless release of voiceless stop clusters in **ptiveu** 'a kind of reed'.

The most unusual Tsou clusters involve **h** and **ʔ**. The upper part of Figure 2 illustrates the clusters **ʔs** and **sʔ**. In the first utterance the initial glottal stop is realized as a couple of creaky glottal pulses. On other occasions an initial glottal stop is just an abrupt onset to the following consonant. In the second picture the glottal stop follows a fricative. On some occasions (although

not on this occasion), this sort of combination is realized as an ejective fricative. On this occasion there is a comparatively long glottal stop forming an abrupt onset to the vowel. Perhaps the most exceptional of all are the clusters with initial **h**, illustrated in the lower part of Figure 2. The initial aspiration is some what stronger than usual in these two examples (chosen so that the acoustic characteristics were clearly visible). Before the alveolar stop in the utterance on the left there is a movement so that the acoustic structure becomes more like that associated with the alveolar burst. In the utterance on the right, during the aspiration before the glottal stop the vocal tract assumes the position for the following vowel.

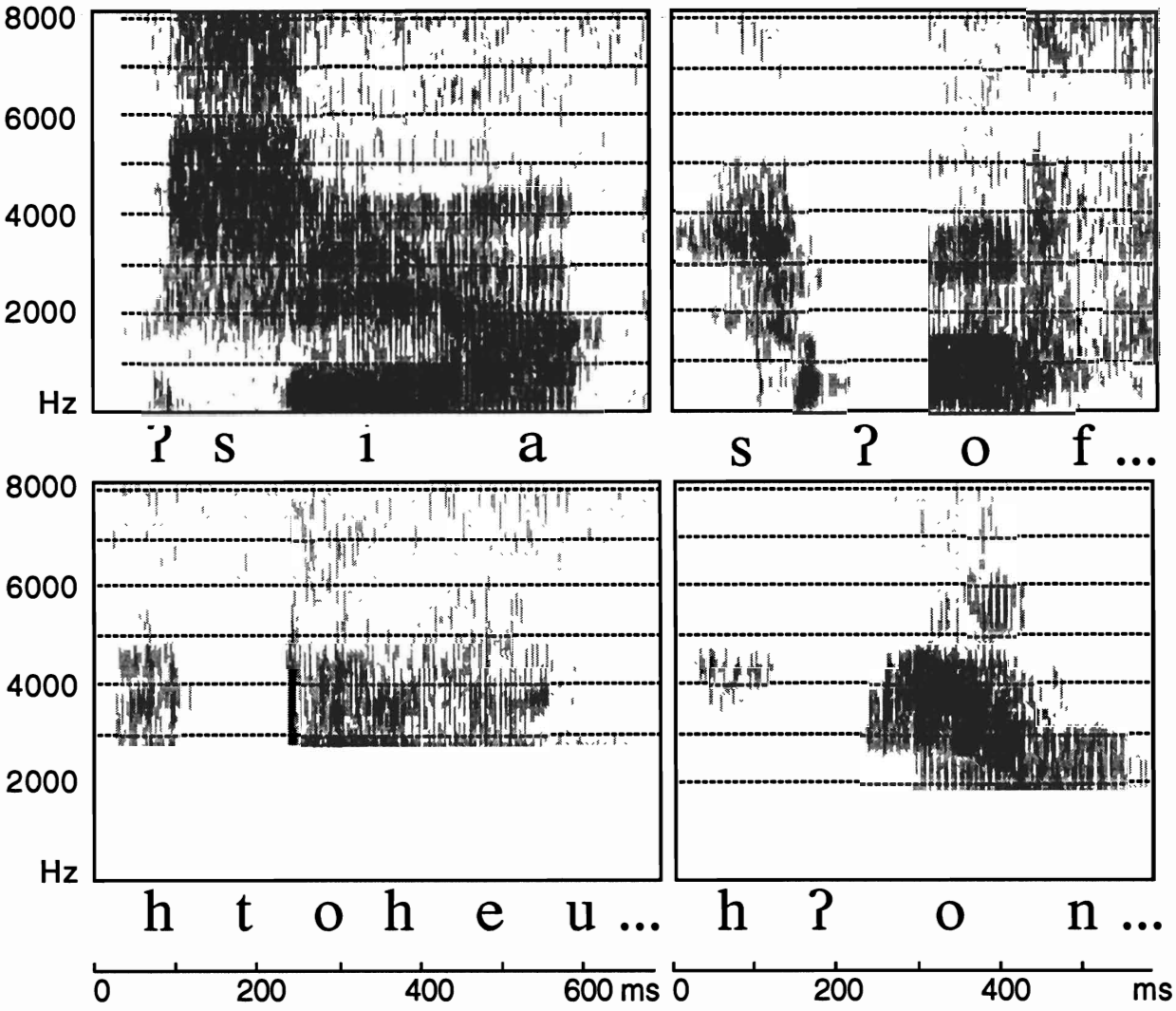


Figure 2. Spectrograms of the Tsou words: **ʔsia** 'September', **sʔofu** 'stick', **htoheuŋsa** 'repeated', **hʔoni** 'liver'

The investigation of these Tsou consonant clusters also illustrates another reason for the necessity of fieldwork in the location where a language is spoken. We need to hear a number of speakers. People speak differently from one another, and a valid phonetic study must be based on analyses of several people. It is clear that it is no longer appropriate to present data based on instrumental records of only a single speaker, as some misguided phoneticians have done in the past (Ladefoged 1968). A previous account of Tsou (Fuller 1990) was based on the speech of a single speaker who was a student in the U.S. at the time. This study reported with understandable

excitement that Tsou had a labiodental pulmonic ingressive fricative **f←**, so that there were words such as **f←usu** 'hair'. If this had been correct it would have been the first documentation of these sounds in any language; and a video recording of the speaker studied showed that it was undoubtedly correct that he did have pulmonic ingressive fricatives in a number of words. But a subsequent investigation (Ladefoged and Zeitoun 1993) in the speaker's home village showed that neither of his parents nor 12 other speakers produced these words in this way. They all used the cluster **f?**, sometimes producing it as a pulmonic fricative followed by a glottal stop, and sometimes as an ejective **f'**. It is important to record a number of speakers who use the language every day. One speaker may use a particular sound; but it does not follow that sound is part of the language.

Fieldwork on another continent also has a bearing on the study of onset consonant clusters containing two stops. Data from Zhuloãsi, a Khoisan language spoken in Namibia, shows that the two stops need not agree in voicing nor in airstream mechanism. Fieldwork illustrating this point was reported by Snyman (1975), who presented considerable phonetic detail. However, the precise mechanisms involved were not known. Working with Jan Snyman, we (Peter and Jenny Ladefoged) undertook further instrumental phonetic fieldwork.

Figure 3 shows waveforms illustrating Zhuloãsi examples in which the first consonant is voiced and the second voiceless. In the top row the initial voiced plosive (with a pulmonic airstream mechanism) is followed by another plosive, in the second it is followed by an ejective (with a glottalic airstream mechanism), and in the third by a click (with a velaric airstream mechanism). The first stop is not released in this language. All the two stop clusters in Zhuloãsi involve the conjunction of homorganic stops, so little additional information concerning the place of articulation would be produced by releasing the first consonant. The duration of the voicing in the first consonant may depend on the airstream mechanism of the following sound. There are four glottal vibrations preceding the voiceless plosive in the top row; before the ejective in the second row there are about 50 ms of regular voicing preceding the creaky voice associated with the glottal closure for the ejective; and before the click in the last row there is a considerable period, over 100 ms, of regular voicing. We have records of five speakers producing these sounds, but we have comparatively few words, so we do not have adequate data for a statistical evaluation of these differences.

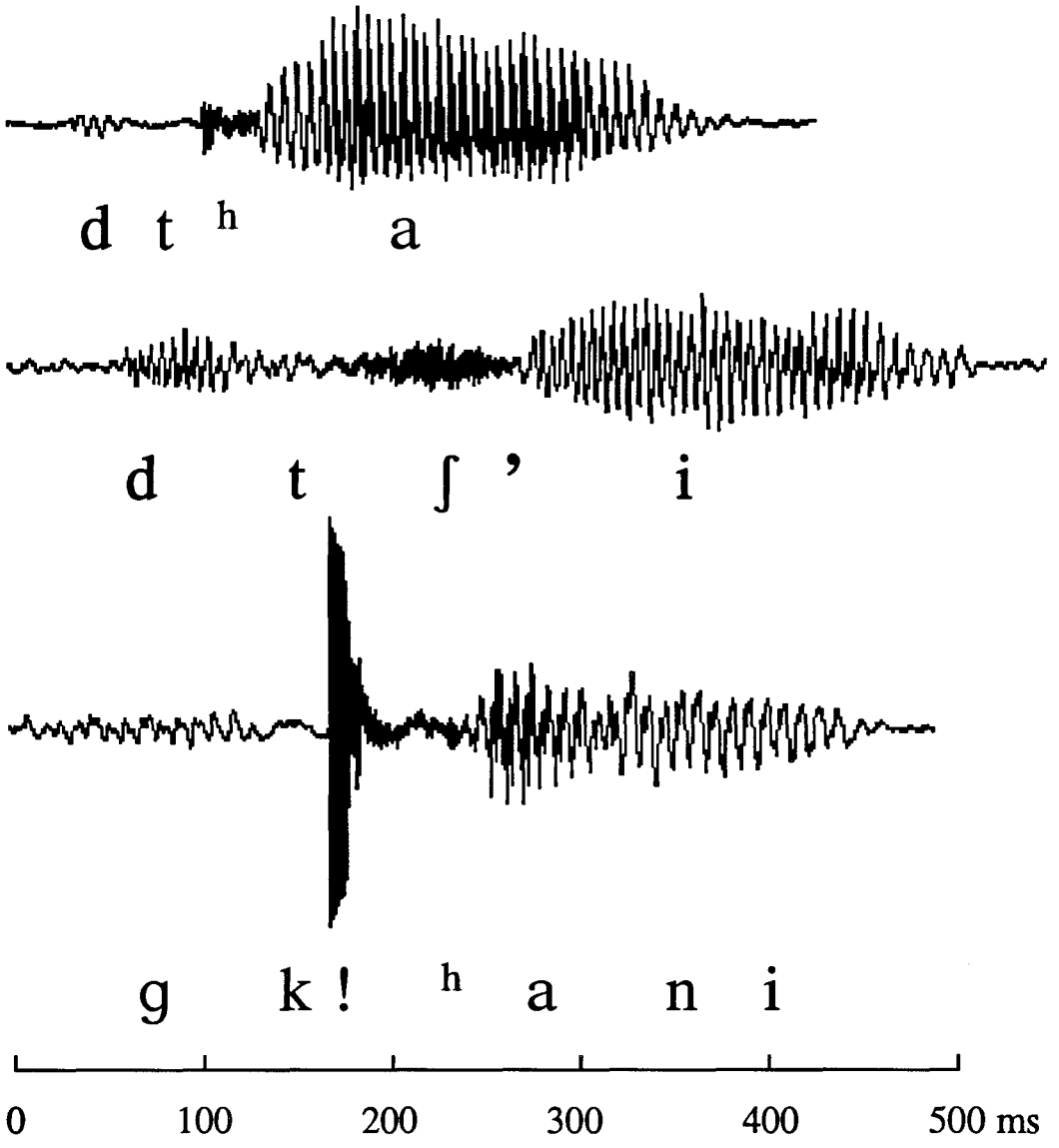


Figure 3. Three examples of consonant clusters with mixed voicing and mixed airstream mechanisms in Zhuloãsi.

Data from Toda illustrates another phonological problem. It is not at all clear how many distinct places of articulation have to be taken into account by a phonological feature theory. Toda has three series of consonants which in some sense exhibit 7 places of articulation. Emeneau (1984), the leading authority on Toda (and, in fact, the only linguist who had done fieldwork on this language) labeled these consonants as shown in Table 2.

Table 2. Part of a chart showing the consonants of Toda, using IPA symbols, but labels as in Emeneau (1984).

	Labial	Dental	Post-dental.	Alveolar	Retroflex	Alveolo-palatal	Velar
Voiceless stop	p	t̪	ts	t	ʈ	tʃ	k
Voiced stop	b	d̪	dz	d	ɖ	dʒ	g
Fricative	f	θ	s̪	s	ʂ ʐ	ʃ ʒ	x

In a preliminary report, Ladefoged and Maddieson (1986), who had not heard Toda themselves, wrongly re-interpreted Emeneau's data and suggested that Toda might contrast two types of retroflex consonants. This notion was based on the findings of Ladefoged and Bhaskararao (1983), who had shown that the retroflex consonants in Hindi and Telugu were phonetically different, those in Hindi being made predominately with the tip of the tongue contacting the posterior part of the alveolar ridge, and those in Telugu using the underside of the tip of the tongue on the hard palate behind the alveolar ridge. Ladefoged and Maddieson reinterpreted Emeneau's work, and suggested that Toda might make a phonemic difference between these two kinds of retroflex consonants.

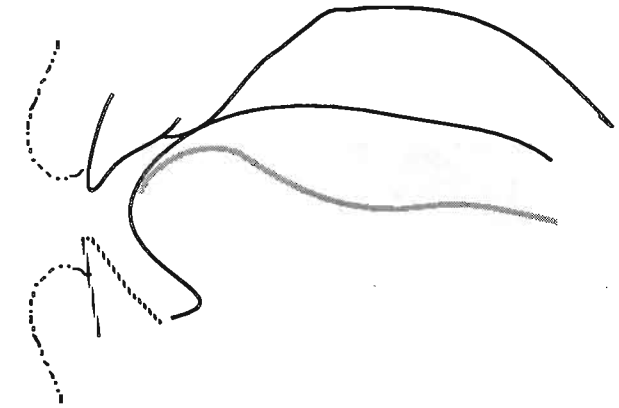
A subsequent field trip demonstrated the importance of obtaining direct articulatory data (Shalev, Ladefoged and Bhaskararao 1993). There were not two degrees of retroflexion, and the stops were substantially as Emeneau (1984) had described them. However there are only six places of articulation among stops, in that both **ts** and **t̪** and their voiced counterparts are laminal denti-alveolars, the former being distinguished by being more affricated.

The Toda fieldwork provided data on the fricatives that was even more interesting, and worth considering in detail. We will concentrate on the sibilants, as there is little that need be said about the non-sibilant fricatives, **f**, **θ**, **x**, which are labiodental, dental and velar as one might expect. Toda has four different articulatory gestures for sibilant fricatives, whereas the other Dravidian languages have only three. Contrasts are shown in Table 3.

Table 3. Words illustrating contrasts among Toda sibilants.

LAMINAL ALVEOLAR		APICAL POST-ALVEOLAR		LAMINAL POST-ALVEOLAR		SUB-APICAL PALATAL	
koːs̪	'money'	**poːs̪**	'milk'	**poːʃ**	'language'	**poːʂ**	(place name)

Figures 4-7 show palatograms, linguograms and sagittal reconstructions of the four sibilants from one speaker. As the shape of the palate is known from dental impressions, the height of the contact at the sides can be observed on the palatograms. We have used this information to make a rough estimate of the position of the center of the tongue (shown by a dotted line). Note that the dark marks on the speaker's front teeth in these photographs are betel juice stains, and are not the result of tongue contact.



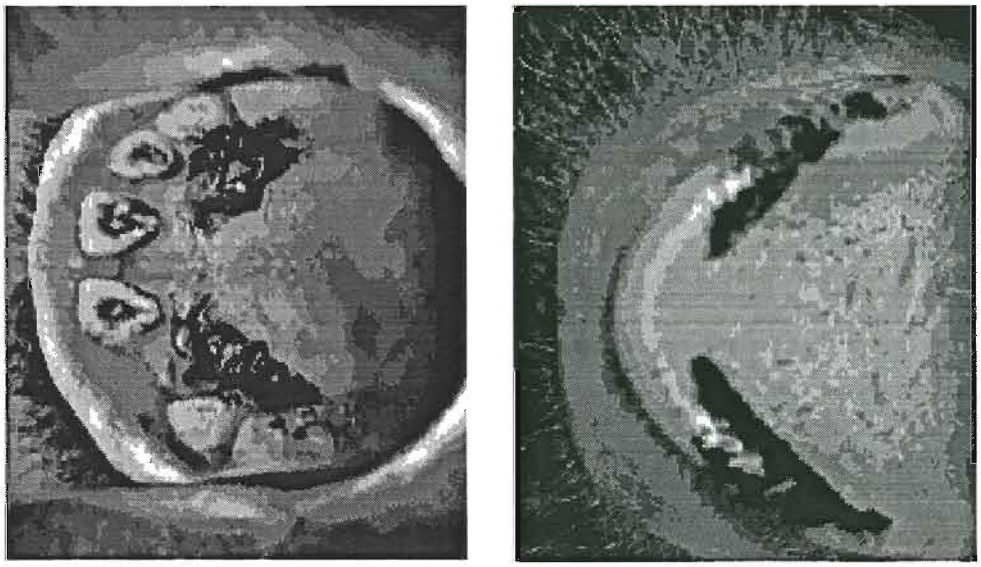
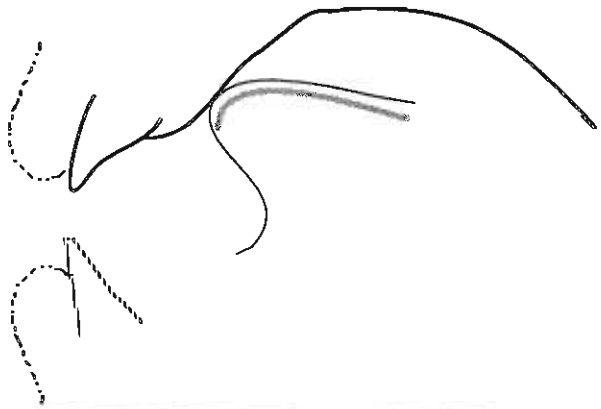
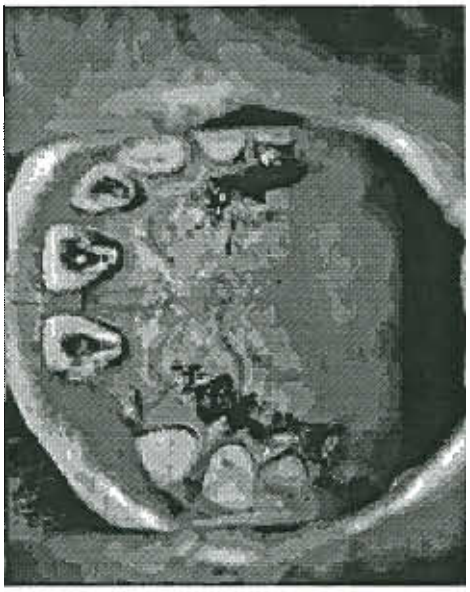


Figure 4. Palatographic and linguographic records of **koːs̪** 'money'. The solid lines on the sagittal section show the known shape of the palate and the observed position of the sides of the tongue. Dashed lines indicate estimated positions of the lips and the center of the tongue.

Toda s̪ has a laminal alveolar articulation. Emeneau (1984) describes it as being "post-dental (pre-alveolar)", and we agree in that the constriction might be said to be closer to the teeth than in English **s**, but it is in the alveolar region. What is most remarkable about this sound is that it is clearly laminal, but nevertheless there is only a narrow part of the blade of the tongue contacting the roof of the mouth. Similar narrow contact areas on the blade of the tongue were observed for all three subjects for whom we have palatographic records. The sides of the tongue touch the hard palate well above the level of the molar teeth. We do not know whether the center of the tongue is hollowed, but from the fact that the distance between the points of contact on the alveolar ridge is slightly smaller than the distance between the comparable points on the tongue, and from direct observation of the production of this sound, we believe that the tongue is slightly grooved, so that it might be in the position shown in the upper part of Figure 4.



poːs̱ "milk"



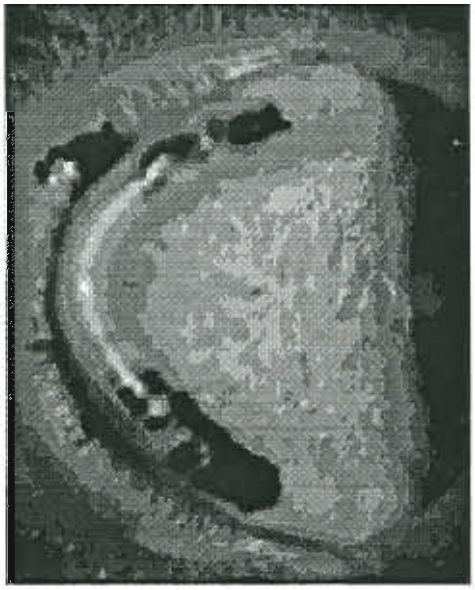


Figure 5. Palatographic and linguographic records as in Figure 4 of Toda **poːs̱** 'milk'.

The laminal contact for **s̱** contrasts with the apical contact for **s̱** in the same language. This sound, which is illustrated in Figure 5, is an apical sibilant. The two sounds also differ in that **s̱** has a wider channel for the airstream, and is articulated slightly further back, on the center of the alveolar ridge, making it post-alveolar. The contact areas at the side of the mouth are closer to the molar teeth, indicating a generally lower position for the tongue. There may also be some hollowing of the tongue in this sound, but it is not as extensive as in **s**.

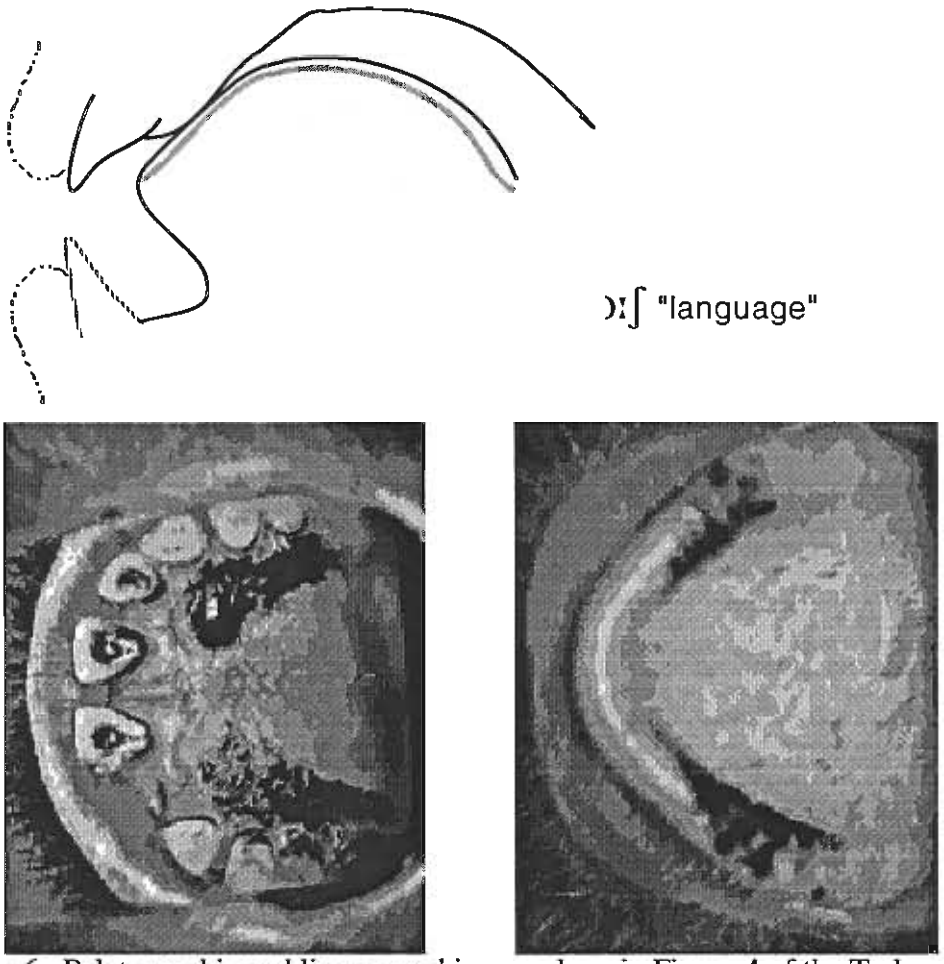
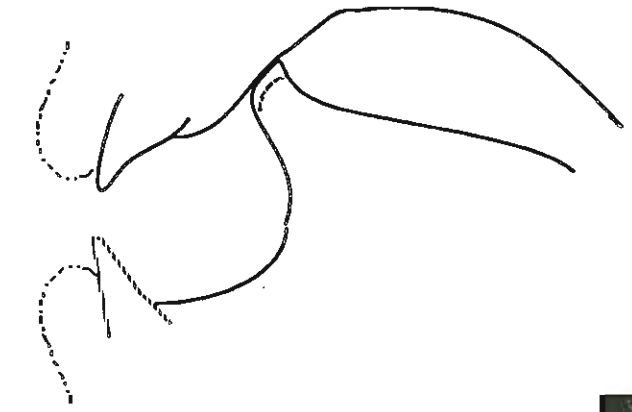
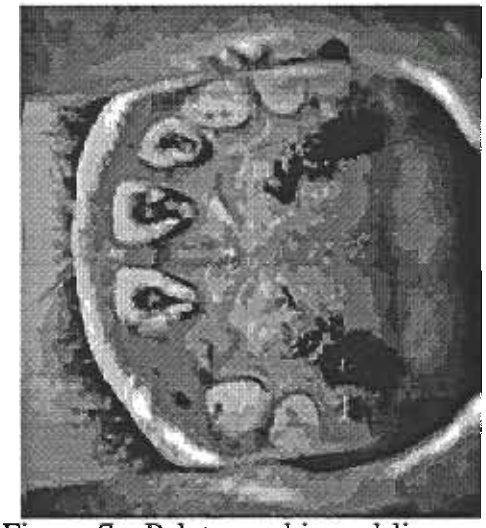


Figure 6. Palatographic and linguographic records as in Figure 4 of the Toda word **poɪʃ** 'language'.

The third Toda fricative, **ʃ**, is a laminal post-alveolar, i.e. a palato-alveolar sibilant, with more contact of the tongue with the palate than either of the preceding sibilants, as can be seen in Figure 6. The laminal tongue contact is similar to that in Toda **s**, but involves a narrower channel, with the tongue sides being much higher in the mouth. The tongue is domed up towards the roof of the mouth, in a way somewhat similar to that in English **ʃ**.



poːʂ "clan"



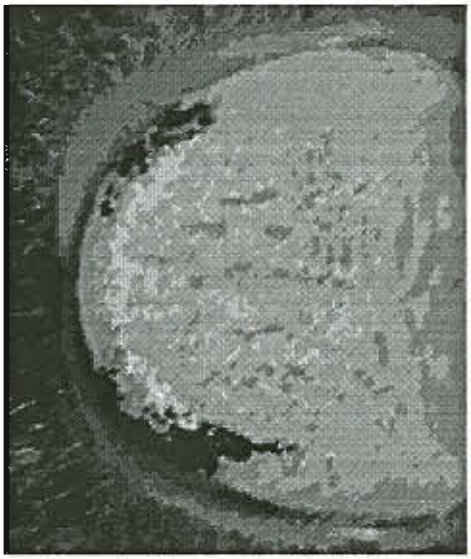


Figure 7. Palatographic and linguographic records as in Figure 4 of Toda **poːʂ** (place name).

The final sibilant in Toda is ʂ, a sub-apical palatal fricative, a genuinely retroflex gesture, is which illustrated in Figure 7. The contact is between the underside of the tip of the tongue (so that it is not all visible in the linguagram) and a point on the roof of the mouth behind the post-alveolar region.

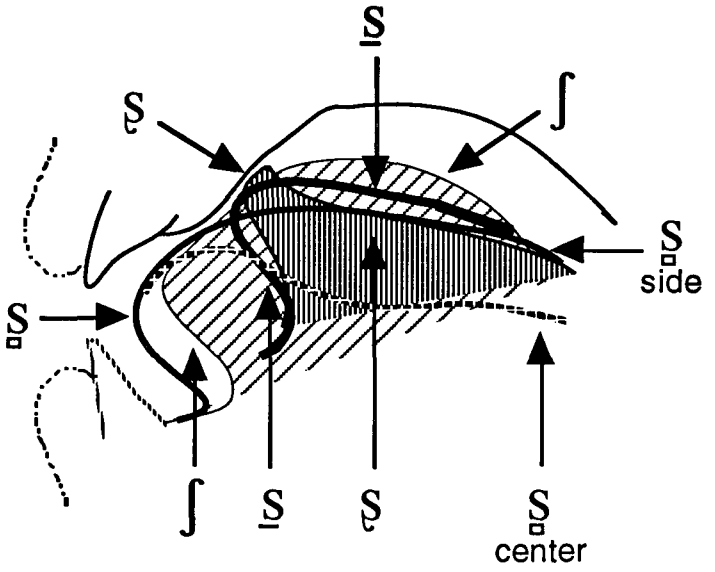


Figure 8. A composite diagram of Toda tongue positions in sibilant fricatives. In the case of s the center and sides of the tongue have been shown separately. The position of the center and sides is taken to be much the same for the other sounds, and only one tongue line is shown for each of them.

So that all these Toda sibilant fricatives might be compared, the tongue positions shown in the preceding diagrams have been superimposed in Figure 8. The major point to note is that only one of these sounds, the sub-apical palatal (retroflex) sibilant, can be readily described in terms of its place of articulation. Each of the others involves subtle distinctions in tongue shape relative to the teeth, but they do not reflect the usual articulatory distinctions associated with these terms. It is not at all clear how a theory of phonology should relate these phonetic differences to a universally applicable set of features.

Finally, we will consider another example of the IPA problem of deciding what counts as a distinct sound. Do we need to recognize the voiceless nasals of Burmese as distinct from the voiceless nasals in Angami? Burmese and Angami are both Tibeto-Burman languages, the former spoken Myanmar and the latter in the Naga Hills in the northeastern parts of India. Both contrast voiced and voiceless nasals; but the voiceless nasals are phonetically different.

Using instrumentation as described in Ladefoged (1993), we recorded the oral and nasal airflow during the pronunciation of 6 Burmese speakers. A typical recording of a word beginning with a voiceless nasal consonant is illustrated in Figure 9 (based on Bhaskararao and Ladefoged, 1991). At the time indicated by arrow (1) the oral airflow increases, forming a breathy as the glottis opens. By time (2) the oral airflow has ceased, and there is only voiceless nasal airflow. At time (3), while air is still coming exclusively out of the nose, voicing starts, so that there is a short interval with a voiced nasal before the vowel. At time (4) the articulators come apart and the vowel begins.

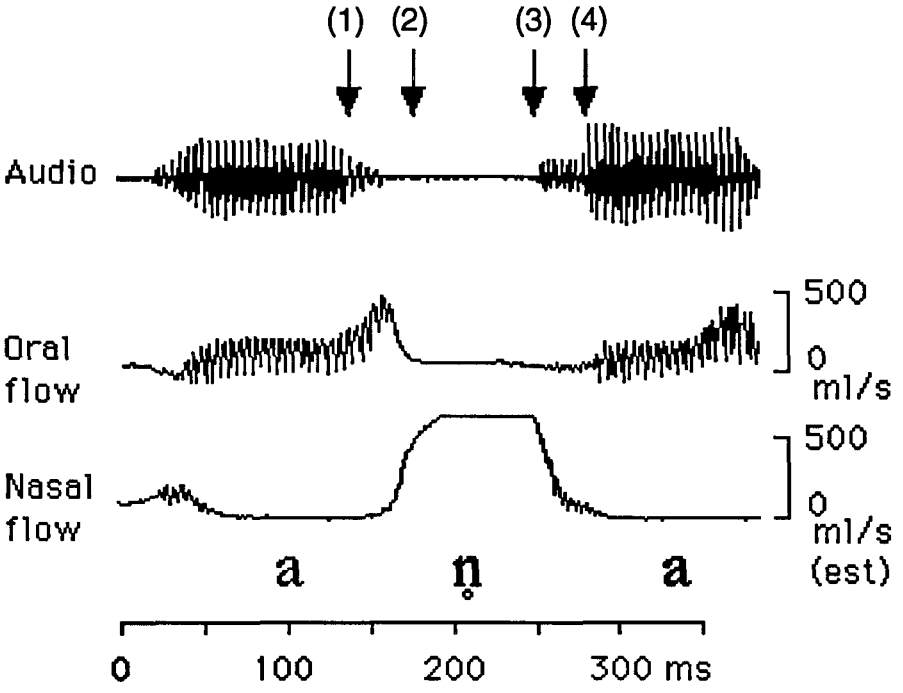


Figure 9. Aerodynamic records of the Burmese word [ŋ̊a] 'nose'.

Using the same instrumentation we also recorded the oral and nasal airflow during the pronunciation of nine Angami speakers. An example of an Angami voiceless nasal is shown in Figure 10 (based on Bhaskararao and Ladefoged, 1991). At time (1) the alveolar articulation is formed, and voicing ceases after a few more vibrations of the vocal cords . Between times (1) and (2) the voiceless nasal airflow increases for this word initial consonant. Then, at time (2), the articulators open and there is a sudden rapid flow of air from the mouth. As air can escape from the mouth, the airflow from the nose drops, but the velum is still lowered so that there is still a considerable flow of air through the nose. The sharp increase in oral airflow in the voiceless alveolar nasal gives rise to the auditory impression of an epenthetic voiceless alveolar plosive. But, although the nasal airflow drops at this moment, it still remains at about 500 ml/s. At time (3) voicing starts, probably with somewhat breathy vibrations, as there is a high rate of airflow through the mouth. For all of the nine speakers of Angami we recorded, oral airflow began about half way through the voiceless section. Unlike the Burmese sounds, in which there was always some voicing during the last part of the nasal, in Angami there was never any voicing during the nasal. This is therefore a voiceless aspirated nasal.

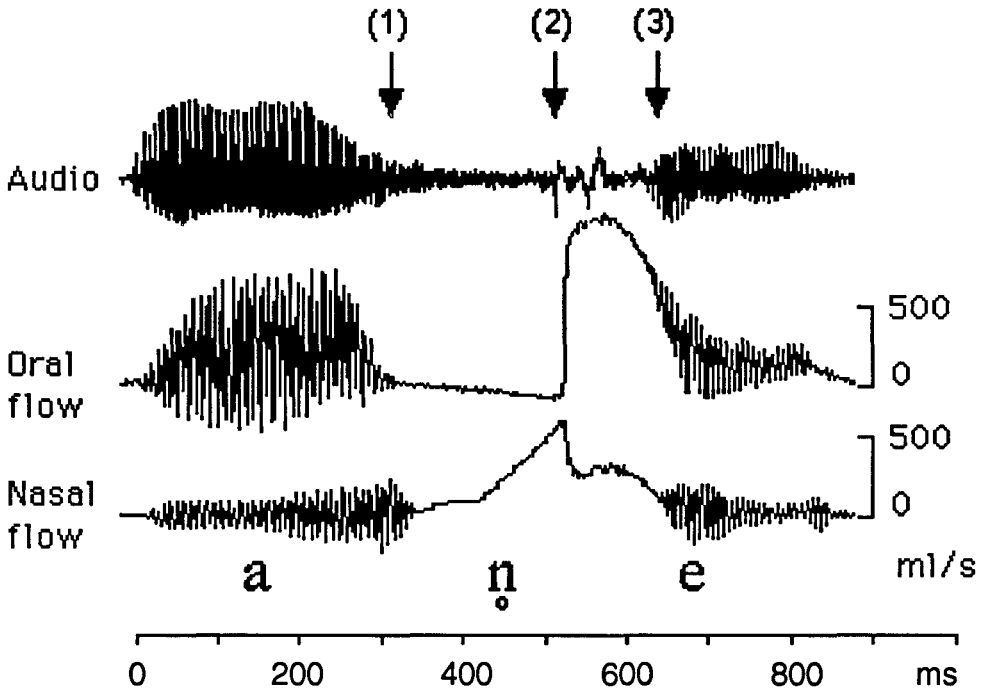


Figure 10. Aerodynamic records of the Angami word [n̥e] 'to blow one's nose'.

Stops in many languages have a three-way voicing contrast: voiced, voiceless unaspirated, and voiceless aspirated. In a few languages, including some forms of Burmese, a similar three-way contrast occurs among fricatives. There may well be a language that has a three way contrast among nasals: voiced (as in both Burmese and Angami), voiceless unaspirated (as in Burmese), and aspirated (as in Angami).

Conclusion

The general conclusion from fieldwork such as that reported above is that linguists underestimate the wide variety of sounds that occur in the world's languages, making their accounts of what constitutes a possible linguistic contrast inadequate. This fieldwork also shows the necessity of obtaining data from several speakers, in that errors arise by relying on observations of a single speaker who may not be producing relevant linguistic data. That theoretical advances and practical fieldwork go hand in hand seems clear. A language is the shared property of a group of individuals, who behave in complex and often unexpected ways. Not even one's own language can be adequately observed from an armchair.

Acknowledgements

Major thanks are due to all the many speakers of the different languages cited in this paper. I have also had considerable assistance from my colleagues in the UCLA phonetics lab.

References

Bhaskararao, P. and P. Ladefoged (1991). "Two types of voiceless nasals." Journal of the International Phonetic Association **21**(2): 80-88.

Emeneau, M. B. (1984). Toda Grammar and Texts. Philadelphia, American Philosophical Society.

Fuller, M. (1990). "Pulmonic ingressive fricatives in Tsou." Journal of the International Phonetic Association **20**(9-14):

Ho, D.-A. (1976). "Tsou phonology." Bulletin of the Institute of History and Philology, Academia Sinica **47**: 245-274.

Ladefoged, P. (1968). A phonetic study of Western African Languages. Cambridge, Cambridge University Press.

Ladefoged, P. (1993). "Linguistic phonetic fieldwork: a practical guide." UCLA Working Papers in Phonetics **84**: 1-24.

Ladefoged, P. and P. Bhaskararao (1983). "Non-quantal aspects of consonant production: A study of retroflex consonants." Journal of Phonetics **11**(3): 291-302.

Ladefoged, P. and I. Maddieson (1986). "Some of the Sounds of the World's Languages: (Preliminary version)." UCLA WPP **64**:

Ladefoged, P. and I. Maddieson (in press). The Sounds of the World's Languages Oxford, Blackwell.

Ladefoged, P. and E. Zeitoun (1993). "Pulmonic ingressive phones do not occur in Tsou." Journal of the International Phonetic Association **23**(1): 13-15.

Li, P. J.-K. (1979). "Variations in the Tsou dialects." Bulletin of the Institute of History and Philology, Academia Sinica **50**(245-74):

Shalev, M., P. Ladefoged and P. Bhaskararao. (1993). "Phonetics of Toda." UCLA Working Papers in Phonetics **84**: 89-126.

Snyman, J. W. (1975). Zul'hoasi fonologie en woordeboek. Cape Town, Balkema.

Starosta, S. (1974). "Causative verbs in Formosan languages." Oceanic Linguistics **13**: 279-369.

Tsuchida, S. (1976). Reconstructionof Proto-Tsouic phonology. Tokyo, Tokyo University of Foreign Studies.

Tung, T.-h. (1964). A Descriptive Study of the Tsou Language. Taipei. Institute of History and Philology, Academia Sinica Special Publication.

Wright, R. and P. Ladefoged. (1994). "A Phonetic Study of Tsou." UCLA Working Papers in Phonetics **87**.

Phonetic studies of American Indian languages

Peter Ladefoged and Victoria Fromkin

The study of Native American Indian languages played and continues to play a major role in the development of American linguistics and linguistic theory in general. A description of Narragansett, an Algonquian language, was published as early as 1643 (Williams, 1963). At the end of the 19th century, J.W. Powell's *Indian Linguistic Families North of Mexico* was important for its contributions to lexicography, but it is generally agreed that the main linguistic contributions of such studies began with the seminal work of Franz Boas.

Boas placed emphasis on careful description of the speech sounds of native languages stating that "One of the most important facts relating to the phonetics of human speech is that every single language has a definite and limited group of sounds" and the investigator has the responsibility to describe this set meticulously as it may and probably will differ from other even closely related languages (Boas, 1911. P 16). His phonetic approach is discussed in *The Handbook of American Indian Languages* (1911) where he provides a methodology for describing the articulation of speech sounds. Using this approach he lists seventeen separate vowels in Kwakwala (Boas 1947) only six or seven of which are phonemic (Swadesh, 1948). Boas provides descriptions of a large number of native languages in addition to Kwakwala, including Keresan, Kutenai, Tsimshian, Bella Bella, Bella Coola, Chinook, Dakota, Salishan, and Haida.

Such an approach had an important influence on Boas's most distinguished student, Edward Sapir, whose studies of American Indian languages contributed decisively to the formulation of his theory of language, including his views on phonology and phonetics. Sapir's American Indian studies were as widespread as Boas's and in many cases more detailed. They include Wishram, Chinook, Takelma, Comox, Nootka, Yana, Navajo, Southern Paiute, Ute, Tutelo, Chimariko, Haida, Sarcee, Kutchin, and Hupa. Sapir was more interested in 'the psychological reality of phonemes' (1949) than Boas and in distinguishing between the set of contrasting abstract phonological segments and their phonetic realizations. But his theory of phonology constrained the set of phonemes of any language to be a subset of the phonetic units of that language, and the units of both inventories – the phonemic and the phonetic – were fully specified as to articulatory phonetic features (Sapir 1921).

Bloomfield – probably the most important pre-Chomsky figure in the history of American linguistics – in his studies of Menomini, Cree, and Fox, further illustrates the role played by Native Indian languages. It is interesting that although he was of the opinion that "the physiologic and acoustic description of acts of speech belongs to other sciences than ours," (Bloomfield, 1926, 153) and believed that phonetics "gives us a purely acoustic or physiologic description (revealing) the gross acoustic features" of speech while phonology "pays no heed to the acoustic nature of the phonemes, but merely accepts them as distinct units" (1933, 137), his work shows a practical reliance on phonetic data. Furthermore, he defined a linguistic form as "a **phonetic** (our emphasis) form which has a meaning." (1933, 138).

Detailed and rich descriptions of the sound systems of American Indian languages were carried on by Sapir's students including Morris Swadesh, Stanley Newman, Mary Haas, Harry Hoijer, and C.F. Voegelin, and by his students' students such as Bill Bright. Bright's major work on Karok (1957) as well as his studies of other American Indian languages such as Cahuilla (1967b), Cupeño (1967a), and Luiseño (1968) continue the Sapirian tradition.

The descriptions of the sounds of these languages remain important in the attempt to delineate all (and only?) the speech sounds of the languages of the world. More recently, comments on phonetic properties of American Indian languages have been less common, but still of importance from a general phonetic point of view. An often cited study is that of Bright (1978), which pointed out that a considerable number of the languages of California contrast **s̱** with **ṣ**. Bright noted that Karok has minimal word-pairs like **s̱úːf** 'creek' vs. **ṣúːf** 'backbone', describing the sound at the beginning of the first of these words as being "a very far-forward, apico-dental sound ... pronounced by younger speakers as **θ**." He describes the sound at the beginning of the second word as "apico-alveolar," and further identifying it as a "retracted ess". Langdon (1976) and Langdon and Silver (1984) have also discussed similar phonetic problems, such as the contrast in Luiseño between **s̱úkat** 'deer' vs. **ṣukmal** 'fawn'.

This paper will show how phoneticians can build on observations such as these, and use them to investigate issues of phonetic theory. There are two steps involved, firstly recording a number of speakers, and secondly using appropriate techniques of instrumental phonetic analysis. As an illustration of what can be done, we will begin by considering a UCLA Phonetics Lab study that was based on Bright's original observations in which Sarah Dart investigated the contrasting sibilants in 'O'odham.

Dart (1991) made palatograms and linguagrams of 8 speakers of 'O'odham as illustrated in Figure 1. The picture on the left shows a view of the roof of the mouth as seen in a mirror, after the tongue had been painted with a mixture of charcoal and olive oil and the word **ha'asa** pronounced. It is plain that the blackened tongue touched the roof of the mouth well behind the upper teeth in the region of the alveolar ridge. The outline of the upper incisors has been added to the photograph from other data. The picture on the right is the result of the reverse process. The roof of the mouth was painted with the black mixture, then the word was pronounced and the tongue protruded over the lower lip so that it could be photographed. Using digital editing techniques, the contrast between the blackened and non-blackened parts of the tongue was enhanced. In this case the blade of the tongue was used to form the narrow groove required for the sibilant. A full account of these palatographic techniques is given in Ladefoged (1993).

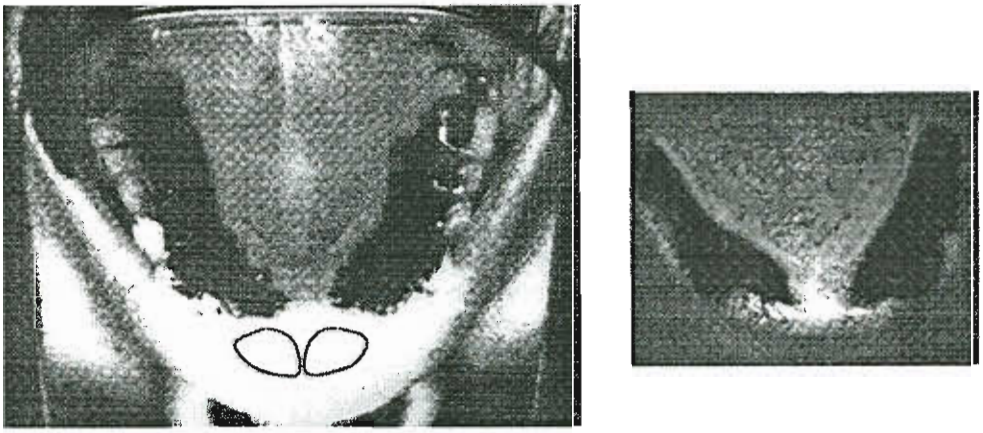


Figure 1. Palatogram and linguagram illustrating the more anterior of the two sibilants in 'O'odham. The ouline of the upper incisors has been added on the view of the palate on the left, and the contrast has been enhanced in the view of the tongue on the left.

Bright and Langdon and Silver were careful to point out that there is individual variation in the articulation of the sibilants in the languages they were discussing. Dart found that this was indeed the case. In comparing the words **haṣaba** 'but' and **ha'asa** 'the end', she found that for

four of the speakers the distinction was maximized and involved both the part of the tongue used and the place on the palate (laminal alveolar vs. apical post-alveolar). Two speakers had only a place of articulation difference (alveolar vs post alveolar). One speaker made the contrast by changing only the part of the tongue used (apical alveolar vs. laminal alveolar), and the last speaker made both fricatives as apicals, with one being in the front part of the alveolar region and the other in the back.

Findings such as these point to the relevance of auditory descriptions of phonological distinctions. What speakers are trying to produce are sounds that differ in certain specific auditory ways. The problem for phoneticians is that we do not know how to specify the auditory characteristics of fricatives. There is no agreed way of interpreting our acoustic descriptions of fricatives in auditory terms. However this research has shown that our articulatory descriptions are not entirely adequate in that some speakers produce these distinctions using different articulations from those of other speakers. We do not want to exaggerate this point and take the position that articulatory descriptions of speech sounds are therefore completely inappropriate. In the production of by far the majority of sounds there is only a very limited range of articulatory possibilities that can be used. Even in the case of these particular fricatives, it is not possible to produce the required contrasts by using widely differing articulations from those used by other speakers. But some variation is possible, and it would be good if phoneticians could specify the auditory targets required, and the range of articulations permitted.

The next phonetic issue we wish to consider is the nature of vowel systems. In this area there is much more agreement on how acoustic specifications should be made. Indeed, the situation is somewhat the reverse of that for sibilants, in that there are no measureable articulatory attributes of vowel quality. (We use terms that seem to specify the position of the highest point of the tongue, but we cannot quantify such descriptions.). We will demonstrate the use of acoustic techniques for specifying vowel quality by reference to Navajo. Note that the earlier phonetic descriptions of Navajo (Sapir, 1942; Hoijer, 1945) were based primarily on the 'good ear' of the investigator, whereas today we have the advantage of the experimental methods of the phonetics laboratory.

The vowels of Navajo have interesting phonetic characteristics. In general, when there are a comparatively small number of vowel qualities in languages, they are fairly uniformly distributed in the available acoustic vowel space (Lindblom 1986). This does not happen in Navajo, as has been shown by McDonough and Austin-Garrison (1994). Figure 2 shows their plot of the vowels of 5 speakers of Navajo, all of whom use virtually only Navajo in their daily life.

There are four pairs of long and short vowels. Each of the long vowels is distinguished from each of the other long vowels, and the same is true of the short vowels, which are by no means evenly distributed in the vowel space. The front vowels **i** and **e** are very close together, and, although the other two vowels are well separated, greater use could be made of the vowel space in that the back vowel is not as high as possible. We should also note that the long and short versions of the front vowels **i** and **e** are very significantly different in quality as well as in duration, but there is no such statistical difference in quality between the long and short versions of the other two vowels. There are no simple phonetic explanations for these facts; we presume that they are due to historical processes in Navajo.

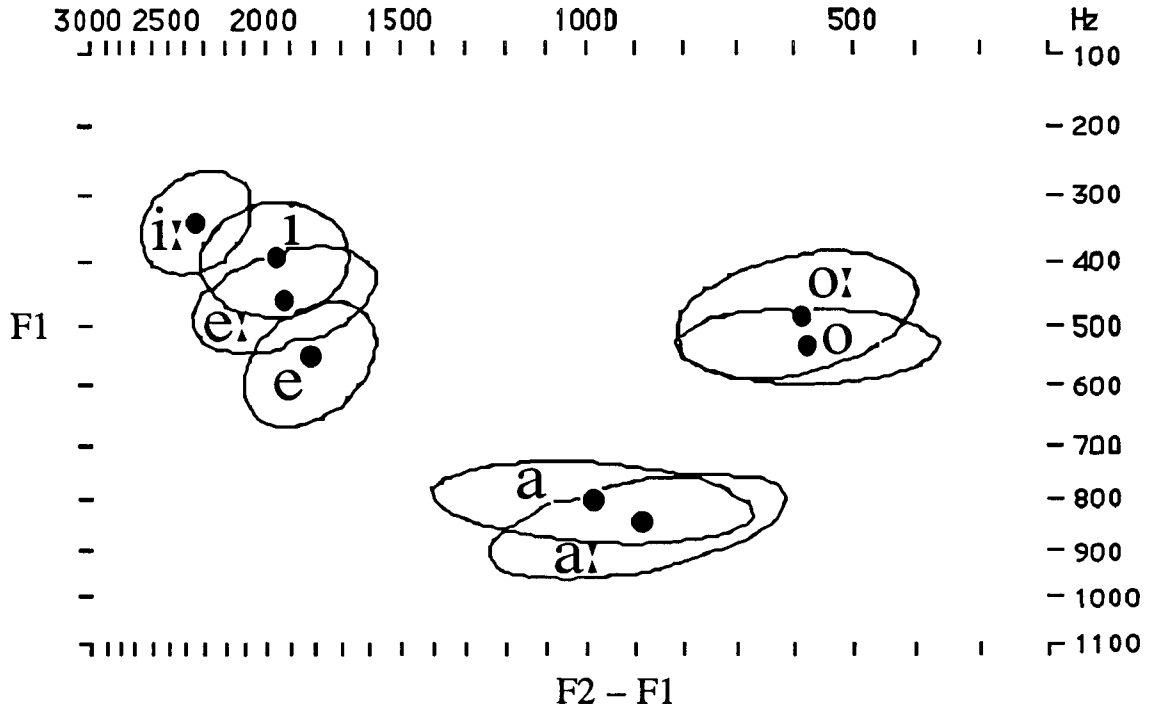


Figure 2. Mean locations of long and short vowels of 5 traditional speakers of Western Navajo. The ellipses ellipses enclose all points withing two standard deviations of the mean (McDonough and Austin-Garrison, 1994).

There is another aspect of vowels in American Indian languages that is of general phonetic interest. Vowels in languages such as Jalapa Mazatec can have different phonation types, as has been shown by Kirk,. Ladefoged, Ladefoged. (1993). According to one phonological analysis, there is a three way contrast between creaky (laryngealized), breathy (murmured) and modal (plain) vowels, as illustrated in Table 1 (Other analyses make the distinction depend on the presence of **h** and **ʔ**).

Table 1. Creaky (laryngealized), breathy (murmured) and modal (plain) vowels in Jalapa Mazatec.

MODAL VOICE	BREATHY	CREAKY
já	**ja̤**	**ja̰**
tree	he carries	he wears
ntʰǽ	**ndǽ̤**	**ndǽ̰**
seed	arse	horse

The acoustic cues distinguishing vowels with different phonation types have been described at length by Ladefoged, Maddieson and Jackson (1988). As a general rule, vowels with stiff voice or creaky voice have more energy in the harmonics in the region of the first and second formants than those with modal voice. Conversely, vowels with slack or breathy voice have comparatively more energy in the fundamental frequency. There is also a tendency (though not in all languages) for vowels with creaky voice to have a more irregular vocal cord pulse rate (more jitter), and for breathy voice vowels to have more random energy (a larger noise component) in the higher frequencies. These points can be seen in the narrow band power

spectra of the creaky, modal and breathy vowels of five speakers of Jalapa Mazatec, shown in Figure 3.

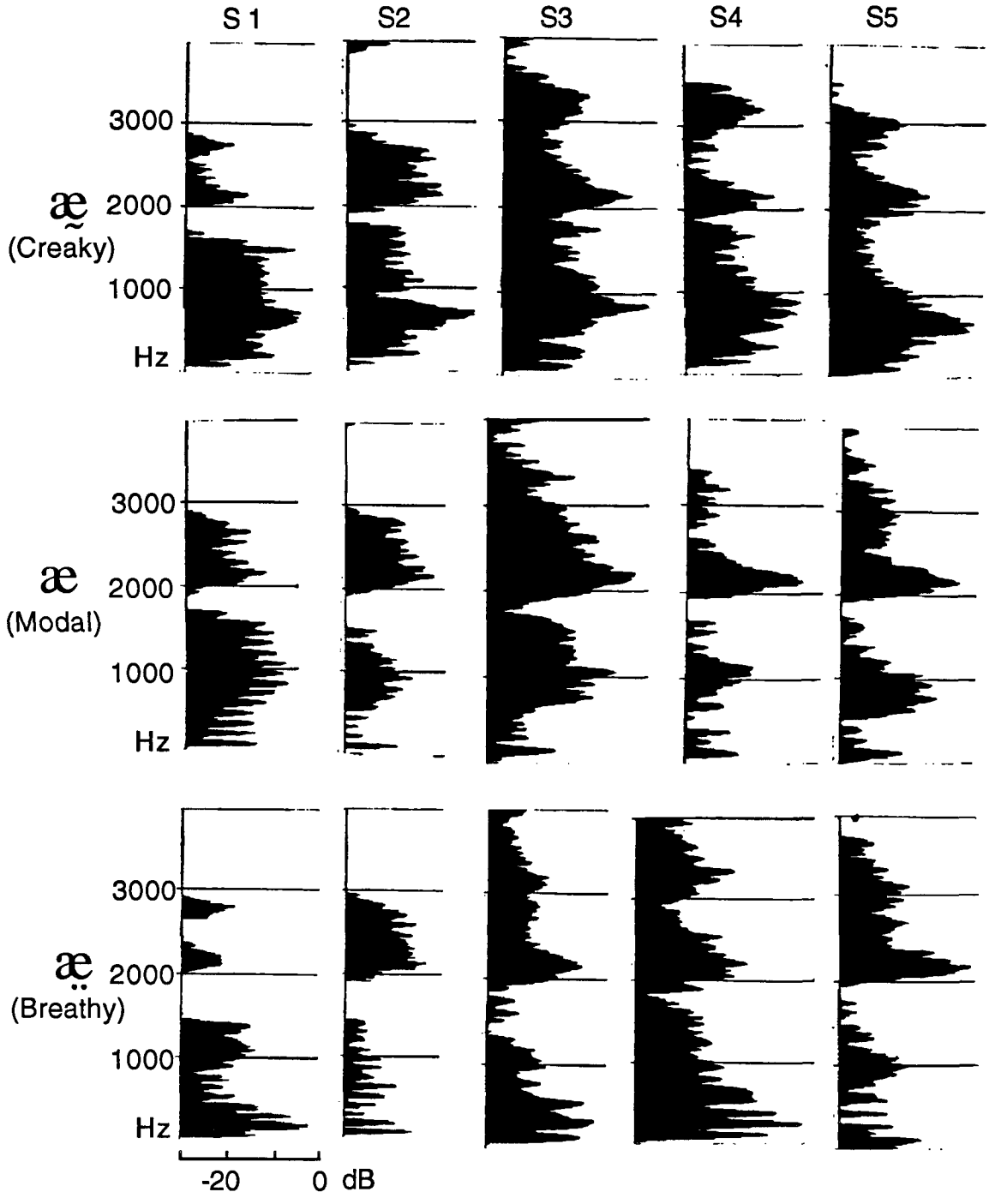


Figure 3. Narrow band power spectra of the creaky, modal and breathy vowels of five speakers of Jalapa Mazatec, taken during the middle of the vowels in the last row of Table 1.

Figure 4 shows the difference between the amplitude of the fundamental and that of the first formant in each of these spectra. It thus illustrates the way in which some of these differences can be expressed quantitatively. Data for each of the five speakers is shown separately, followed by the mean for all five. The lowest set of three bars shows that the mean difference in amplitude between the fundamental and the first formant for creaky voice (black bar) is -17 dB, i.e. the fundamental has 17 dB less amplitude than the first formant, which is thus considerably stronger. The mean for modal voice (shaded bar) is -7 dB, and that for breathy voice (white bar) is +5 dB (i.e. for breathy voice there is a comparatively large amount of energy in the fundamental rather than in the first formant). There is considerable variation from speaker to speaker in the value for each of the three phonation types, but for all speakers the value for breathy voice is higher than that for modal voice for any speaker. Creaky voice and modal voice show some overlap of values, but it is still true that for every speaker creaky voice has a relatively lower value than modal voice. The point of phonetic theory to note here is that voice quality is relative in just the same way as pitch: what is breathy (or high tone) for one speaker may be modal (or mid tone) for another.

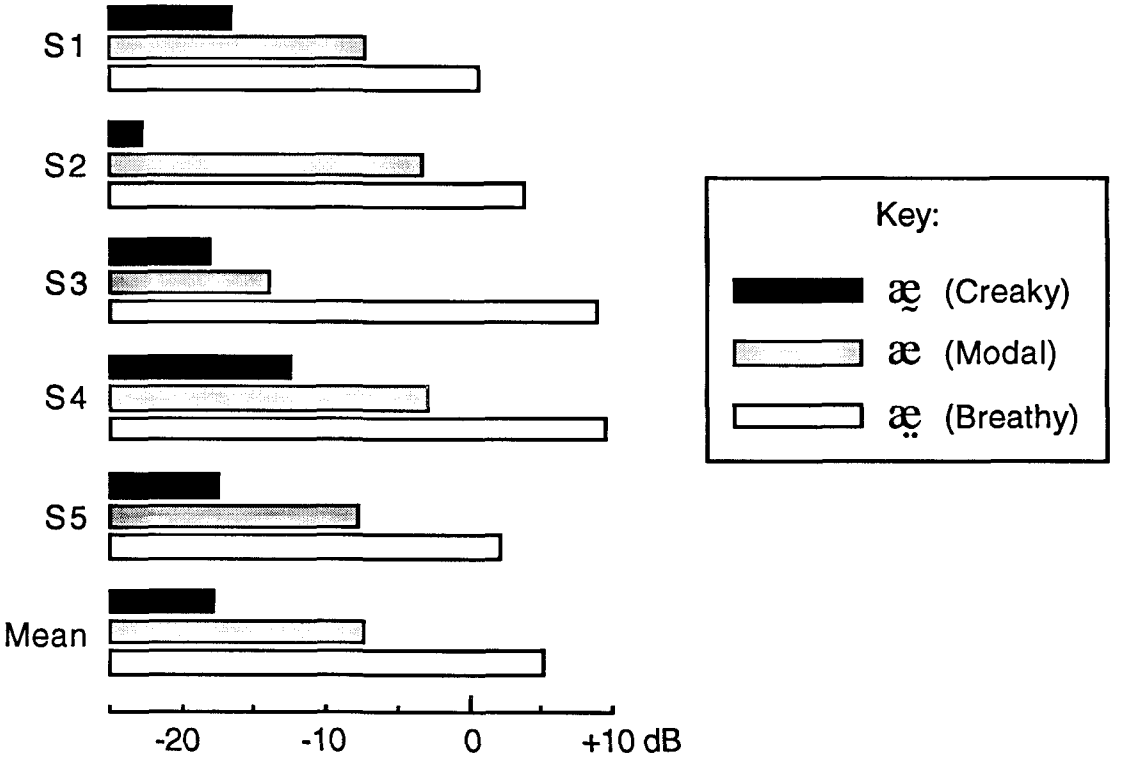


Figure 4 Relative amplitude of the fundamental and the first formant in Jalapa Mazatec vowels for five speakers and the mean of all five.

The last American Indian language we want to consider is Montana Salish. The phonetic structures of this language have been described by Flemming, Ladefoged and Thomason (1994). There are many points that could be taken to illustrate the theme of this paper. Here we will simply illustrate the analysis of ejective stops and pharyngeal approximants.

Figure 5 is an example of a set of aerodynamic records of the phrases **tsu páʕs** 'He said pale face' and **tsu p'əʕáp** 'He said the grass/timber caught fire'.. The top line is an electroglottographic (EGG) record from the larynx. This type of record cannot be quantified (except in the time domain), but it provides a good indication of the degree of closure of the vocal cords. The middle line is the oral pressure as recorded by a small tube inserted between

the lips, with its open end behind the alveolar ridge, and the third line is a record of the oral air flow. The arrows at the top show the moment of release of the bilabial closure. In the case of the ejective on the right of the figure, it may be seen that there is considerable laryngeal activity both at that time and slightly before it. This activity is followed by an interval of about 100 ms before vocal cord vibrations begin. There is far less activity for the plosive on the left of the picture, and vocal cord vibrations begin almost immediately. There is considerably greater oral pressure for the ejective in the second word than for the plosive in the analogous position in the first word. The plosive at the end of the second word also has less oral pressure. After the release of the plosive the oral flow rises to above 500 ml/s, whereas in the ejectives there is a comparatively small burst of oral flow, followed by a period in which there is no flow while the glottal closure is maintained. In this sound the vowel begins abruptly as the glottal closure is released.

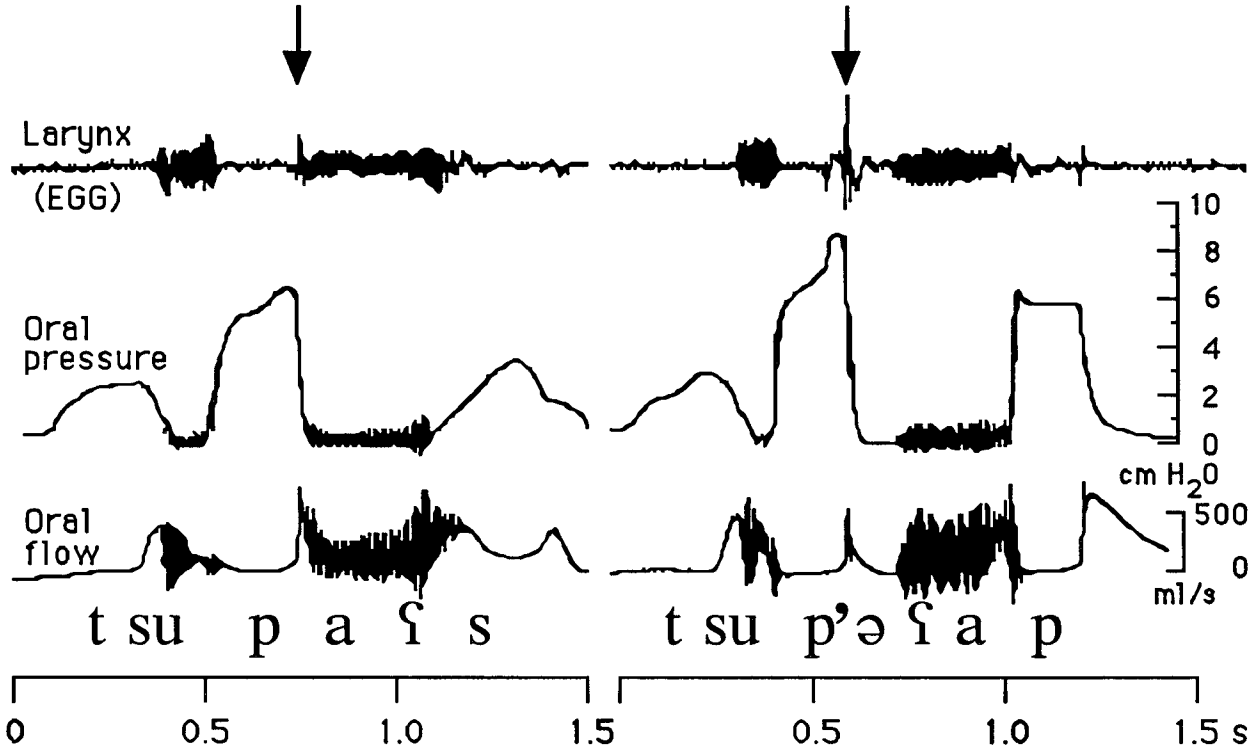


Figure 5. Aerodynamic records of contrasting plosive and ejective in Montana Salish **tsu páʕs** 'pale face' vs. **tsu p'əʕáp** 'the grass/timber caught fire'. (from Flemming, Ladefoged and Thomason, 1994)

The final point that we will consider is the investigation of pharyngeals in Montana Salish through the use of sound spectrograms. The pharyngeal consonants in this language are characterized by two different phonetic events, a retraction of the root of the tongue and a lowering of the pitch (which is probably due to a lowering of the larynx). Sometimes one of these possibilities occurs, sometimes both are evident, and sometimes the articulation is so weak that the only evidence of a pharyngeal is the lengthening of the accompanying vowel. Retraction of the root of the tongue causes a lowering of the second formant and a raising of the first formant, as can be seen in the wide band spectrogram on the left in Figure 6. Lowering of the fundamental frequency is more apparent from the movement of the tenth harmonic in the narrow band spectrogram of the same utterance on the right. What is particularly interesting from a phonetic point of view is that the two events do not occur at the same time. As may be seen from the location of the arrows below the spectrograms, the maximum lowering of the second formant associated with the tongue retraction occurs after the minimum in the pitch. Ths indicates that the

pitch lowering is not an automatic consequence of the tongue retraction, and there are two independent gestures involved.

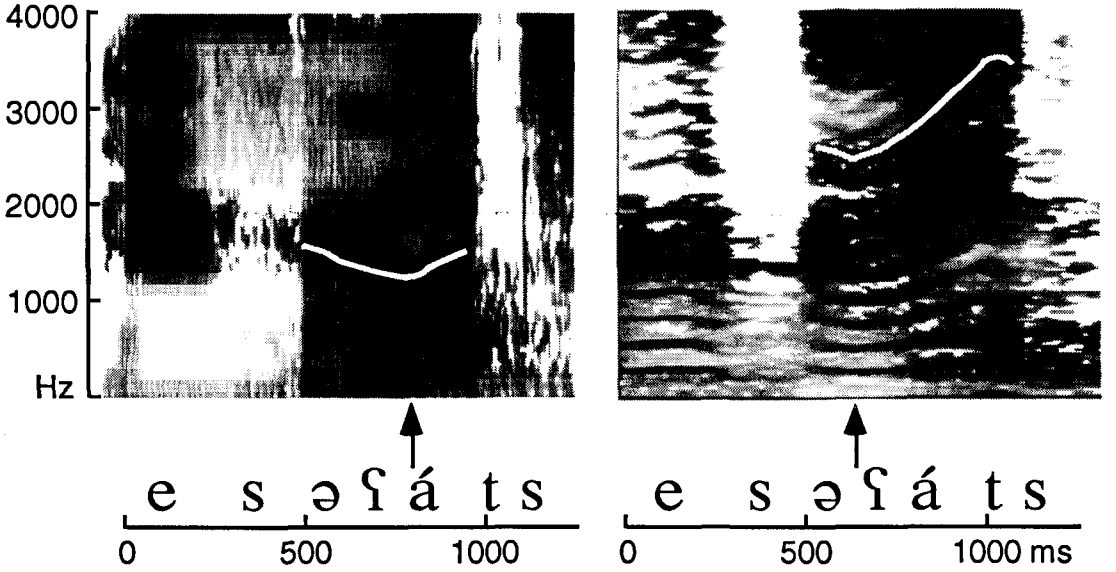


Figure 6. Wide and narrow band spectrograms of the word **esəʕáts** 'it's tied, staked'. Maximum F1 raising and F2 lowering occur at the time of the arrow below the wide band spectrogram on the left. Maximum F0 lowering occurs at the time of the arrow below the narrow band spectrogram on the right.

We hope that the above discussion on some of the sounds of some American Indian languages shows why the study of such languages, which constituted an important part in the history of American linguistics, should continue. Only by a careful phonetic study of speech sounds can we begin to understand what may constitute a universal phonetic set of all the possible sounds of human language. If we are interested in what is a possible human language, that is, a theory of language, such a theory must include a theory of phonetic universals, supported by empirical evidence from languages such as the ones that make up the families of American Indian languages.

Bloomfield, Leonard (1926) A Set of Postulates for the Study of Language. Language 2:153-64.
Bloomfield, Leonard (1933) Language. New York: Holt.
Bloomfield, Leonard (1962) The Menomini Language. New Haven: Yale Univ. Press.
Boas, Franz, ed. (1911) Handbook of American Indian Languages, vol. 1 Bureau of American Ethnology, Bulletin 40. Part I.
Boas, Franz (1947) Kwakiutl Grammar: with a Glossary of the Suffixes. Transactions of the American Philosophical Society, 37 (part 3): 201-377.
Bright, William (1957) The Karok Language. University of California Papers in Linguistics 13.
Bright, W. (1967a) [in collaboration with Jane Hill] The linguistic history of the Cupeño. Studies in Southwestern ethnolinguistics. Ed. By D. Hymes. The Hague: Mouton. 351-71.
Bright, W. (1967b) The Cahuilla language. The ethnobotany of the Cahuilla Indians of Southern California (new edition). Ed. D.P. Barrows. Banning, California: Malki Museum xxi-xxix.
Bright, W. (1968) A Luiseño Dictionary. University of California Papers in Linguistics 51.
Bright, W. (1978). "Sibilants and naturalness in aboriginal California." Journal of California Anthropology. Papers in Liguistics 1: 39-63.
Dart, S. (1991). "Articulatory and Acoustic Properties of Apical and Laminal Articulations." UCLA Working Papers in Phonetics. 79: 1-155.

Flemming, E. , P. Ladefoged and S. Thomason (1994) "The phonetic structures of Montana Salish." UCLA Working Papers in Phonetics. 87:

Hoijer, Harry. (1945) Navaho Phonology. Publications in Antrhopology, no.1. Albuquerque, NM: University of new Mexico Press.

Kirk, P. L., J. Ladefoged, and P. Ladefoged. (1993). Quantifying acoustic properties of modal, breathy and creaky vowels in Jalapa Mazatec. American Indian Linguistics and Ethnolography in honor of Laurence C. Thompson. University of Montana. 435-450.

Ladefoged, P. (1993). "Linguistic phonetic fieldwork: a practical guide." UCLA Working Papers in Phonetics 84: 1-24.

Ladefoged, P., I. Maddieson, and M. Jackson. (1988). Investigating phonation types in different languages. Vocal Physiology: Voice Production, Mechanisms and Functions. New York, Raven. 297-317.

Langdon, M. (1976). "Metathesis in Yuman languages." Language 52(4): 866-883.

Langdon, M. and S. Silver (1984). "California t/t." Journal of California and Great Baisin Anthropology Papers in Linguistics 4: 139-165.

Lindblom, B. (1986). Phonetic universals in vowel systems. In Experimental Phonology (ed. J. Ohala and J. Jaeger), Orlando, Academic Press. 13-44.

McDonough, J. and M. Austin-Garrison (1994). "Vowel enhancement and dispersion in the vowel space of Western Navajo: a study of traditional speakers." UCLA Working Papers in Phonetics 87:

Powell, J.W. (1891) Indian Linguistic Families of America North of Mexico Annual Report of the United States Bureau of American Ethnography. 7: 1-142.

Sapir, Edward.(1921) Language. New York: Harcourt, Brace & World.

Sapir, Edward (1942) Navaho Texts, with Supplementary Texts by Harry Hoijer (Edited by Harry Hoijer). Philadelphia: Linguistic Society of America.

Sapir, E. (1949) The psychological reality of phonemes. In D.G. Mandelbaum, ed., Selected Writings of Edward Sapir in Language, Culture, and Personality. UC Press, Berkeley and Los Angeles. 46-60.

Swadesh, Maurice. (1948) Review of Boas 1947. Word 4:58-63.

Williams, Roger (1643) A Key Into the Language of the Americas. Reprinted in 1963. New York: Russell & Russell.

The phonetics of Kazakh and the theory of synharmonism

Zhoumaghaly Abuov
visiting IREX scholar from Kazakhstan

1. Introduction

Kazakh is the official language of the newly independent Republic of Kazakhstan. It is spoken by about 9 million people not only in Kazakhstan but also by Kazakh minorities in neighboring countries such as Uzbekistan, Kyrgyzia, Mongolia, Russia, and Sinkiang Uigur Autonomous Region of People's Republic of China where close to a million Kazakh live. Kazakh is one of about 20 modern languages that make up the Turkic language family. These languages are spoken from the Adriatic Sea in the west to the Altai mountains in the east. Together with the Turkic languages, Mongolian and Manchu-Tunguz languages they form a distinctive group known as the Altaic Language family. Some scholars add Japanese and Korean to this family, but also some scholars argue that due to lack of enough cognate words the Altaic languages do not constitute a language family, and they call the Turkic languages a separate language family.

The Turkic languages divided into various sub-groups which may be summarized as follows:

a. Southern (Oghuz) Group: it includes Turkish, Azerbaijani, and Turkmen.
b. Eastern (Chaghatai) Group: it includes Uzbek and Uighur.
c. North-Western (Kipchak) Group: it includes Kazakh, Kyrgyz, Tatar, Bashkir, Karachai, Karakalpak, etc.
d. Northern (Altai) Group: it includes Altai, Tuvinian, Khakas, etc.
e. Chuvash Group: the only representative is the Chuvash language.
f. Yakut (Sakha) Group: the only representative is the Yakut language.

The oldest written samples of Turkic languages date back to the 8th century AD. The formation of the Kazakh people and their language started after the 12th century, and modern Kazakh took its present form during the early part of this century. Written in the Arabic alphabet until the end of the 1920's and then in the Latin script for a brief period, Kazakh began to be written in a modified Cyrillic alphabet in 1940. Following the disintegration of the Soviet Union, Kazakhstan became an independent state, and the Kazakh language became the official state language. Now all Kazakh citizens, regardless of their ethnic backgrounds are required to learn Kazakh. After its independence, Kazakhstan gained immediate importance due to its geopolitical location and its rich natural resources, including especially large amounts of oil and natural gas deposits, and vast potential for business opportunities. Thus today it has become ever more important to learn the Kazakh language.

Kazakh used to be written in a modified Arabic alphabet until the 1920s, then for a brief period of time a form of modified Latin alphabet was used. Since 1940 a modified Cyrillic alphabet has been used. For about one million Kazakhs who live in the Sinkiang-Uighur Autonomous Region of the People's Republic of China the official alphabet is a modified Arabic alphabet. Those Kazakhs who live in Turkey as a small minority use a Latin alphabet which is similar to the one used for Modern Turkish. Nowadays there are four movements in Kazakhstan. The representatives of the first movement would like to keep the present Cyrillic alphabet. They argue that an alphabetical change would be to costly culturally as well as financially. The proponents of the second group would like to change to a Latin script. They argue that the Latin alphabet conforms with the linguistic properties of Kazakh better. They say that Turkish and all the civilised societies are using the Latin script, and already some of the other Turkic states have decided to change their alphabets to Latin, such as Azerbaijan, Uzbekistan, Turkmenustan etc. The third group wants to establish a modified Arabic script in order to be closer to our co-religionists. The representatives of the final group would like a change to the Orkhon-Yenisei runic script after modifying it on the basis of some phonological (synharmonic) principles.

2. Major Linguistic Features of Kazakh

Kazakh is a 'subject-object-verb' language. Verbs come at the very end of a sentence, and modifiers precede the modified (head) nouns. It is an agglutinating language, with exclusive suffixation. It has postpositions and no prepositions. The most common way of producing new words is to add derivational and inflectional suffixes to roots or stems. Nouns are inflected for case, and passive, causative, reflexive, reciprocal, and negative forms are obtained by suffixes that are added to the verb. There is no gender and no equivalent to the definite article. From a phonological point of view the major feature is a form of vowel harmony. Generally, initial and final consonant clusters do not occur, although there are some exceptions for the final consonant clusters.

3.1 Phonetic structure of Kazakh from the traditional point of view

Currently there are two opposing hypotheses regarding Kazakh phonetics and phonology. The first theory states that Kazakh has a word stress feature in speech. This theory describes the phonetic system (vowels, consonants, allophones, tense and lax segments, articulation and acoustic correlates) of Kazakh speech sounds from a traditional point of view. As a result of its Indo-European bias, in this theory not only has the Kazakh word prosody been completely reduced to the equivalent of Russian word stress, but the well-known phenomena "vowel harmony" has been neglected or mis-stated. Insofar as the idea of vowel harmony has presupposed the presence of at least two syllables in a Kazakh word, it has played a misleading role as far as relevant research is concerned. Monosyllabic words have been excluded from the field of study.

Kazakh has nine vowels, all of which are of equal length. There are no long vowels. In the roman alphabet the vowels are: a. e, ä, o, u, ı, i, u, ü. Figure 1 shows the vowels in a phonetic transcription.

Table 1. The Kazakh vowels.

	FRONT		BACK	
	UNROUNDED	ROUNDED	UNROUNDED	ROUNDED
HIGH	i	ʏ		ʊ
MID	e	ɵ	ɤ	ɔ
LOW	æ		a	

As shown in Table 2, there are 19 native consonants in Kazakh, and 6 consonants that are borrowed from other languages. The borrowed consonants are **f, v, c, tʃ, dz, x** from Russian, and **h** from Arabic. They are not in bold in Table 2.

Table 2. Kazakh Consonants

	BILABIAL	LABIO-DENTAL	DENTAL	PALATO-ALVEOLAR	VELAR	UVULAR	GLOTTAL
STOP	**p b**		**t d**		**k g**	**q**	
(AFFRICATE)			**c**	tʃ dZ			
NASAL	**m**		**n**		**ŋ**		
FRICATIVE		**f v**	**s z**	**S Z**	**x**	**ʁ**	**h**
TRILL				**r**			
APPROXIMANT	**w**			**j**			
LATERAL			**l**				

The last syllable of a word usually carries the stress:
balá, balalár, balalarím, balalarımá

3.2 Vowel Harmony in Kazakh

Kazakh vowels are divided into two groups:

1. Back (velar) vowels: a, o, u, ı
2. Front (palatal) vowels: ä, ö, ü, i, e

When a new suffix added to nominal and verbal roots and stems, certain phonological rules affecting the vowel of the new suffix should be observed. According to the rules of the vowel harmony base roots or stems with velar vowels can only have suffixes containing the velar vowels. Similarly base roots or stems with palatal vowels can only have the suffixes that contain the palatal vowels. In the case of the foreign words the vowel of a suffix is determined on the basis of the final vowel of the base word.

bala 'child'
bala + lar 'children'
bala + lar + ım 'my children'
bala + lar + ım + da 'in my children'

yel 'country'
yel + der 'countries'
yel + der + im 'my countries'
yel + der + im +de 'in my countries'

Labial harmony is also observed in Kazakh:

köl 'lake'
köl + dör 'lakes'

qʊl + ʊn 'foals'

3.3 Consonant Assimilation

When a root or stem ends in a voiced consonant, the initial consonant of a suffix added has also to be a voiced consonant. In the same way if it is a voiceless consonant, the initial consonant of a suffix has to be a voiceless consonant as well.

at 'horse'
at + tar 'horses'

özön 'river'
özön + dör 'rivers'

4.1 Phonetic structure of Kazakh from the point of view of the theory of synharmonism

This theory was developed and established by the Kazakh linguist A. Zhunisbekov (1987). He has been working on the problem of synharmonism for the past 20 years, and has published extensively on the topic. His research can be summarized as follows. Synharmonism is not an ordinary phonetic phenomenon, but the basis of the whole linguistic structure of the Turkic languages. It is a specific language unit forming the integrity of syllables and words in Turkic speech. There are four synharmonic timbres in Kazakh. They are: hard non-labial, soft non-labial, hard labial. and soft labial syllables. Palatal and labial synharmonism do not function separately and four synharmonic timbres are formed out of their combination.

The *constitutive function* of synharmonic timbres provides proper recognition of a Kazakh word. For example: **[bas]** 'head,' **[b´es´]** 'five,' **[b°os°]** 'empty,' and **[b´°ös´°]** 'boast.' These words are characterised not only by a certain linear combination of sounds, but also by the unique quality of each word's synharmonic timbre, both vowels and consonants alike being synharmonic timbre bearers. The *culminative function* of synharmonism is that all syllables are

organised according to one of the synharmonic timbres. This function plays an important role in the formation of the general phonetic image of Kazakh words and it proves synharmonism to be characteristic not only of polysyllabic words but of monosyllabic ones as well. It means that the terms "synharmonism" and "vowel harmony" are not synonymous, the latter being inexact both as a term and a phenomenon. Vowels play only a syllable forming role in Kazakh without being "harmonisers" and moreover, without performing word distinctive functions. The word-*distinctive function* of synharmonism is significant as well. Kazakh words distinguish not only by a vowel synharmonism, but also by a consonant one. For example: **[tas]** 'stone,' **[t´än´]** 'body,' **[t°os°]** 'wait,' **[t´°ös´°]** 'chest,' **[tıs]** 'outside,' **[t´is´]** 'tooth,' **[tus]** 'side,' **[t´°üs´°]** 'dream,' **[t´es´]** 'drill.' Participation of all sounds comprising the word in word distinction (contrast) is strictly obligatory. It is impossible for any synharmonic variant of one consonant to be replaced by another one. The acoustic features of synharmonic consonants are different. Their four timbres are distinguished from one another by this or that order of placing vowel and consonant formants, because they are articulated at the same places. Since synharmonic phonology allows one to distinguish four synharmonic timbres (hard, soft, labial, non-labial) the whole Turkic languages can be called timbral ones. Thus the language functions inherent in stress of accentual languages and in tone of syllabic languages are found in the Turkic languages an synharmonism. This shows their functional identity in general linguistics and seems to represent important typological features distinguishing language groups.

A. Zhunisbekov formulated the following principles:

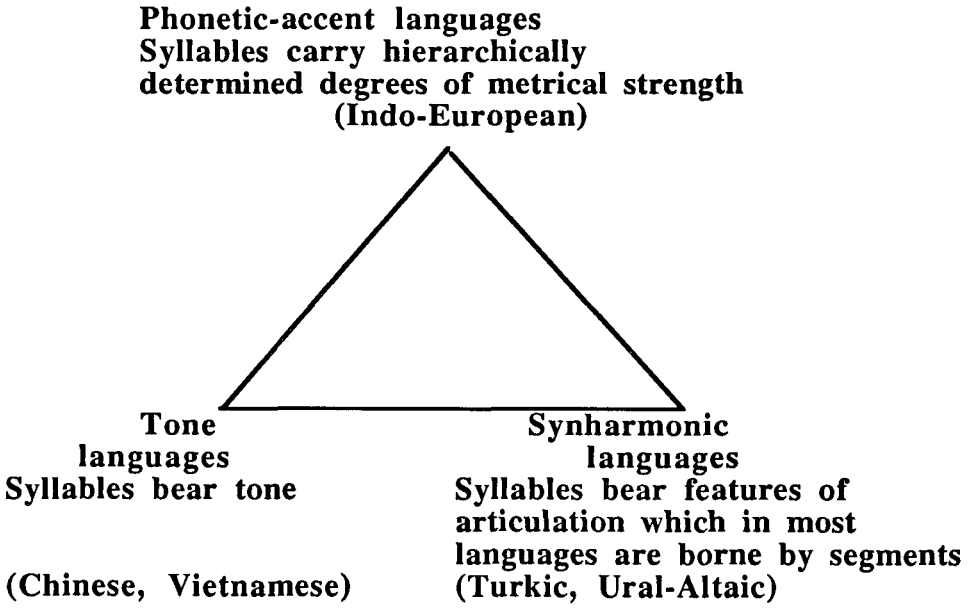
1- Word-synharmonism is a unit on the same level as a word-stress, and word-tone. All these units combine articulatory segments into words.

2- These units are prosodic features of the word. They regulate the phonetic gradation of syllables. Word synharmonism controls timbral (hard non-labial, hard labial, soft non-labial. and soft labial timbres), word-stress controls accent (stressed, unstressed), and word-tone controls tonal features (high, low, medium).

3- The common basic phonetic unit for all three typological groups (Indo-European, Chinese, Uralic-Altaic) is the syllable, but their phonetic realizations and phonological functions are semantically different.

4- Each of the three units regulates articulatory interaction (co-articulation) of sounds in a syllable.

The phonological properties of syllables among languages can be represented as shown below.



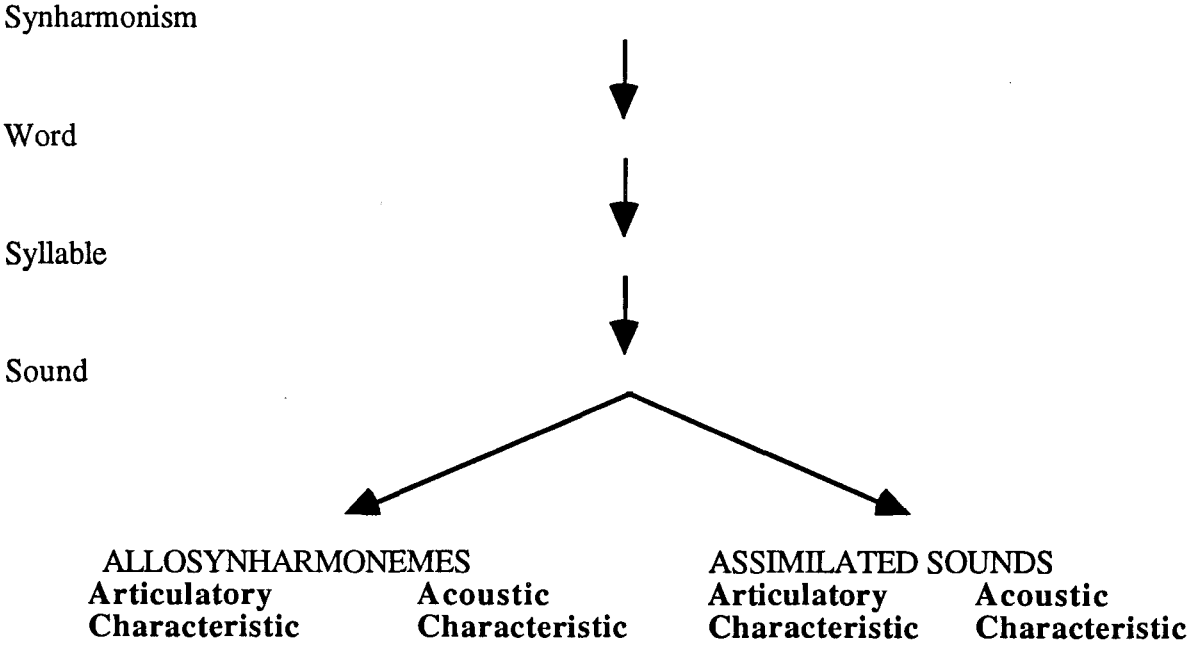
Back and Round are syllable-level features in Turkic. In most languages these features are vowel features; in some languages one or both are distinctive for consonants (e.g. Back in Russian) or for consonants only (e.g. Back in some Causcasian languages). When we observe that these features seem to affect the entire syllable — the consonants as much as the vowels — we can take either of two approaches: (1) The features are carried by the syllable node and therefore all segments are affected equally. (2) the features are carried distinctively by the vowel; Turkic languages have very strong coarticulation, so the consonants are more affected by the vowel than is generally true in other languages. Zhunisbekov suggests that the first approach is more appropriate.

Typological Characteristics of Languages

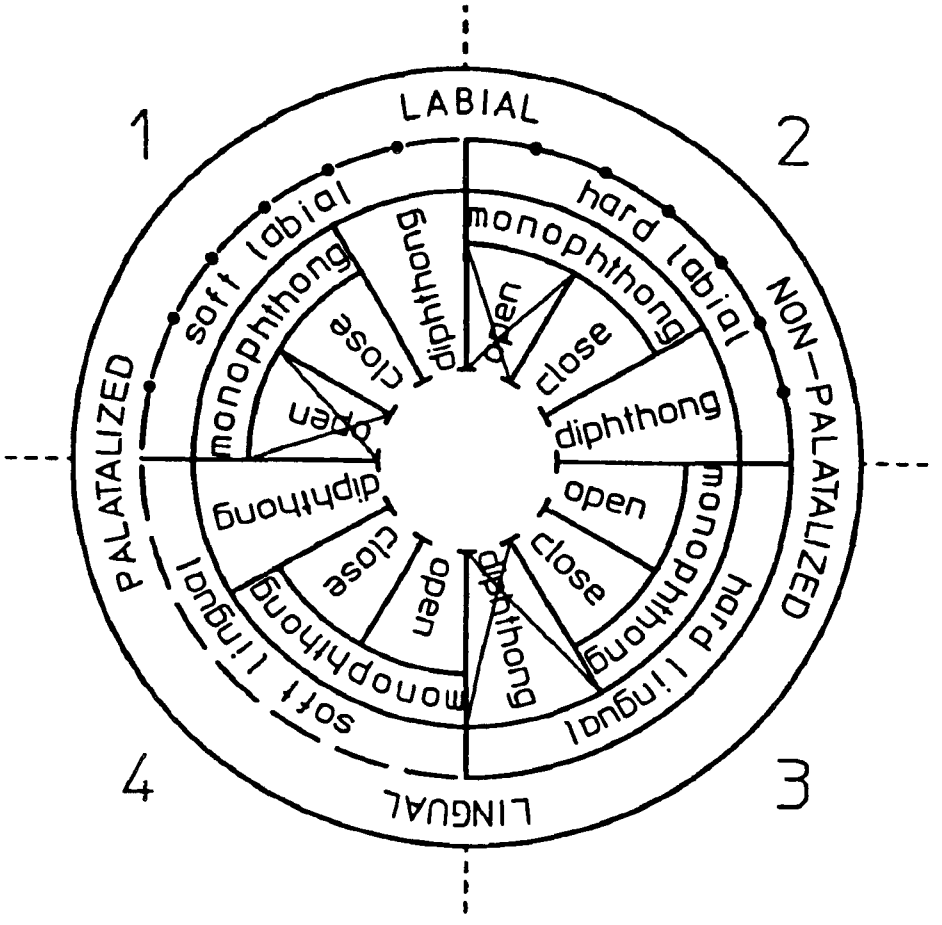
Features	LANGUAGES		
Geographical	Indo-European	South-East Asian	Ural-Altaic
Genetical	Romanic-Germanic	Chinese - Thai	Turkic-Mongolic
Morphological	Inflected	Isolated	Agglutinative
Prosodic	Accentual	Tonal	Synharmonical
Structural	Polyaccentual	Polytonal	Polytimbral
Segmental	Phoneme	Toneme	Synharmoneme
Allosegmental	Allophone	Allotone	Allo-synharmoneme

According to Prof. A. Zhunisbekov, since agglutinative languages have always been investigated on the basis of the linguistic patterns of analyses of the inflected languages (Indo-European), true prosodic, structural, segmental, and allosegmental features of the agglutinative languages could not be established so far. His sketch of the properties of the Kazakh word is as shown below.

Synharmonic Analyses of the Kazakh Word



Zhunisbekov's model of the phonology of synharmonism can be summarized as shown in the diagram below.



5. Spectral Characteristics of Kazakh Vowels

Spectral analyses of vowels and consonants in monosyllabic Kazakh words were made for one male and one female native speaker. Both speakers have a literary pronunciation of Kazakh. In addition, the formants of one speaker previously reported in the literature (Zhunisbekov 1972) were noted for comparison. The formant frequencies of the nine Kazakh vowels were measured from recordings of tokens uttered in isolation in various contexts. The complete set of words used is given in the Appendix. The measurements were made on a Kay Elemetrics Computerized Speech Lab (CSL), using Linear Predictive Coding (LPC) analysis. In accordance with Ladefoged (1993), the formant frequencies may be regarded as determining vowel quality in terms of height and backness, taking F1 as a measure of height, and either F2· or F2·-F1 as a measure of backness. F2· is a weighted average of F2, F3 and F1, and computed as follows (Fant, 1973:52):

$$F2 = F2 + 0.5\ (F3\text{-}F2)\ \frac{F2\ \ F1}{F3\ -\ F1}$$

If the measured formant frequencies are converted from Hertz to Bark, then equal intervals in Bark may be taken as equivalent to equal perceptual intervals. A suitable formula for this purpose is:

B = 13 arctan (0.76f) + 3.5 arctan f/7.5), where B is the critical band value in Bark and f is the frequency in kHz (Syrdal and Gopal, 1985:1088).

5.2 Results

Figure 1 shows the formant frequencies of all the Kazakh vowels analyzed. Obviously, not much can be concluded when all the data is presented as a single picture.

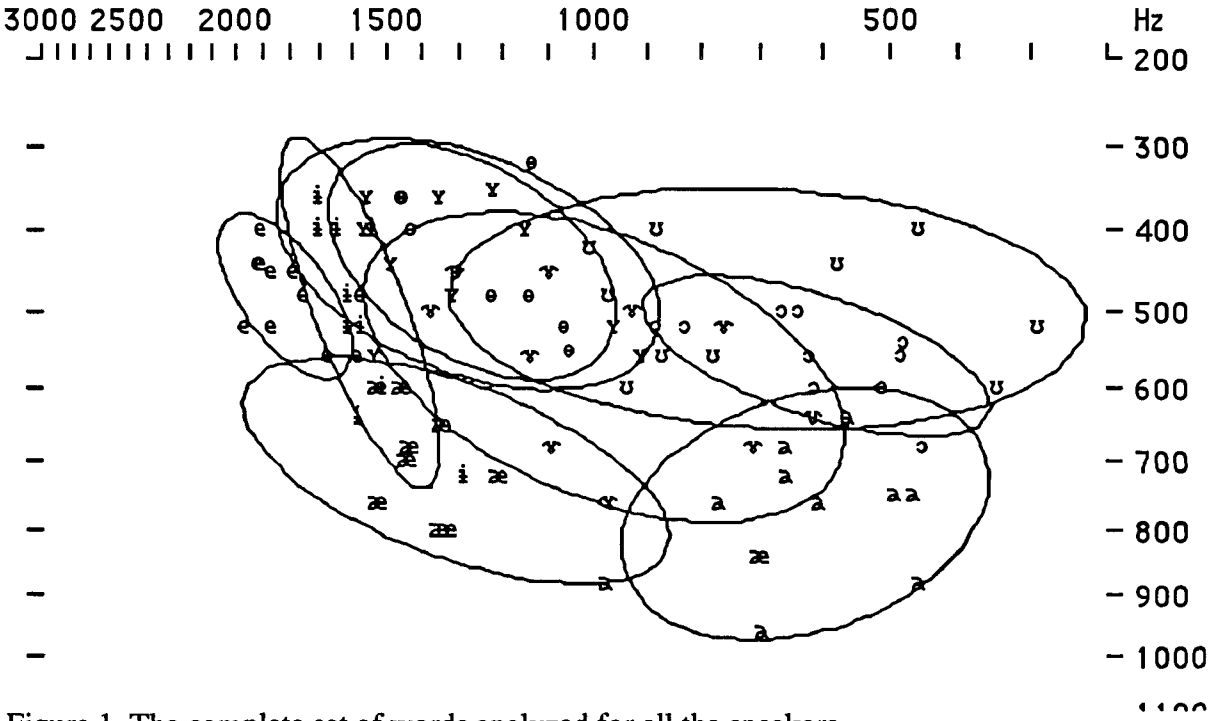


Figure 1. The complete set of words analyzed for all the speakers.

Figure 2 shows the Kazakh vowels in isolation as spoken by one female speaker. It may be seen that there are three vowel heights: [i ɤ ɵ ʊ] are high vowels, [e ɤ ɔ] are mid vowels and [a æ] are low vowels. There are also phonetically three degrees of backness: [i e æ ɤ] are front vowels, [ɵ ʊ ɤ] are central vowels and [a ɔ].

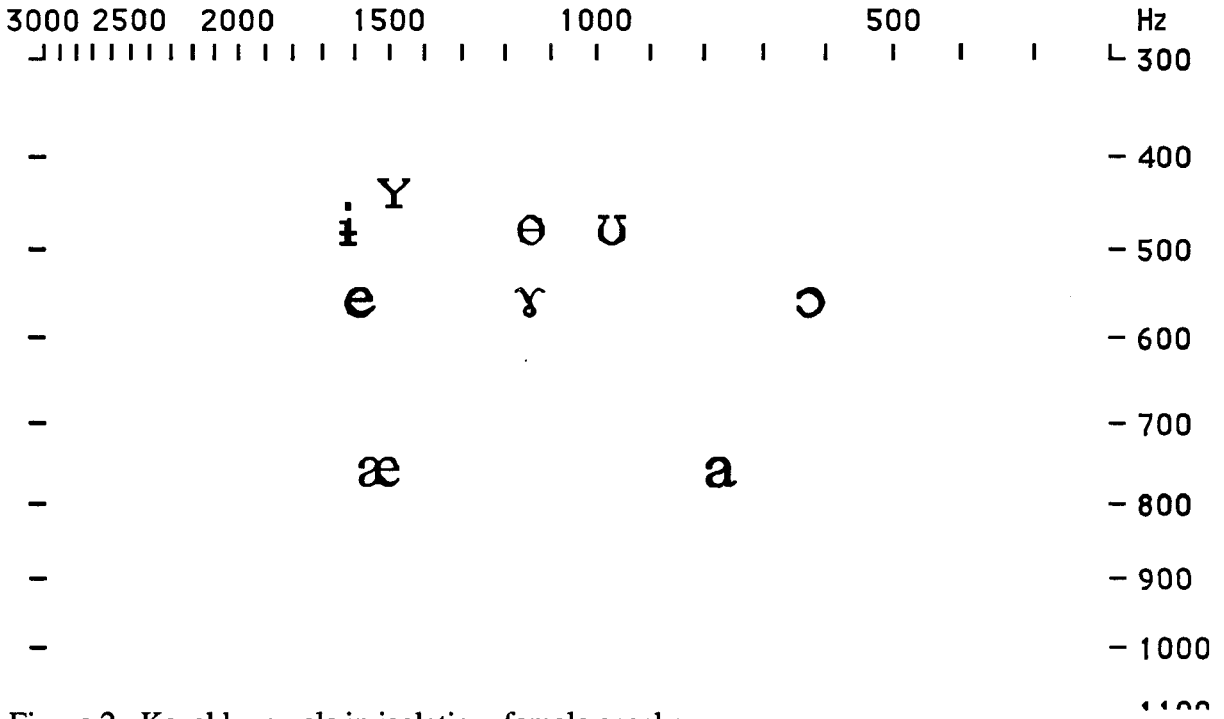


Figure 2. Kazakh vowels in isolation, female speaker

Figure 3. illustrates the Kazakh vowels before [k] or [q] for the female speaker. The back vowels never occur before [k], and the front vowels never occur before [q]. Note that [æ] becomes very back. The other vowels are moving in different directions from their position in isolation. The vowels [a æ ɤ ʏ ʊ ɨ] are moving back and low, and [e ɵ ɔ] are moving front and high.

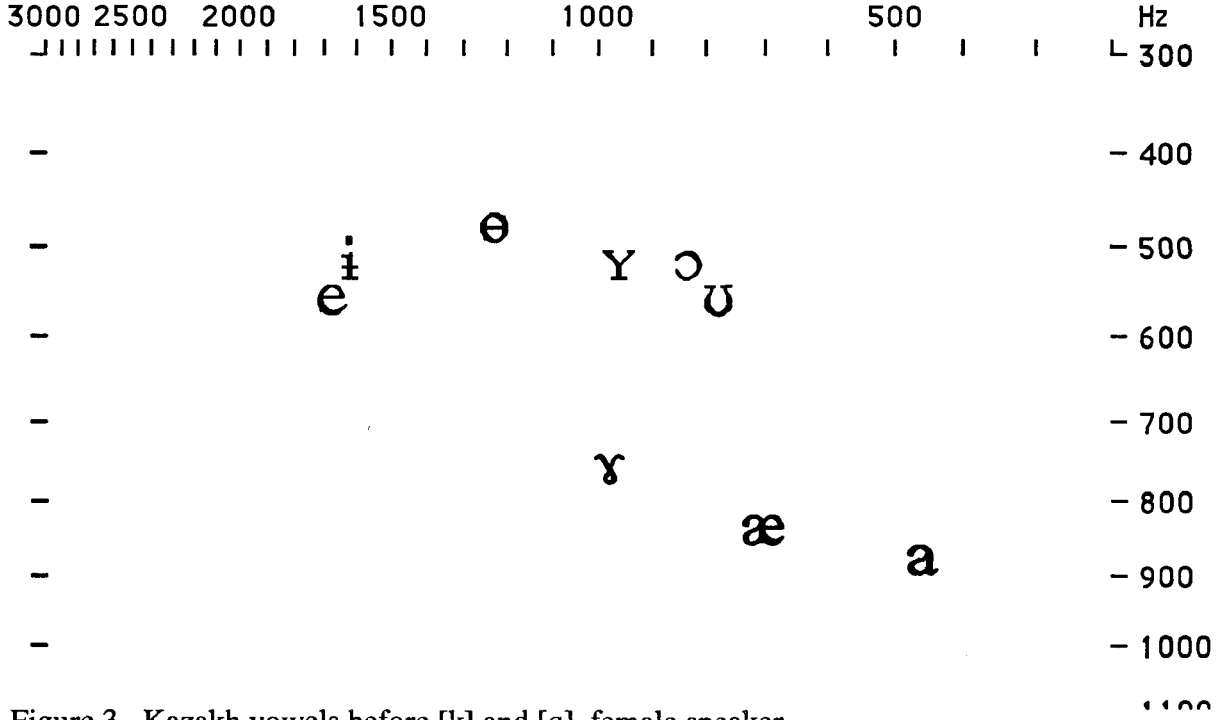


Figure 3. Kazakh vowels before [k] and [q], female speaker

When the vowels are after [b] and before alveolars for the female speaker, there are various differences from their positions in isolation, as shown in Figure 4. The vowels [æ ɤ a i ʊ ʏ], become back and low, and [e ɵ ɔ] become front and high. Similar movements occur for the Kazakh vowels after [t] and before [s] as shown in Figure 5. In this case [æ ɵ a i ɔ ʊ ʏ] become back and low, and [e ɤ] become front and high.

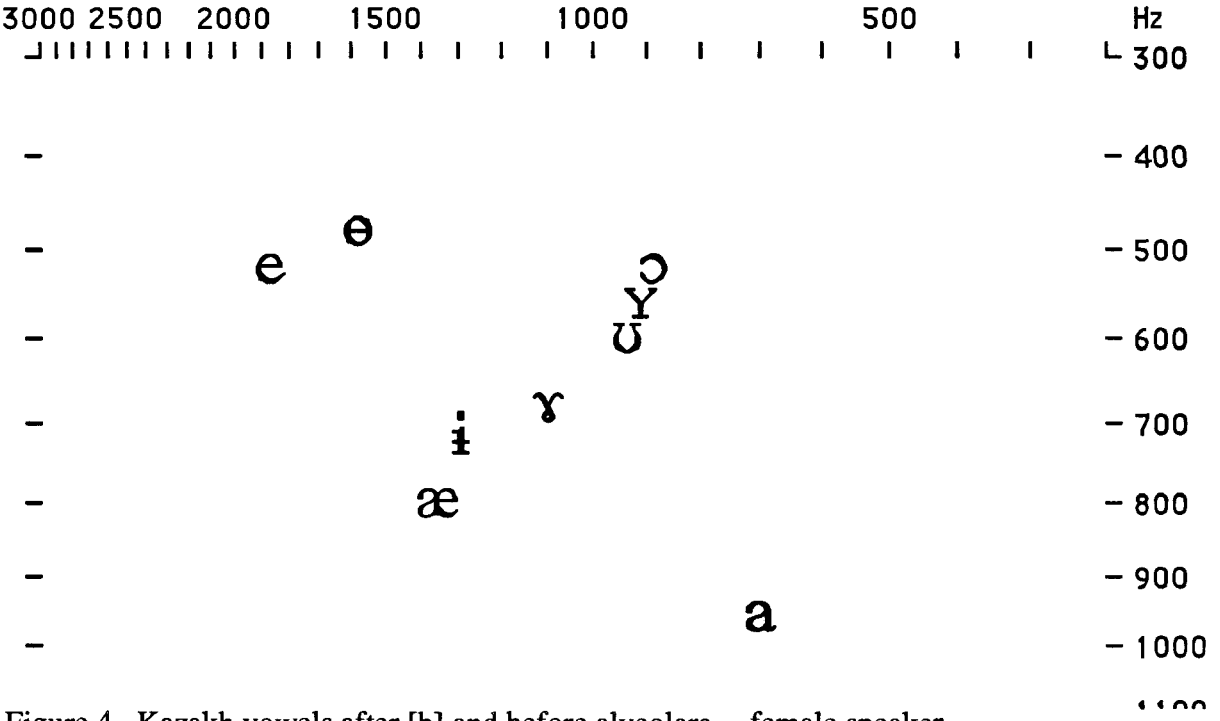


Figure 4. Kazakh vowels after [b] and before alveolars , female speaker

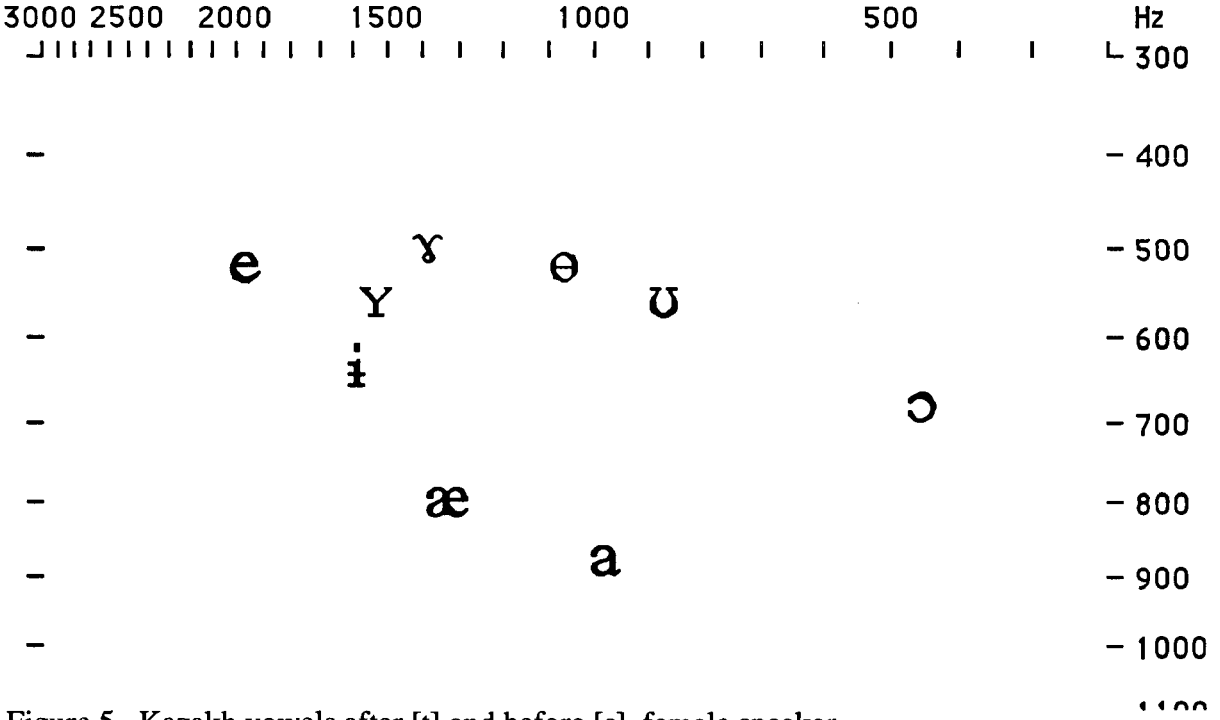


Figure 5. Kazakh vowels after [t] and before [s], female speaker.

Figure 6 summarizes the data for the female speaker.

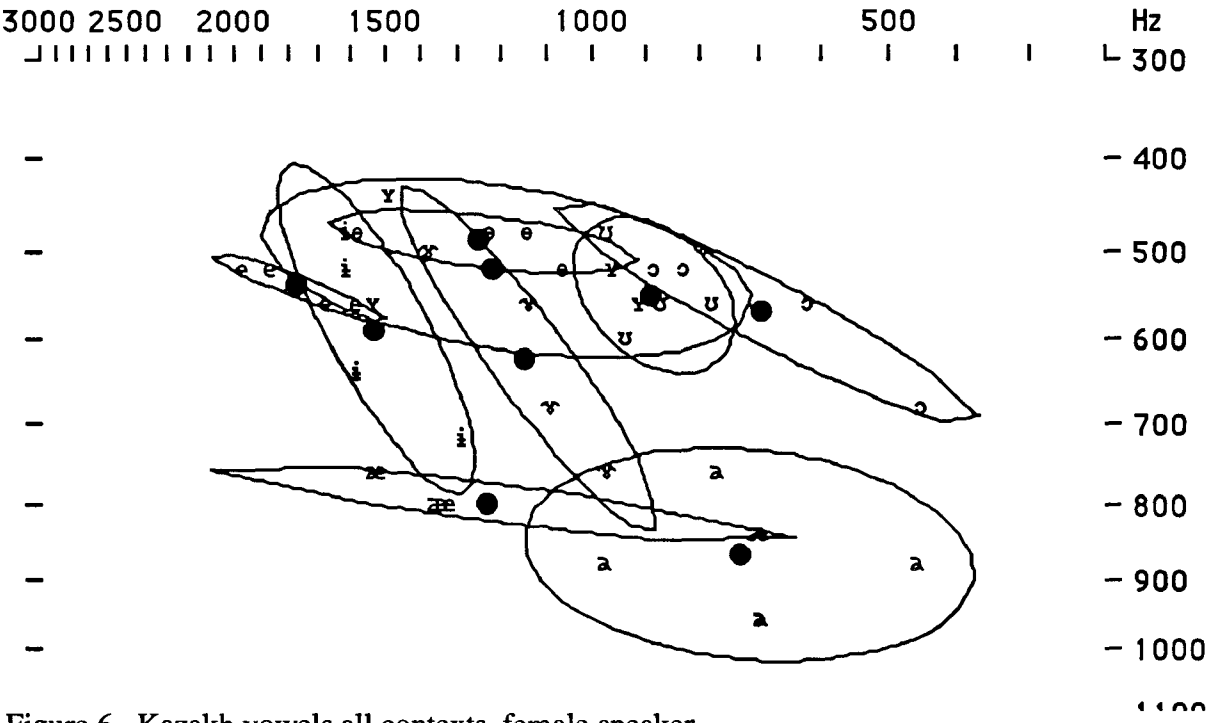


Figure 6. Kazakh vowels all contexts, female speaker

Figure 7 shows the Kazakh vowels in isolation for a male speaker. There are some differences from the female speaker. The high vowels [ɨ ʏ ɵ ʊ] and the mid vowels are not so clearly separated. On the front–back dimension, [ɵ] is more front, and [ʊ] is more back, leaving only [ʏ] as a central vowel.

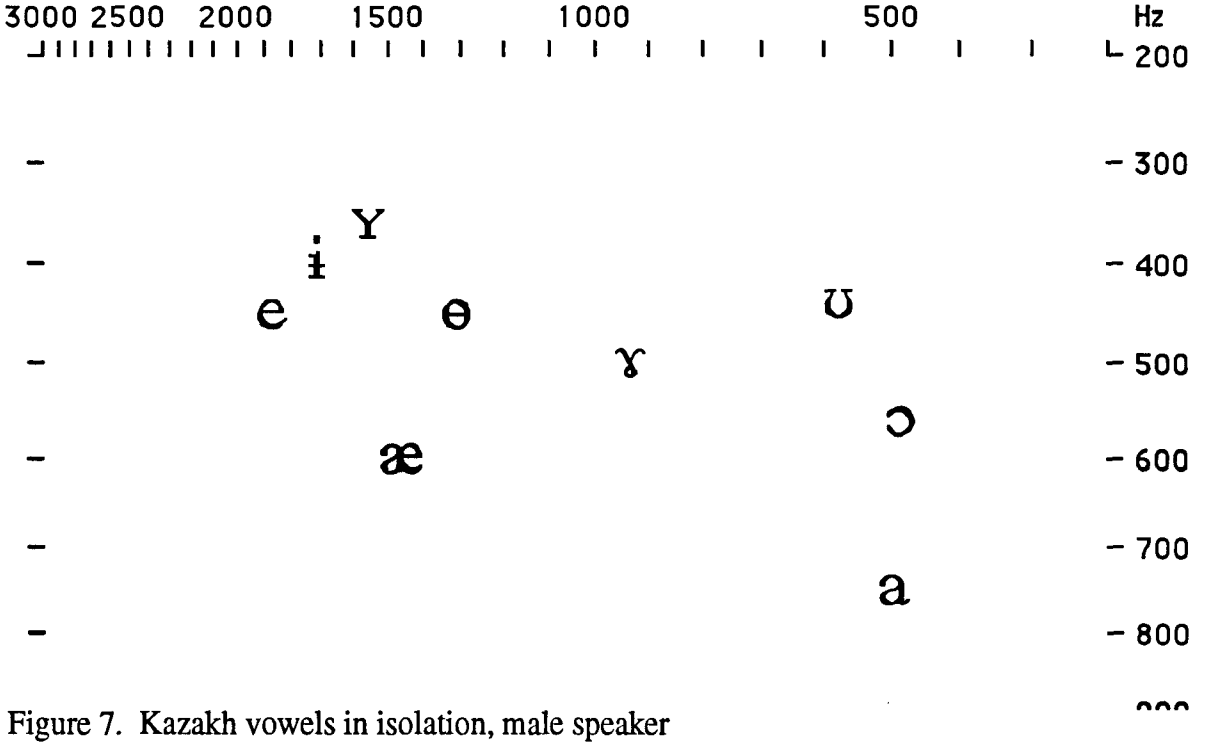


Figure 7. Kazakh vowels in isolation, male speaker

Figure 8 shows the Kazakh vowels before [k] or [q] for the male speaker. As in the case of the female speaker, [æ] becomes very back. In addition the other vowels move, [ɵ] becomes higher, [e ɵ ɔ] become more front, [a] becomes slightly more front, and [ɤ ɨ] lower.

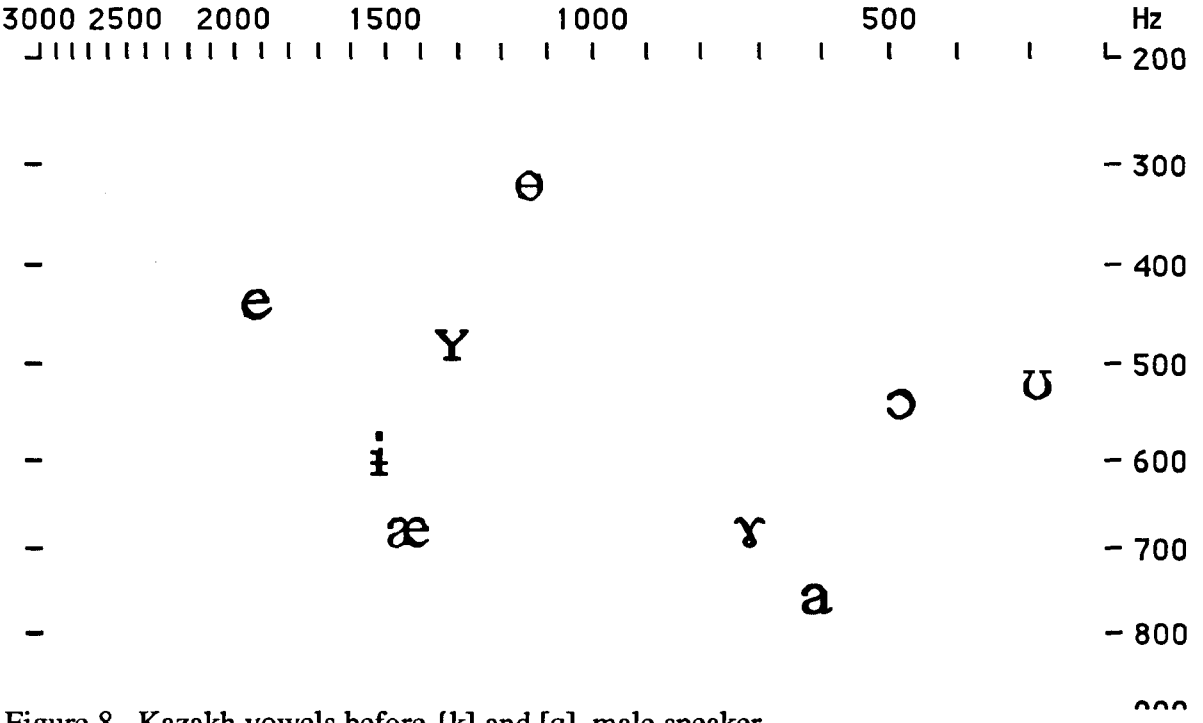


Figure 8. Kazakh vowels before {k] and [q], male speaker.

Figure 9 shows the Kazakh vowels after [b] and before alveolars for the male speaker. Again the vowels move in various directions. The vowels [æ ɵ a ɨ ɔ ʊ e] move up and forwards and [ʏ ɤ] move back.

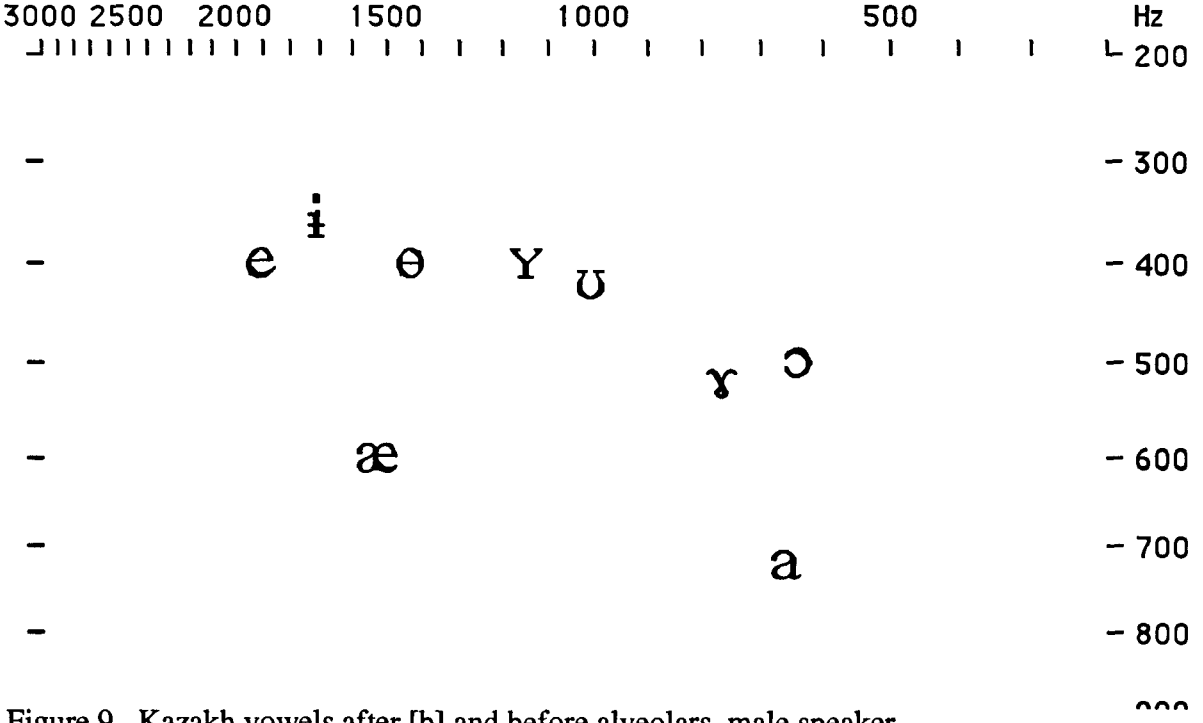


Figure 9. Kazakh vowels after [b] and before alveolars, male speaker

Figure 10 shows the vowels after [t] and before [s]. The vowels [ɵ ɐ ɔ] move up and forwards and [æ a ɨ ʊ ʏ ɤ] move down and back.

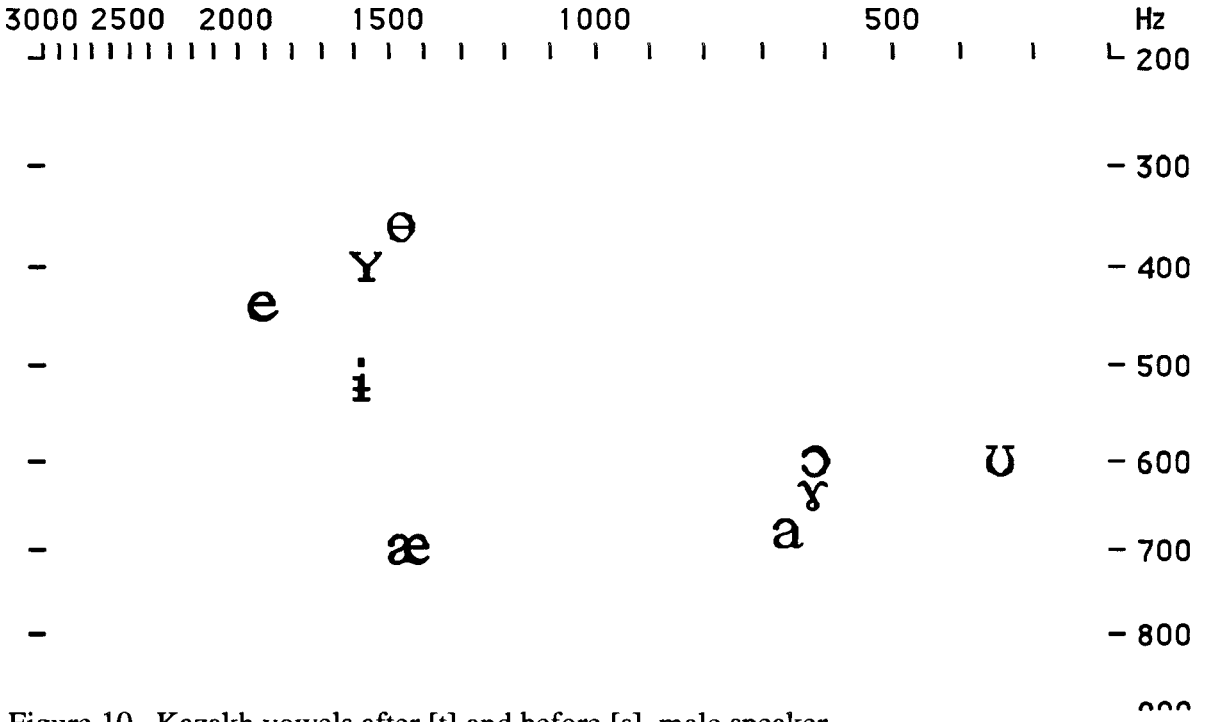


Figure 10. Kazakh vowels after [t] and before [s], male speaker

The complete set of vowels for the male speaker are shown in Figure 11.

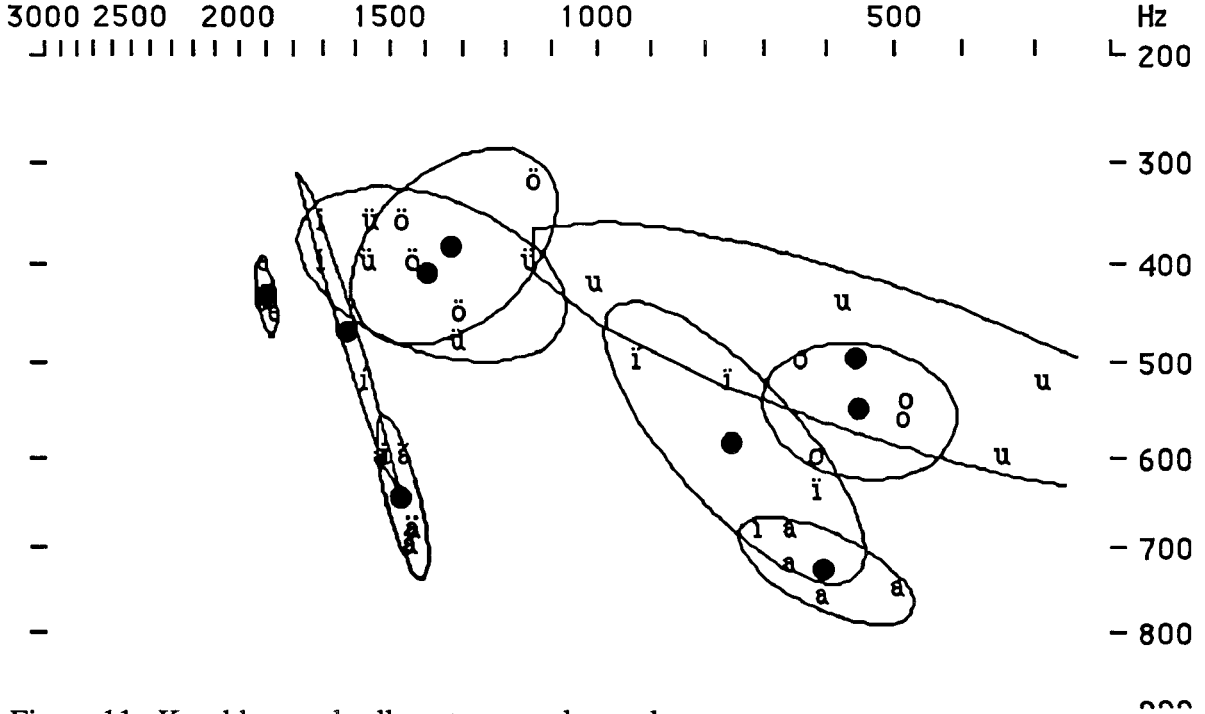


Figure 11. Kazakh vowels all contexts, male speaker

In the literature (Zhunisbekov 1972) there is a report of the vowels of another male speaker, as shown in Figure 12. It may be seen that [ɵ] is much lower, and [ɤ] is higher.

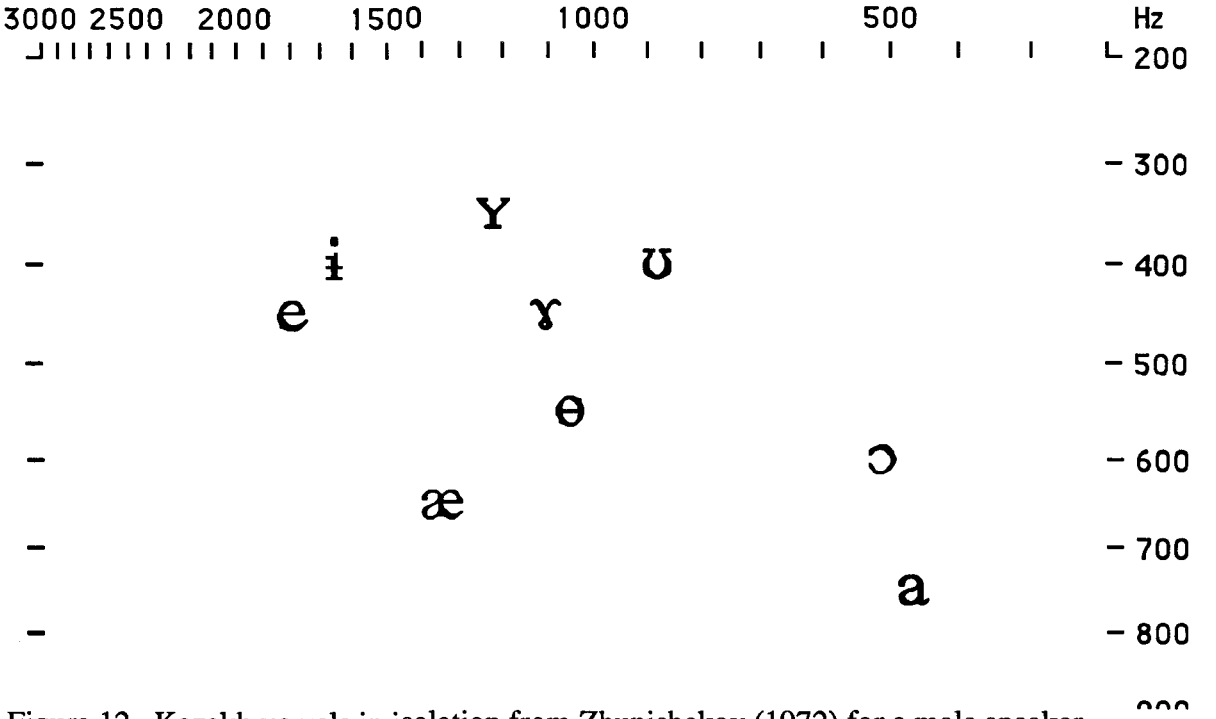


Figure 12. Kazakh vowels in isolation from Zhunisbekov (1972) for a male speaker

A comparison of the vowels as shown in the literature and as found in this investigation is given in Figure 13.

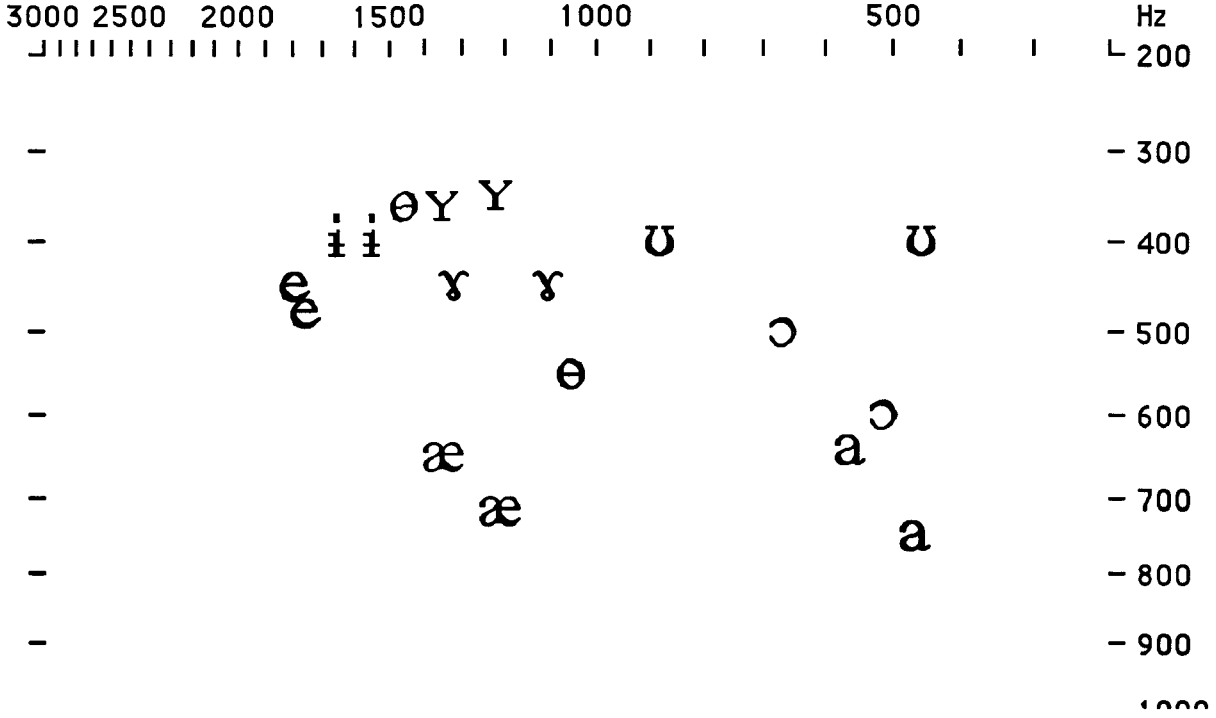


Figure 13. Kazakh vowels in isolation, two male speakers

Conclusions

1. Synharmonism or vowel harmony must be studied from a morphphonological point of view. Sounds are affected in all grammatical forms, such as nouns, verbs, etc.
2. Speech perception experiments should be used to find out which segments or parts of words convey the meaningful information.
3. The releases of the stop consonants should be analyzed to determine how distinct the allophones are in the different synharmonic forms.

References

Fant, G. (1973). Speech sounds and features. Cambridge, MA, MIT Press.

Ladefoged, P. (1993). A Course in Phonetics. New York, Harcourt, Brace, Jovanovich.

Syrdal, A. K. (1985). "Aspects of a model of the auditory representation of American English vowels." **4**: 121-135.

Zhunisbekov, A. (1972), Glasnie kazakhskogo jazyka Alma-Ata, Nauke.

Zhunisbekov, A. (1987), 'The Turkic word prosody problem' Proceedings of the XIth International Congress of Phonetic Sciences **1**: 321-323.

Acknowledgments

I would like to thank Peter Ladefoged, Pat Keating, Ian Maddieson, Jenny Ladefoged, Sinisa Spajiç, Richard Wright for their help in working with various acoustic data, in addition to the other members of the Phonetics laboratory of UCLA.

I also would like to thank Kurtulus Öztopçu of Turkic Languages at UC Berkeley, for his help on theoretical and practical matters,

Appendix
List of the Kazakh words used in the analyses

/aq/	'white'	**/bas/**	'head'	**/tas/**	'stone'	**/tastï/**	'stone-obj. case'
/äk/	'lime'	**/bäs/**	'bet'	**/tän/**	'body'	**/tändi/**	'body-obj. case'
/oq/	'bullet'	**/bos/**	'empty'	**/tos/**	'wait'	**/tostï/**	'wait-past tense'
/ök/	'command'	**/bös/**	'boast'	**/tös/**	'chest'	**/tösti/**	'chest-obj. case'
/uq/	'understanding'	**/bul/**	'this'	**/tus/**	'side'	**/tustï/**	'side-obj. case'
/ük/	'break to pieces'	**/bür/**	'snatch'	**/tüs/**	'dream'	**/tüsti/**	'dream-obj. case'
/iq/	'side'	**/bït/**	'break'	**/tïs/**	'outside'	**/tïstï/**	'outside-obj. case'
/iz/	'trace'	**/biz/**	'we'	**/tis/**	'tooth'	**/tisti/**	'tooth-obj. case'
/ek/	'sow'	**/bes/**	'five'	**/tes/**	'drill'	**/testi/**	'drill-obj. case'

Three types of American /r/[1]

Robert Hagiwara

0. Abstract

Ten monolingual speakers of American English repeated monosyllabic words with /r/ in initial, vocalic (syllabic), or final position. A probe was inserted into the mouth during articulation of the [ɹ] until contact was made with the tongue. The location of the probe contact on the tongue was recorded. For a given speaker and a given syllabic position, probe contact occurred in one of three locations: on the upper surface of the tongue, on the underside of the tongue, or on the tongue apex. These three classes of probe contact are related to three different tongue shapes used to produce American [ɹ].

1. Introduction

When describing American English, writers will sometimes make reference to the use of two distinct tongue shapes associated with American /r/: one *retroflex*, with the tip of the tongue raised, the other *bunched,* with the tip of the tongue retracted and pointed down, commenting that these two articulatory configurations, though very different, both have similar effects on the acoustic speech signal. For instance, Ladefoged (1993:84) writes,

> Some speakers have the tip of the tongue raised, as in a retroflex consonant, but others keep the tip down and produce a high bunched tongue position. These two gestures produce a very similar acoustic effect.

However, Delattre & Freeman (1968) demonstrated that a wide variety of tongue shapes are used by American speakers to produce approximant [ɹ], rather than only two. These they divided into six types, varying chiefly in the location and shape of the oral constriction.

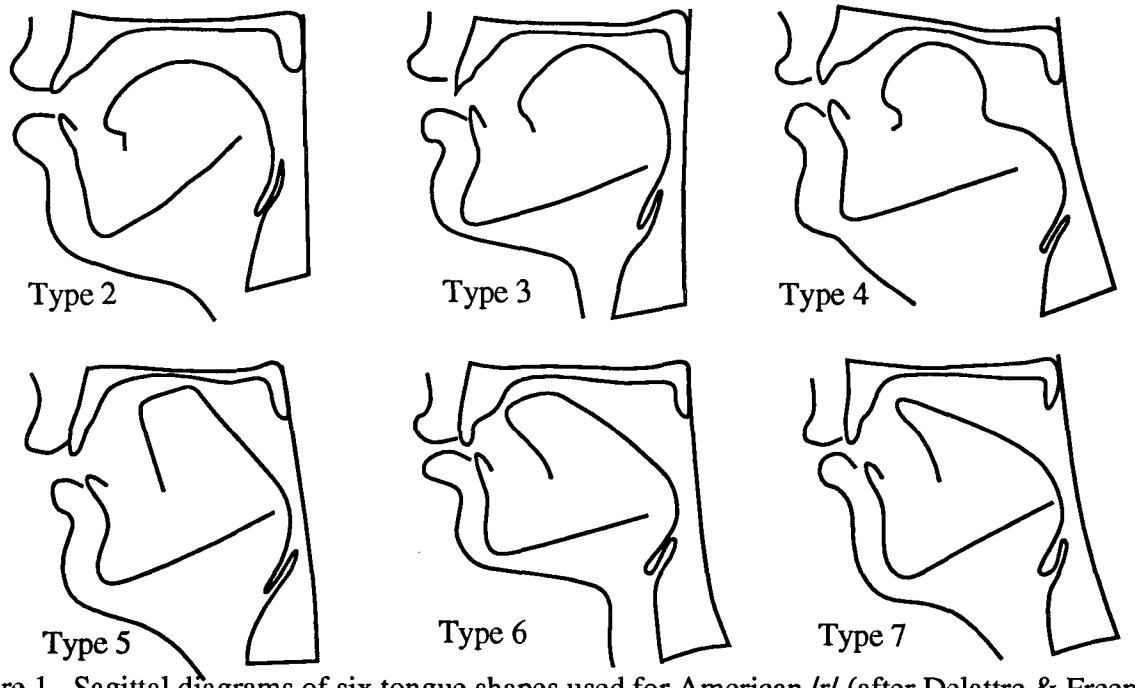


Figure 1. Sagittal diagrams of six tongue shapes used for American /r/ (after Delattre & Freeman, 1968)

[1]Portions of this paper were presented as Hagiwara (1994).

Figure 1 is derived from the X-ray tracings published by Delattre & Freeman (1968), and corresponds to Delattre & Freeman's Types 2 through 7. (Delattre & Freeman present examples of a total of eight tongue shapes, Types 1 and 8 being used only by the British English speakers in their study.) Type 7 is the only one of the six tongue shapes in Figure 1 which can properly be regarded as 'retroflex'. Type 3 is perhaps the most 'canonical' example of a bunched tongue shape. Type 2 is probably best regarded as a schwa-like configuration, such as are used for post-vocalic /r/ in some 'non-rhotic' dialects of English.

Delattre & Freeman apparently regarded the six tongue shapes as a continuum. This continuum may be defined as lying between two extremes, the traditionally described bunched and retroflex tongue shapes. However, writers may refer to the two extremes, ignoring or omitting reference to the 'intermediate' tongue shapes described by Delattre & Freeman. Unfortunately, this may leave many readers with an oversimplified view of the articulations of American /r/. This difficulty is compounded by the idea that retroflex and non-retroflex articulations produce an acoustic effect more properly labeled 'rhoticity', the characteristic lowering of the third formant, which has itself been referred to as 'retroflexion'.

It should be remembered that any description of the articulation of American /r/ that relies on observations of the anterior mouth will necessarily be incomplete. Delattre & Freeman (1968) show in their X-ray tracings that American /r/ is accompanied by a constriction in the pharynx. It can also be seen in Figure 1 that precise shape of the tongue in the posterior mouth shows considerable variation. Imaging the pharynx and the posterior mouth is extremely difficult. Traditional X-ray cineradiography is dangerous to the subject; X-ray microbeam and electromagnetic articulometry systems do not allow tracking of movements of the tongue root. Magnetic resonance imaging and other systems are extremely expensive and do not have the time-resolution necessary to analyze real-time speech. Thus, a simplified view of American /r/ as either bunched or retroflex might be useful descriptively, as long as the articulations of American /r/ can be accurately classified by using what can be observed in the anterior mouth — that is, whether the tongue tip points up or down. The position of the tongue tip can be deduced by inserting a probe into the mouth during articulation.

The purpose of the study reported here was to investigate how probe-contact testing can be used to reveal the characteristics of the oral constriction during the articulation of American /r/. Can probe-contact testing readily distinguish tip-up from tip-down tongue shapes, and what can the results of such testing reveal about the 'intermediate' tongue shapes?

2. A probe-contact study of American English /r/

Procedure

In this study, a probe was inserted into the mouth while the speaker articulated [ɹ] in a variety of contexts. Cotton swabs were used as probes. The tip of the swab was soaked in an iodine preparation to ensure sanitary conditions. Under the direction of the investigator, the subject was asked to insert the probe between the upper and lower incisors, in the midline of the occlusal plane, while articulating the target sound. The subject was then asked to hold the probe against the tongue and to show the investigator where the contact is made. This procedure was repeated until the subject and the investigator were satisfied that a consistent pattern was identified, never more than five repetitions per test allophone (initial, final, syllabic). Typically, three trials were more than ample.

Because the sound must be sustained long enough for the probe to be inserted, the easiest allophone of /r/ to test in this manner is the syllabic [ɻ], as in 'bird' or 'fur'. However, with a minimum of training, the speakers in this study were able to isolate the initial ('reap', 'rap') and final ('peer', 'bar') productions. A relatively low jaw position is preferable. Many speakers in the present study produced /i/ (as in 'peer') with a sufficiently open jaw position as to make the test feasible.

As Delattre & Freeman (1968) indicated, a speaker may use different tongue configurations for the various allophones of /r/. For this reason, the probe technique was used to determine the

gross tongue shape used by each speaker for these three allophones of /r/: initial, final and syllabic.

Subjects

The results from ten subjects (5 men and 5 women) are included in this report. These ten were the first ten subjects in an acoustic study of American /r/ allophones to be reported fully in Hagiwara (in preparation). They were selected to represent, as well as possible, speakers of a single dialect of American English. All were monolingual, between 18 and 26 years of age, and had lived all or most of their lives in southern California.

Results

Three classes of probe contact were noted. The probe sometimes contacted the underside of the tongue, touching the fleshy surface not covered by taste buds. Sometimes, the probe contacted the upper surface of the tongue, usually a centimeter or more (on a protruded tongue) behind the tongue apex. The probe could also contact the tongue tip. The tongue tip can be defined as that part of the tongue blade anterior of a coronal plane defined by the sub-apical border where the fleshy part of the underside of the tongue meets the taste-bud covered surface. In some speakers, this border is on a muscular ridge. The tongue tip, under this definition, may include several millimeters flanking the apex of the tongue.

Table I describes the results of the probe-contact test. The first column identifies the speaker (Speakers 1-5 are female, Speakers A-E are male), and the following columns indicate the location of the probe contact for each of the three allophones of /r/ under study.

Table I. Location of probe contacts on the tongue for each speaker.

Spkr	*Initial /r/*	*Syllabic /r/*	*Final /r/*
S1	underside	underside	underside
S4	underside	underside	underside
S5	underside	underside	underside
Sa	underside	underside	underside
Sc	underside	underside	underside
Sd	underside	underside	underside
S2	tip	upper surface	upper surface
Sb	tip	upper surface	upper surface
Se	tip	upper surface	upper surface
S3	upper surface	tip	tip

Table I indicates the contact location used most often by each speaker in each context. In almost all cases, two or three trials produced identical results. The exceptions to this generality were Speaker A and Speaker E. Speaker A showed a consistent subapical contact in all cases, as indicated by the center of the swab. However in the syllable final allophone, the edge of the swab overlapped the sub-apical ridge. Had the center of the swab also overlapped this line, the contact would have been classified as a tongue-tip contact. Speaker E showed preferred probe contacts as noted in Table I, and indicated these as 'normal' articulations. Upon noting that his initial /r/ was different from his final /r/, however, he prodigiously began to play with various tongue shapes, ultimately producing several in every syllabic position.

Discussion

Probe contact with the underside of the tongue indicates a retroflex tongue shape, since the tongue curls up and exposes the underside of the tongue to the front of the mouth.

Probe contact with the upper surface of the tongue indicates a bunched tongue shape, where the tip is pulled down, out of the way of the probe.

For the probe to contact the tongue tip, the tongue apex can neither be pointed toward the roof of the mouth, as with a retroflex tongue shape, nor can it be pointed down and retracted, as described for the bunched tongue shape. The blade of the tongue is raised, but the primary constriction is not made with the tongue tip, but some other, more posterior, portion of the tongue, such as the posterior tongue blade and/or the anterior tongue body. Two tongue configurations fitting this general description were noted by Delattre & Freeman (1968); these were Types 5 and 6, on the retroflex end of their implied continuum.

Figure 2 is a schematic diagram of the presumed mid-sagittal tongue shapes as revealed by the probe-contact technique. To avoid the confusing terminology often used in discussions of either the articulation or the acoustics of American /r/, these three tongue shapes will be referred to as ***'tip down'*** (bunched, as indicated by probe contact on the upper surface of the tongue), ***'tip up'*** (apical retroflex, as indicated by probe contact on the underside of the tongue), and ***'blade up'***, (previously unnamed, but indicated by probe contact on the tip of the tongue). In Figure 2, the arrow indicates the direction of probe insertion. 'Tip down' (bunched) is in black, 'blade up' is in dark gray and 'tip up' (apical retroflex) is in light gray.

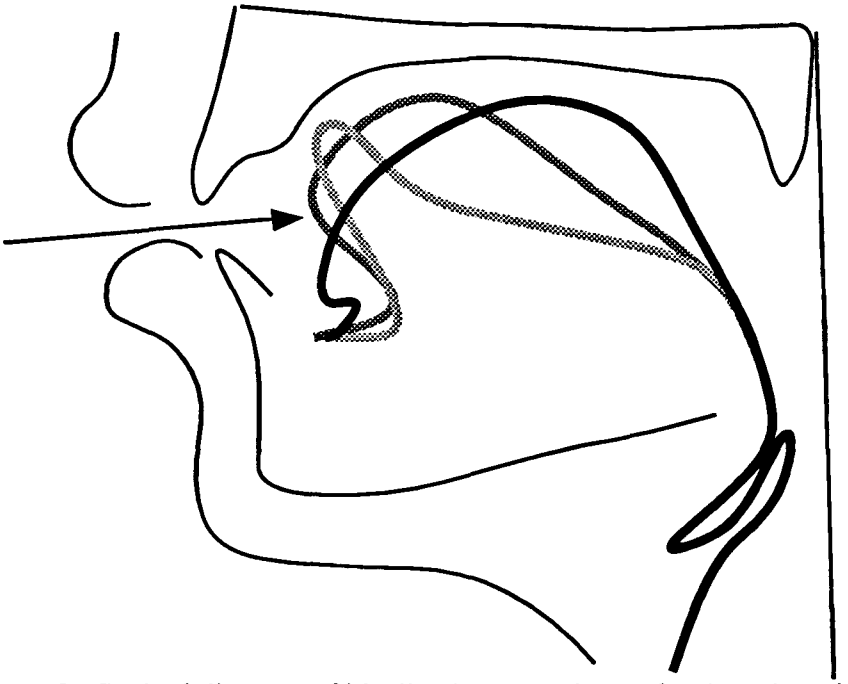


Figure 2. Sagittal diagram of idealized tongue shapes for American /r/.

The tongue shapes illustrated in Figure 2 are inferred from a variety of information, including the probe contact data, Delattre & Freeman's published tracings, some direct observation of the oral portion of the articulation of American /r/ in a variety of speakers, and some separate investigations of the musculature of the tongue (UCLA Phonetics Laboratory, 1990).

Table II translates the probe-contact data from Table I into the presumed tongue shapes used.

Six of the ten speakers in this study show probe-contact on the underside of the tongue for /r/ in all three syllabic positions, indicating an invariably 'tip up' articulation for their productions of /r/. The others used both 'blade up' and 'tip down' tongue shapes, depending on syllabic position. Three used the 'blade up' shape for initial /r/ and 'tip down' for final and syllabic /r/. Speaker 3 used the opposite pattern, reserving 'tip down' for the initial /r/, and using 'blade up' for the other two.

Table II. Tongue shapes used for /r/ by each speaker

Spkr	*Initial /r/*	*Syllabic /r/*	*Final /r/*
S1	tip up	tip up	tip up
S4	tip up	tip up	tip up
S5	tip up	tip up	tip up
Sa	tip up	tip up	tip up
Sc	tip up	tip up	tip up
Sd	tip up	tip up	tip up
S2	blade up	tip down	tip down
Sb	blade up	tip down	tip down
Se	blade up	tip down	tip down
S3	tip down	blade up	blade up

The number of speakers who preferred the 'tip up' tongue shape for syllabic [ɻ̩], six of the ten, was somewhat unexpected. Lindau (1985) provided X-ray tracings of six speakers of American English, all of whom Lindau describes as using a bunched tongue shape. Lindau's tracings have been re-drawn in Figure 3.

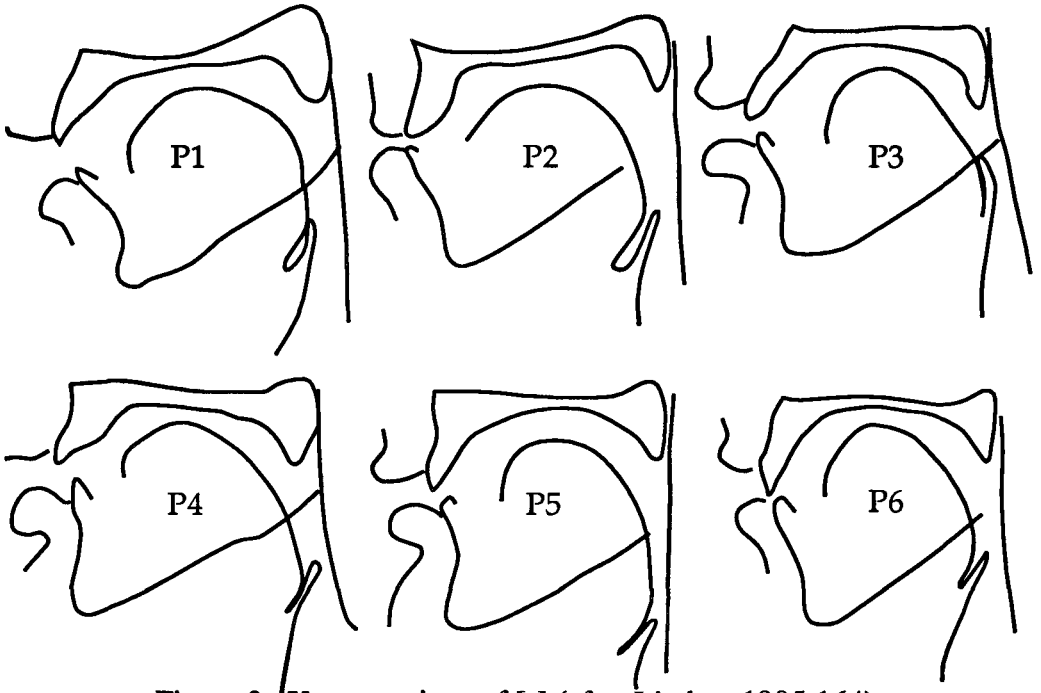


Figure 3. X-ray tracings of [ɻ̩] (after Lindau, 1985:164).

Lindau does not discuss the intermediate tongue shapes described by Delattre & Freeman (1968), but examination of the tracings suggests that at least one of her speakers (P4) and as many as three (P1, P4, P6) might have shown tongue-tip contact in the probe contact test described above. That is, it is possible these speakers used the 'blade up' tongue shape for syllabic /ɹ̩/. Tracing soft tissues such as the tongue is extremely difficult, particularly in the anterior mouth, where the image is confused by the much clearer shadows of the teeth. Unfortunately, the original

X-rays are not available for examination, so it is impossible to tell if such an interpretation of these tracings is consistent with the more detailed evidence that would have been available on the X-ray.

Returning to the main point, even without examining the original X-rays, it is clear that none of Lindau's speakers used a 'tip up' articulation for syllabic [ɻ̩], in which highest point of the tongue should be the tongue apex. Even allowing for tracing errors, this is not true of any of Lindau's speakers.

The interpretation of the probe-contact tests given in Table II, however, supports the traditional view of the 'tip up' or retroflex articulation of American /r/ as the more common tongue shape, a view which until now was largely unsupported by any quantifiable evidence and flatly denied by Delattre & Freeman. However, Delattre & Freeman's examples of Types 5 and 6 both have tongue tip locations which are higher than the occlusal plane. These speakers might have produced tongue-underside contacts if they had been subjected to the probe-contact test. Thus, it is obviously premature to accept the interpretation in Table II without question. In particular, modern imaging methods, such as electromagnetometry, magentic resonance imaging and ultrasound,, should be explored as a source of more conclusive evidence. Still, additional, unquantified observations made with the speakers during the course of the present study suggest that at least some, if not all, actually *did* use tip-up retroflexion. Until more conclusinve evidence can be brought to bear, the question of the relative distribution of the various tongue shapes must remain open.

3. Is 'blade up' a variant of 'tip up'?

By itself, the probe-contact data does not invalidate the view that there are two primary ways to articulate an American /r/. In Delattre & Freeman (1968), the tongue shapes that correspond to what the present study calls 'blade up' are placed on the retroflex end of the continuum. The 'blade up' articulation may represent merely a variation of the extreme tongue shape commonly called retroflex. Perhaps it should be called 'laminal retroflex' to distinguish it from 'apical retroflex' (in which the constriction is formed by the tongue apex) and 'sub-apical retroflex' (in which the constriction is formed by the underside of the tongue blade curling back over the tongue) articulations as described by Ladefoged & Maddieson (in press). However, there is little about the 'blade up' tongue shape which is suggestive of what is normally meant by 'retroflexion'. There is some further, indirect evidence that 'blade up' should not be classed with the other truly retroflex articulations, in particular what here is being called 'tip up'.

The distribution of tongue shapes among speakers seems particularly interesting. The six speakers who used 'tip up', or truly retroflex, articulations appear to do so to the exclusion of other tongue shapes, or at least other tongue shapes are very highly marked. The remaining four speakers who use either the 'blade up' or 'tip down' articulations appear never to use only one or the other, but both, the choice being conditioned by the syllabic positions or other factors not tested in the present study. In this way, the 'blade up' and 'tip down' articulations appear to form a 'natural class', distinct from the 'tip up' articulation.

Another piece of more anecdotal evidence also suggests this trend. When subjects were asked to show the investigator where the probe made contact with the tongue, the 'blade up' and 'tip down' configurations allowed the speaker to protrude the tongue easily — speakers often did so spontaneously and without prompting. With the sublaminal contact of the probe, the tongue could not be easily protruded, and so the subjects usually tipped their heads back and lowered their jaw to permit the investigator to see into their mouths. The muscles used to produce the 'tip up' or retroflex configuration appear to prohibit tongue protrusion (which would entail uncurling the tongue) as an obvious option. As this prohibition is not in force when the probe contacts the tongue apex, this suggests that producing the 'blade up' tongue shape has the same or similar kinesiological requirements as the 'tip down' tongue shape, and that these constraints are different from those placed on the truly retroflex, 'tip up', configuration.

4. Summary of the probe contact data

The probe test results in three classes of probe contacts: contact with the underside of the tongue, contact with the tip of the tongue, and contact with the upper surface of the tongue. These

contacts indicate three different tongue configurations, which have here been characterized as 'tip up', 'blade up', and 'tip down', respectively. Speakers seem to be able to vary between 'blade up' and 'tip down', but speakers in this study do not appear to vary between the truly retroflex 'tip up' configuration and either of the other two. The speaker's ability to protrude the tongue in the 'blade up' configuration (as well as the 'tip down' configuration) suggests strongly that the 'blade up' tongue shape is not best characterized as a modification of 'retroflex', but perhaps represents a separate category or a variant of 'tip down'.

The functional labels for the tongue shapes are preferable to the more traditional terms because they make specific reference to the observable shape of the tongue in the anterior mouth, without the possibly confusing acoustic connotations of words like 'retroflex'.

Among the various tongue shapes associated with American /r/, there are clearly more than the two traditionally described classes (retroflex and bunched), and an attempt must be made at distinguishing a third class of 'blade up' tongue shapes. If anything, the above evidence suggests that 'blade up' is not a variation or modification of 'tip up', but is more closely related to the 'tip down' tongue shape.

The evidence presented here suggests that it is an error to assume that American /r/ is always made with one of the two traditionally described tongue shapes, and that the probe technique will always reveal one or the other. At least three categories of tongue shape must be recognized in articulatory investigations of American /r/. Whether these three can be reduced to the traditional two, and how this is to be done, is not a trivial issue.

References

Delattre, Pierre & Donald C. Freeman (1968). "A dialect study of American r's by X-ray motion picture". *Linguistics, an international review,* **44**, pp. 29-68.

Hagiwara, Robert (1994). "Speaker sex and formant frequencies of American [ɹ]". Poster presented at the 127th Meeting of the Acoustical Society of America, 7 June 1994, Cambridge, MA.

Hagiwara, Robert (in preparation). *Acoustic realizations of American /r/ as produced by women and men.* Ph.D. Dissertation, UCLA.

Ladefoged, Peter (1993). *A course in phonetics (third edition).* New York: Harcourt Brace Jovanovich, Inc.

Ladefoged, Peter and Ian Maddieson (in press). *The sounds of the world's languages.*

Lindau, Mona (1985). "The story of /r/". In Victoria A. Fromkin (ed.), *Phonetic linguistics, essays in honor of Peter Ladefoged.* Orlando: Academic Press, Inc.

UCLA Phonetics Laboratory (1990). "Dissection of the speech production mechanism". *UCLA Working Papers in Phonetics,* **77**.

Sex, syllabic [ɻ̩], and the American English vowel space[1]

Robert Hagiwara

0. Abstract

Formant frequencies of eleven vowels, including syllabic [ɻ̩], were measured for ten (five male and five female) speakers of a single dialect of American English. Mean frequencies of all three formants showed significant differences according to sex within each vowel category. Women's formants were higher overall, but showed greater variability. Moreover, the distribution of vowel categories within the space is different for men and women, women showing unequal distribution of vowel categories in the height dimension as well as centralization of low /æ/. In both men's and women's productions, the back vowels are quite central compared to other dialects, in spite of syllabic [ɻ̩], which has F_1 and F_2 values of a higher-mid, central vowel, overlapping the /o/ and /ʊ/ categories as a result of centralization. American /r/ is usually described as having a third formant below 2000 Hz; this study suggest that the third formant of syllabic [ɻ̩] is 'usually' lower than that for men (1775 Hz) and much higher (2436 Hz) for women.

1. Introduction

In this paper, presents a portion of the data collected from a larger, ongoing study of sex-specific phonetic variables (Hagiwara, in preparation). This paper has two immediate goals. The first is to describe sex-specific asymmetries in the distribution of vowels in the acoustic vowel space of a single dialect of American English. In general, women's formants are higher than men's. This is often regarded as a simple, linear transformation, the result of relatively shorter vocal tracts in women. Shorter vocal tracts modeled as a simple tubes produce higher neutral resonances, as predicted by general acoustic theory. Fant (1973) showed that in fact calculating women's vowel formants from men's formant frequencies was somewhat more complicated, involving different scaling factors depending on which formant is being scaled and which vowel is being calculated. Fant concluded, however, that vocal tract size factors are the fundamental determinant of relative vowel formant scaling. Comparison of Fant's data with Peterson & Barney's (1952) study of American English vowels suggests that, while non-linear and multidimensional, relationships among vowel categories are similar for men and women. The data presented in this study suggest that the situation in southern California English is still more complicated, with differential distribution of vowel categories within the men's and women's vowel spaces.

The second goal is to describe syllabic [ɻ̩] and its relationship to the plain (non-rhoticized) vowels of American English as spoken by women and men. Syllabic [ɻ̩] is unquestionably a vowel in this dialect of American English, but is often overlooked in descriptions of American vowels. As noted by Peterson & Barney (1952) and Lehiste (1962), among others, American [ɹ] sounds are characterized by an extremely low third formant. However, the literature contains little discussion of the first and second formants of syllabic [ɻ̩] (Peterson & Barney, 1952, being a notable exception) and their relationship to the rest of the vowel space.

2. Procedure

Subjects

Students at UCLA were asked to respond to a Speaker Survey Form if they were willing to participate in a phonetic study of /r/ in dialects of American English. From the respondents to the survey, five men and five women were selected for their similarity in age and geographic background. All were between 18 and 26 years of age, had lived all or most of their lives in southern California, and were monolingual. Each was compensated $10 US for participating in the study.

---

[1]Portions of this paper were presented as Hagiwara (1994).

Tokens

The words used in this study are listed in Table I. These words illustrate the eleven vowels in this dialect of English (excluding diphthongs) in three environments: /b_t/, /t_k/, and /h_d/. Real English words and proper nouns were used; where a word of the appropriate phonological shape did not exist a word as close in shape as possible to the target was substituted, as with 'put', 'duke' and 'hut'.

Table I. Words illustrating the 11 vowels of Southern Californian English.

beat	teak	heed
bit	tick	hid
bate	take	hate
bet	tech	head
bat	tack	had
boot	duke	hoot
put	took	hood
boat	toke	Hode (hoed)
bought	tock	hod
but	tuck	hut
Bert	Turk	herd

Recording and measurement techniques

The 33 words presented in Table I were added to 36 others illustrating [ɹ] in initial and final positions. Each word was presented in the frame 'Cite ___ twice.' Each of the 69 sentences randomized together and included three times in a single recording script consisting of a total of 207. Each speaker was recorded reading from the script in a sound-treated room on professional quality equipment.

The subjects' speech was digitized from the audio cassette tape of the recording session at 10 kHz using Kay Elemetric's Computerized Speech Laboratory (CSL). Frequencies were measured for the first three formants of each syllabic nucleus in the 33 words illustrating syllabic [ɻ] and other vowels. Formants were determined by simultaneous evaluation of wide band spectrograms and narrow band FFT spectra averaged over a 30 msec window through the steady state portion of the vowel (if there was one). If no steady state was present, the 30 millisecond window was placed in the center of the vowel. The spectrogram and FFT spectra were supplemented at times by an LPC formant history and/or an LPC slice taken in the center of the FFT window. The number of LPC poles used varied between 10 and 14 depending on the sex of the speaker and the number of formants visible in the spectrogram.

3. Results and Discussion

In this section, formant averages are discussed for each speaker in the study. More general conclusions to be drawn for these data are discussed in Section 4. In this and the following sections, female speakers are designated numerically (i.e., Speakers 1-5, or S1-S5) and male speakers are designated alphabetically (Speakers A-E, or Sa-Se).[2] In the figures which follow, axis labels are in Hertz, but are plotted in a non-linear (Bark) scale. F_1 is plotted down the vertical axis and F_2 (or F_3) is plotted right-to-left along the horizontal axis. Ellipses enclose areas defined by two standard deviations from the mean.

[2]This scheme was selected for the convenience of avoiding possible confusions between, for instance, F2 (female speaker two) and F_2 (second formant).

Speaker 1

Speaker 1's formant averages are given in Table II. Her F_1xF_2 vowel space is illustrated in Figure 1; her F_1xF_3 vowel space is illustrated in Figure 2.

Table II. Formant averages for Speaker 1. Units are Hertz; standard deviations are in parentheses; n=9 for each vowel.

	F_1	F_2	F_3	
i	316 (22)	3109 (127)	3601 (344)	i
ɪ	487 (35)	2503 (61)	3309 (71)	ɪ
e	470 (45)	2778 (97)	3359 (95)	e
ɛ	932 (112)	2174 (80)	3208 (54)	ɛ
æ	1112 (45)	1912 (58)	2948 (131)	æ
u	365 (32)	1670 (396)	3169 (86)	u
ʊ	550 (66)	1759 (202)	3180 (35)	ʊ
o	587 (81)	1302 (120)	3273 (36)	o
ɑ	1004 (41)	1326 (52)	2680 (56)	ɑ
ʌ	942 (49)	1802 (170)	3037 (156)	ʌ
ɹ	547 (49)	1528 (59)	1925 (78)	ɹ

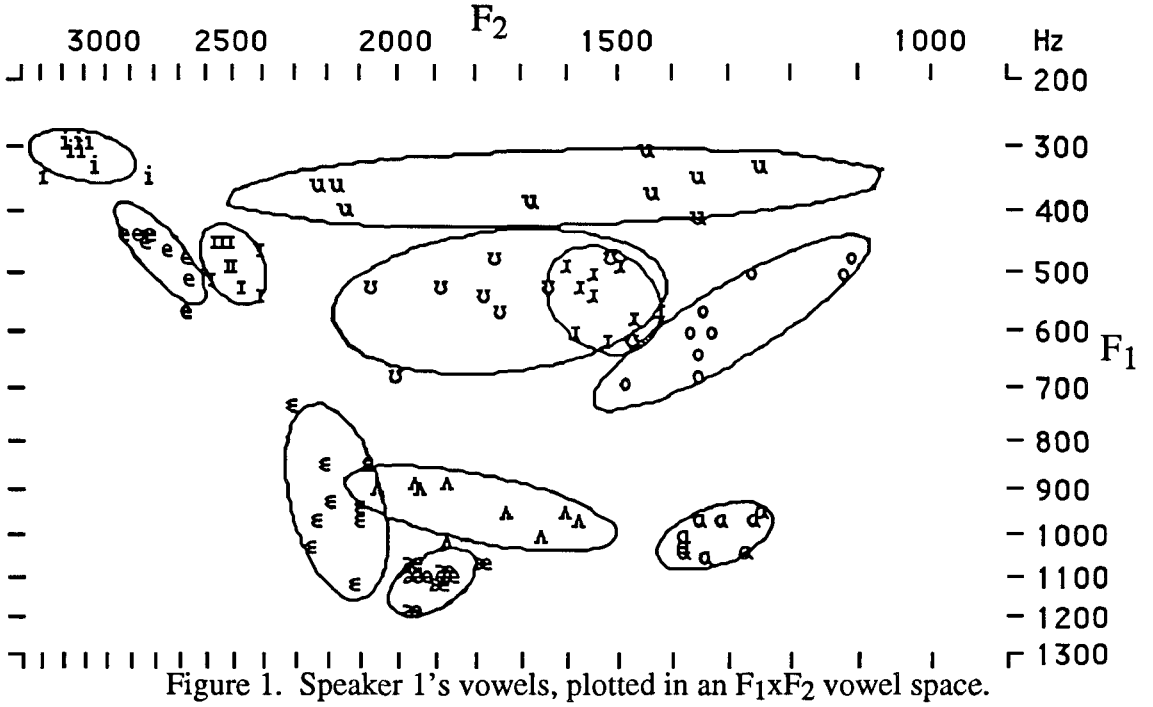


Figure 1. Speaker 1's vowels, plotted in an F_1xF_2 vowel space.

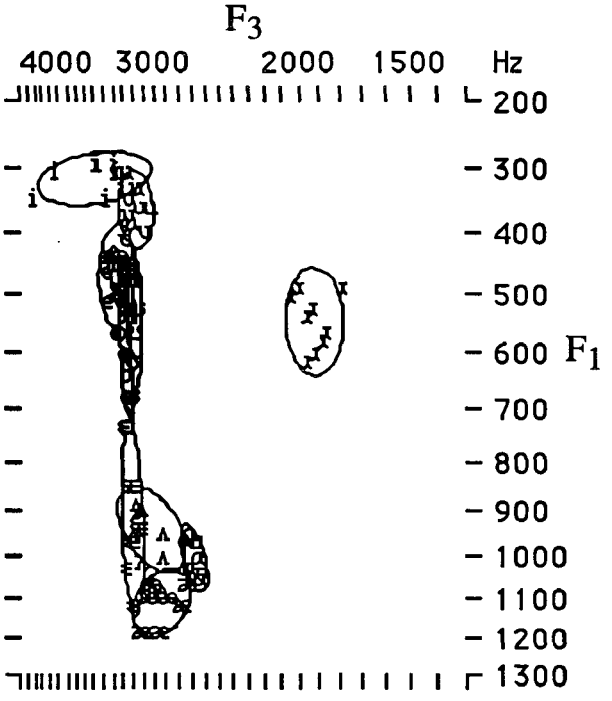


Figure 2. Speaker 1's vowels, plotted in an F_1xF_3 vowel space.

Johnson, Ladefoged and Lindau (1993) suggested that token-to-token variation in vowel formants is quite limited within an individual speaker. For Speaker 1, this appears to be true of the /i/, /ɪ/, /ɛ/, /æ/, /A/ and /ɹ/ categories. However the remaining five categories seem to show considerable token-to-token variation. In the case of /u/, three tokens have F_2 above 2000 Hz (see Figure 1) as a result of the fronting of /u/ after coronals. That is, this is a reflex of the /ju/ diphthong after coronals in other dialects. In the word "duke", S1 has a front on-glide, resulting in a higher measured second formant than the other /u/ tokens. However, the remaining six /u/ tokens are still relatively widely distributed in the higher-back portion of the space. This, as well as the relatively wide distribution of /ʊ/ and /o/ tokens, may be the result of variation in lip rounding. It is typical in southern California speech for the back vowels to be unrounded (i.e. "good" is often pronounced [gɯd]) in spontaneous speech, even though they may be fully round in isolation. In the non-natural laboratory environment, some variation in rounding may be expected. Note that in S1's /o/ category, backness and height co-vary, in the sense that the tokens with the lower F_1's are also the ones with the lower F_2's, suggesting a correlation between height and backing/rounding.

However, variable rounding is not a viable explanation for the relatively disparate productions of /ɛ/ and /ʌ/ for Speaker 1. /ɛ/ varies a great deal in F_1, as /ʌ/ does in F_2.

Speaker 2

Speaker 2's formant averages are given in Table III. Her F_1xF_2 vowel space is illustrated in Figure 3; her F_1xF_3 vowel space is illustrated in Figure 4.

Table III. Formant averages for Speaker 2. Units are Hertz; standard deviations are in parentheses; n=9 for each vowel.

	F_1	F_2	F_3	
i	359 (30)	2692 (128)	3448 (110)	i
ɪ	494 (78)	2338 (25)	3138 (200)	ɪ
e	445 (35)	2626 (76)	3216 (258)	e
ɛ	901 (125)	2153 (58)	2975 (270)	ɛ
æ	1050 (34)	1650 (86)	2673 (95)	æ
u	364 (38)	1697 (378)	2527 (95)	u
ʊ	504 (48)	1652 (237)	2583 (140)	ʊ
o	540 (66)	1322 (75)	2528 (39)	o
ɑ	999 (37)	1315 (52)	2551 (106)	ɑ
ʌ	882 (41)	1809 (175)	2793 (247)	ʌ
ɝ	534 (97)	1419 (123)	1733 (140)	ɝ

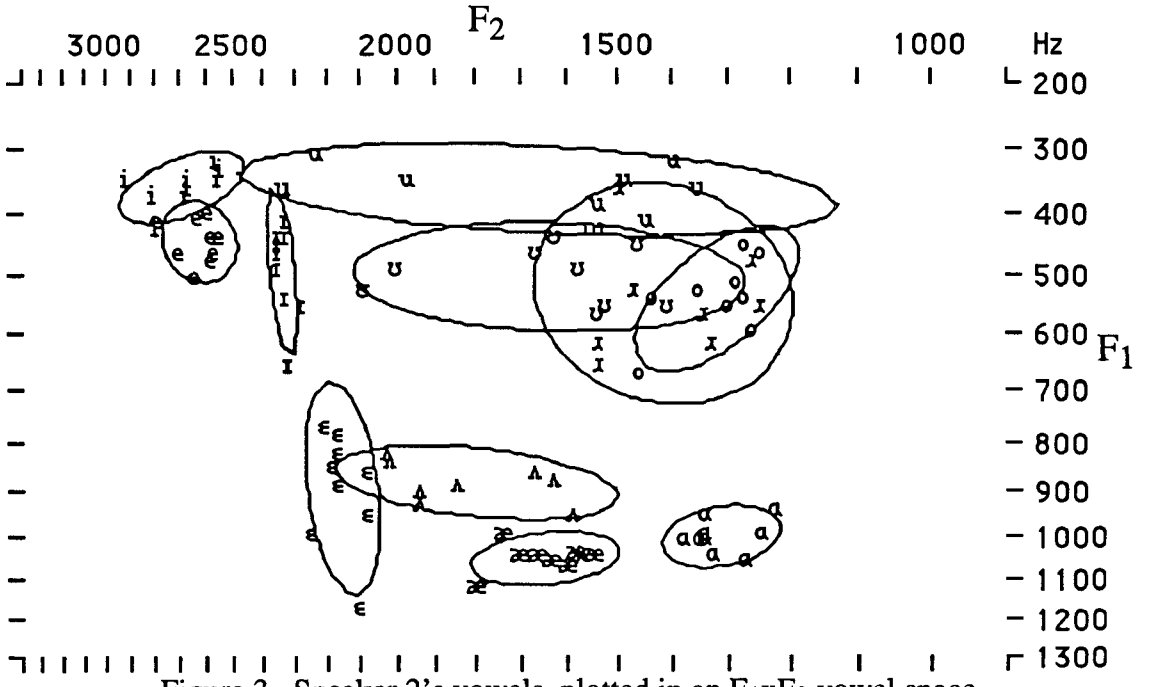


Figure 3. Speaker 2's vowels, plotted in an F_1xF_2 vowel space.

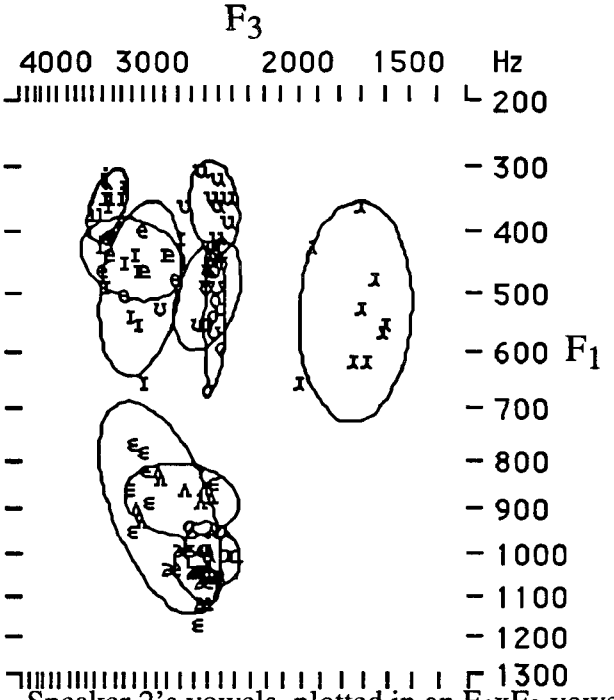


Figure 4. Speaker 2's vowels, plotted in an F_1xF_3 vowel space.

Speaker 2 shows a similar pattern to S1, in that variation in F_1 of /ɛ/ and F2 of /ʌ/ is relatively wide. S2 has broader variation in F_1 and F_2 of /ɪ/ than S1, resulting in greater overlap with /ʊ/ and /o/ in the F_1xF_2 space. However, as can be seen in Figure 4, the third formant of /ɪ/ is much lower than the third formant of any other vowel. Unlike S1, however, S2's third formant seems to correlate more with the frequency of the second formant, the fronter (higher second formant) vowel showing higher F_3's as well.

Speaker 3

Speaker 3's formant averages are given in Table IV. Her F_1xF_2 vowel space is illustrated in Figure 5; her F_1xF_3 vowel space is illustrated in Figure 6.

Table IV. Formant averages for Speaker 3. Units are Hertz; standard deviations are in parentheses; n=9 for each vowel.

	F_1	F_2	F_3	
i	387 (29)	2664 (102)	3392 (231)	i
ɪ	544 (37)	2298 (59)	2960 (253)	ɪ
e	473 (17)	2567 (62)	3086 (445)	e
ɛ	927 (33)	2135 (174)	2837 (338)	ɛ
æ	1076 (35)	1652 (80)	2515 (57)	æ
u	467 (17)	1799 (323)	2697 (119)	u
ʊ	687 (122)	1601 (75)	2564 (119)	ʊ
o	723 (207)	1511 (98)	2632 (73)	o
ɑ	1014 (49)	1388 (35)	2507 (95)	ɑ
ʌ	942 (37)	1712 (97)	2656 (131)	ʌ
ɹ	507 (63)	1643 (96)	2141 (124)	ɹ

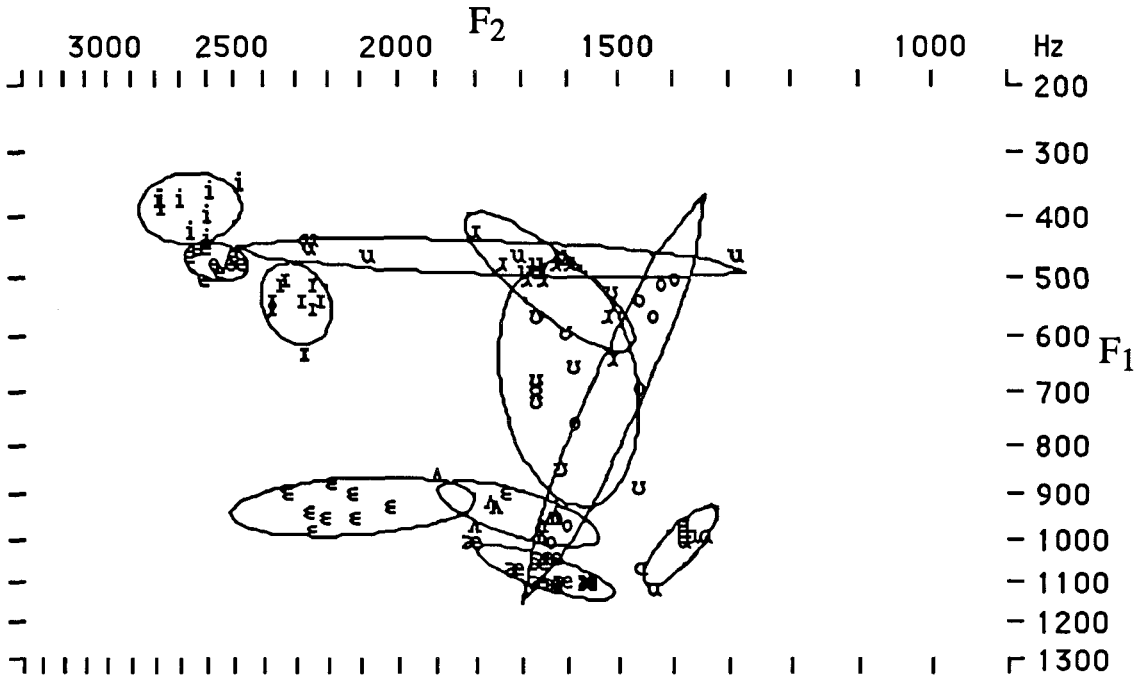


Figure 5. Speaker 3's vowels, plotted in an F_1xF_2 vowel space.

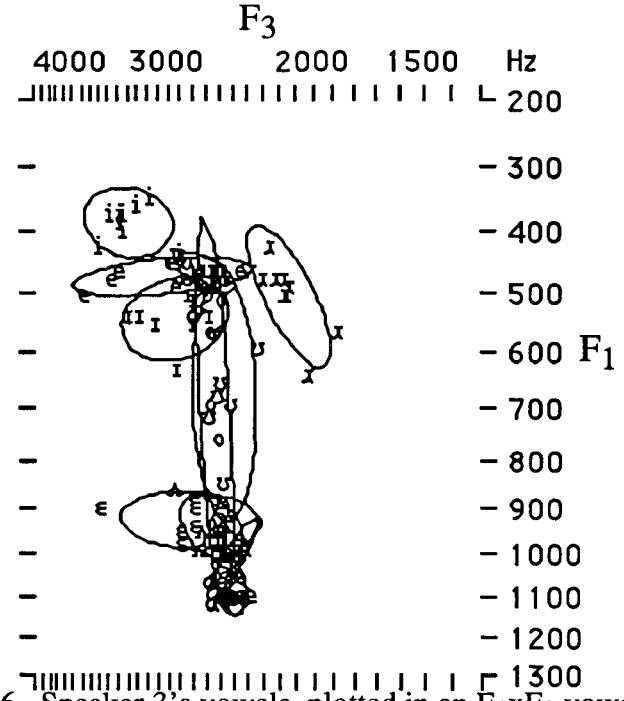


Figure 6. Speaker 3's vowels, plotted in an F_1xF_3 vowel space.

Speaker 3 again shows relatively little variation in many vowels. However, her /ʊ/ and /o/ categories vary widely in F_1. Unlike the previous speakers, S3's /ɛ/ category shows relatively little variation in F_1, but one token of /ɛ/ has a very low second formant, producing a relatively large standard deviation in what is otherwise a vowel with relatively little variation.

Speaker 4

Speaker 4's formant averages are given in Table V. Her F_1xF_2 vowel space is illustrated in Figure 7; her F_1xF_3 vowel space is illustrated in Figure 8.

Table V. Formant averages for Speaker 4. Units are Hertz; standard deviations are in parentheses; n=9 for each vowel.

	F_1	F_2	F_3	
i	362 (28)	2824 (66)	3569 (94)	i
ɪ	428 (39)	2433 (84)	3184 (216)	ɪ
e	412 (23)	2619 (160)	3186 (144)	e
ɛ	963 (152)	2081 (267)	2933 (263)	ɛ
æ	1079 (76)	1858 (59)	2766 (174)	æ
u	371 (27)	1482 (253)	2811 (43)	u
ʊ	419 (51)	1588 (128)	2854 (218)	ʊ
o	412 (29)	1218 (82)	2866 (69)	o
ɑ	1020 (58)	1375 (104)	2825 (150)	ɑ
ʌ	934 (96)	1674 (97)	2878 (154)	ʌ
ɹ	408 (33)	1389 (53)	1680 (62)	ɹ

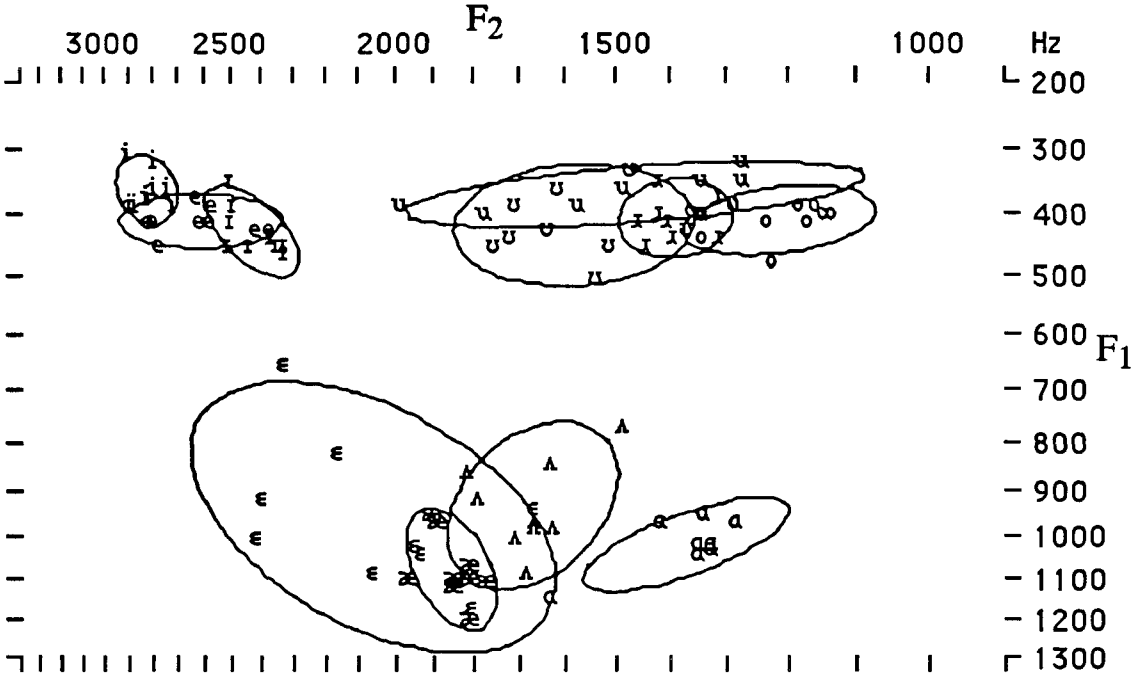


Figure 7. Speaker 4's vowels, plotted in an F_1xF_2 vowel space.

Speaker 4 shows extremely disparate tokens of /ɛ/, particularly in comparison to /æ/, to the point that the relatively small /æ/ space is completely encompased by the /ɛ/ space.. This variation is not the result of the three different contexts in which the vowel appeared; that is, it is not the case that three tokens of /ɛ/ are relatively higher than or backer than the other six. It looks almost as if S4 is producing tokens randomly in the lower-front space for /ɛ/ instead of producing a vowel with formants at particular, specific frequencies. Note also that S4 does show clear separation between the 'fronted' tokens vs the other tokens of /ʊ/, as did S1-3. S4 also differs from the preceeding speakers in that she shows very little variation in F_1 of /ʊ/ and /o/. This in turn produces a clear

separation of the higher vowels from the lower ones. This trend is less obvious in the preceeding speakers.

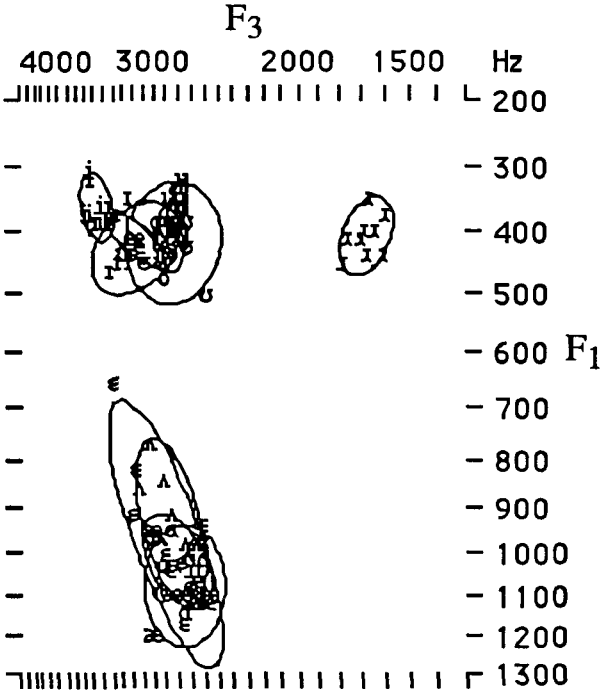


Figure 8. Speaker 4's vowels, plotted in an F_1xF_3 vowel space.

Speaker 5

Speaker 5's formant averages are given in Table VI. Her F_1xF_2 vowel space is illustrated in Figure 9; her F_1xF_3 vowel space is illustrated in Figure 10.

Table VI. Formant averages for Speaker 5. Units are Hertz; standard deviations are in parentheses; n=9 for each vowel.

	F_1	F_2	F_3	
i	365 (42)	2946 (91)	3774 (126)	i
ɪ	491 (33)	2567 (107)	3719 (220)	ɪ
e	469 (71)	2813 (111)	3732 (247)	e
ɛ	691 (119)	2352 (177)	3481 (309)	ɛ
æ	961 (79)	1939 (72)	3063 (388)	æ
u	427 (41)	2163 (338)	3132 (164)	u
ʊ	543 (62)	1811 (118)	3289 (143)	ʊ
o	546 (50)	1884 (93)	3163 (211)	o
ɑ	918 (137)	1454 (118)	2931 (170)	ɑ
ʌ	733 (84)	1959 (97)	3416 (226)	ʌ
ɹ	544 (78)	1871 (73)	2725 (85)	ɹ

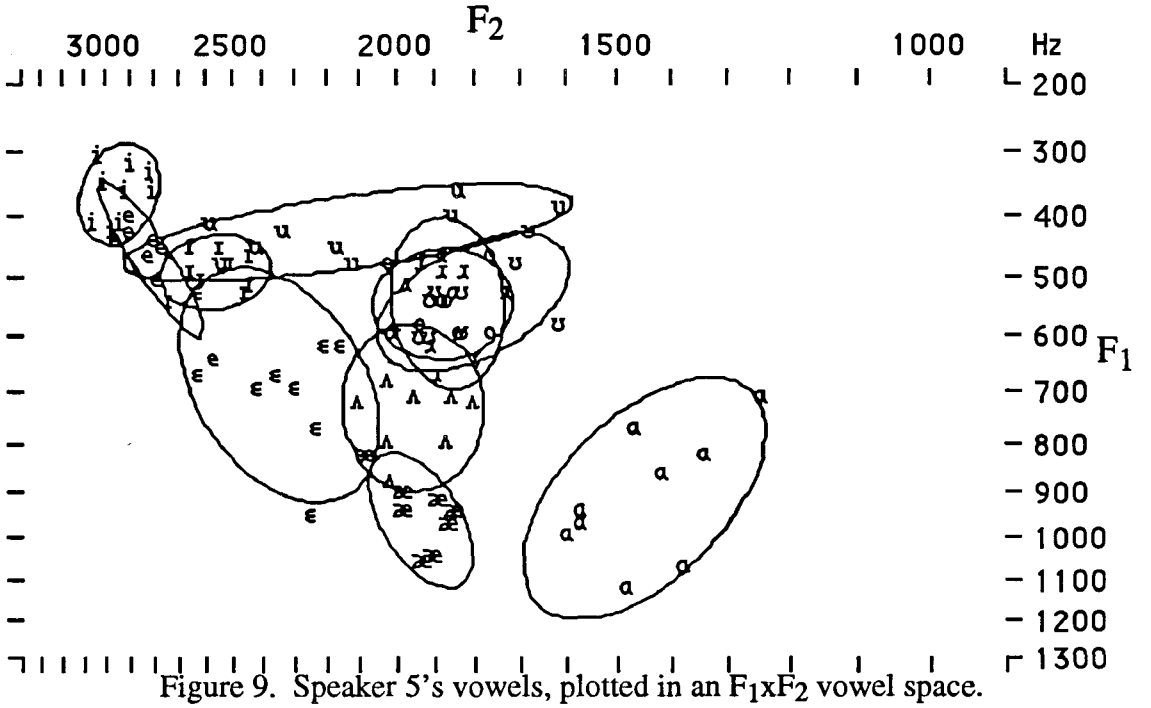


Figure 9. Speaker 5's vowels, plotted in an F_1xF_2 vowel space.

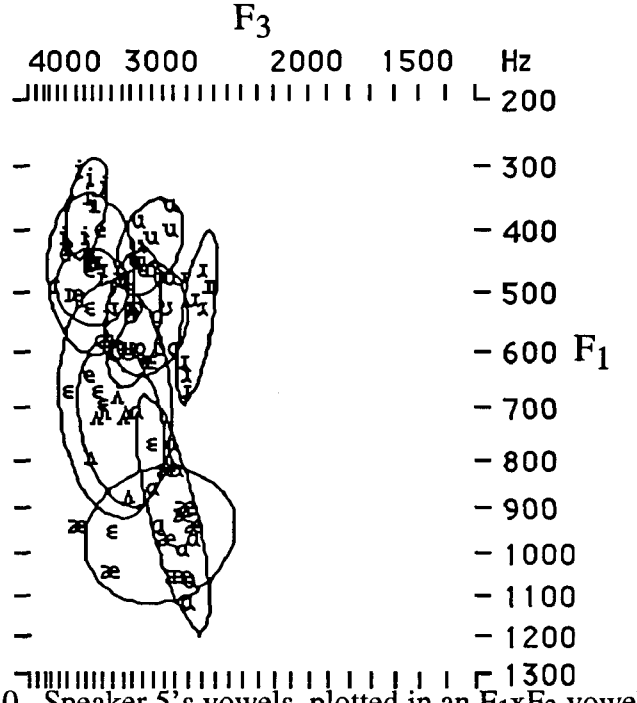


Figure 10. Speaker 5's vowels, plotted in an F_1xF_3 vowel space.

Speaker 5's /u/, /ʊ/, and /o/ tokens all have second formants above 1500 Hz, suggesting that these vowels are never fully back or round. Further, she has nearly complete overlap between her /ʊ/ and /o/ tokens whereas the preceding speakers had much more separated /ʊ/ and /o/ categories. However, these vowels could probably be distinguished by dynamic (i.e. transitional) factors not recorded in this study. Similar to S4, S5 had relatively wide variation in her /ɛ/ category. Unlike the preceding speakers, she also has a lot of variation in her /ɑ/ tokens. This

may be due, at least in part, to a knowledge of a distinction between /ɑ/ and /ɔ/, which is not normally maintained in this dialect (but see Speaker E).

Unlike all four previous speakers, S5 does not clearly separate the third formant of /ɹ/ from the third formants of the other vowels. That is, while the F_3 of /ɹ/ is lower than most of the other vowels, it is not nearly so low as it is for other speakers, whose third formants of /ɹ/ were often 1000 Hz or more lower than the more neutral F_3 of the other vowels. S5's /ɹ/'s in all contexts were auditorily 'weaker' than for the other speakers in the study.

Speaker A

Speaker A is the first of the male speakers to be presented. Sa's formant averages are given in Table VII. His F_1xF_2 vowel space is illustrated in Figure 11; his F_1xF_3 vowel space is illustrated in Figure 12.

Table VII. Formant averages for Speaker A. Units are Hertz; standard deviations are in parentheses; n=9 for each vowel.

	F_1	F_2	F_3	
i	262 (27)	2422 (38)	3003 (171)	i
ɪ	385 (21)	1831 (62)	2503 (89)	ɪ
e	359 (27)	2215 (135)	2643 (132)	e
ɛ	573 (78)	1692 (61)	2440 (113)	ɛ
æ	782 (94)	1636 (54)	2437 (105)	æ
u	286 (23)	1379 (210)	2383 (56)	u
ʊ	400 (17)	1306 (81)	2363 (73)	ʊ
o	402 (11)	1080 (57)	2450 (34)	o
ɑ	782 (45)	1217 (44)	2389 (74)	ɑ
ʌ	611 (73)	1414 (94)	2372 (77)	ʌ
ɹ	416 (25)	1321 (60)	1656 (29)	ɹ

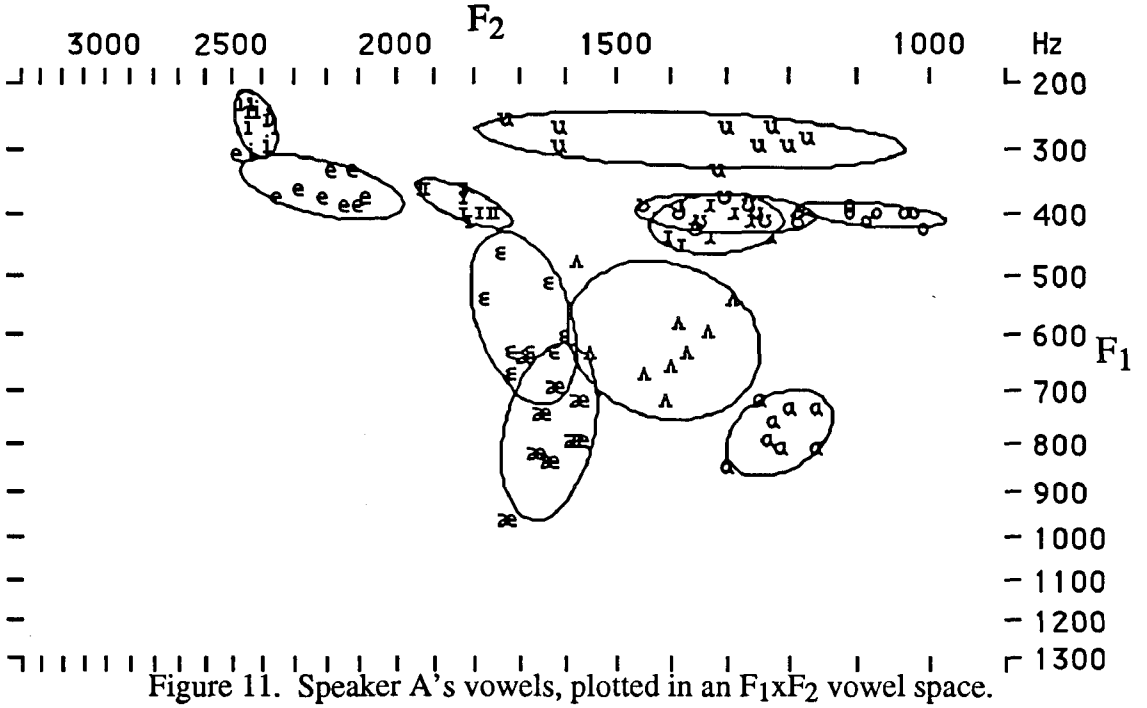


Figure 11. Speaker A's vowels, plotted in an F_1xF_2 vowel space.

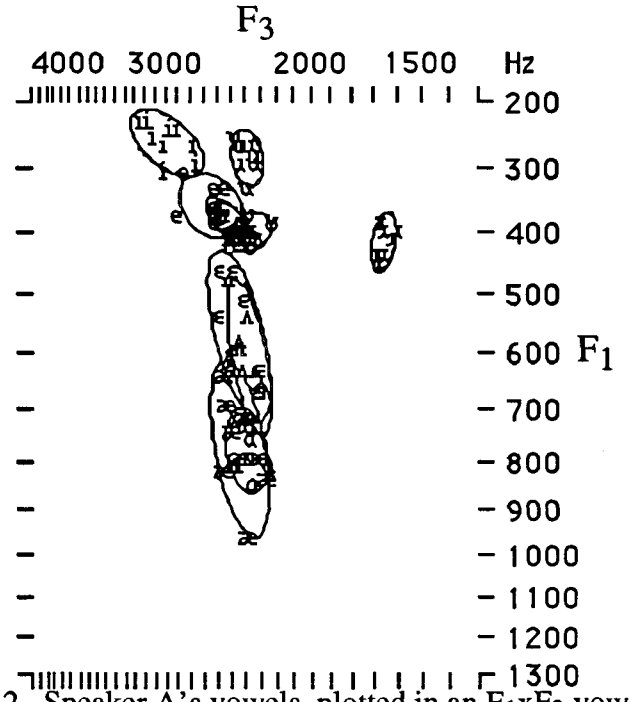


Figure 12. Speaker A's vowels, plotted in an F_1xF_3 vowel space.

As predicted by acoustic theory as well as previous research in sex-differences in vowel production (e.g. the classic studies by Peterson & Barney, 1952, and Fant, 1973, among others), Sa shows lower formants over all than the women, resulting in a smaller overall space located slightly 'up and right' of the women's space (in an F_1xF_2 plot with the origin in the upper right, such as those used in this paper).

Speaker A shows very little variation in F_1 in the higher vowels, but more in /ɛ/, /æ/ and /ʌ/. Some variation in F_2 is observable in several vowels, but not to the degree as in /u/, which is primarily the product of fronted /u/ after coronals.

Speaker B

Speaker B's formant averages are given in Table VIII. His F_1xF_2 vowel space is illustrated in Figure 13; his F_1xF_3 vowel space is illustrated in Figure 14.

Speaker B shows verly little separation of vowles in the F_1 dimension, producing relatively high first formant values even for the 'low' vowels /A/ and /æ/. That is, Sb does not appear to exploit as full a range of first formant values as possible.

Table VIII. Formant averages for Speaker B. Units are Hertz; standard deviations are in parentheses; n=9 for each vowel.

	F_1	F_2	F_3	
i	293 (21)	2710 (62)	3251 (100)	i
ɪ	398 (27)	1866 (38)	2651 (45)	ɪ
e	386 (31)	1967 (64)	2629 (89)	e
ɛ	498 (37)	1727 (26)	2620 (52)	ɛ
æ	532 (13)	1640 (44)	2740 (302)	æ
u	323 (29)	1545 (216)	2665 (354)	u
ʊ	428 (13)	1500 (63)	2606 (135)	ʊ
o	418 (13)	1338 (128)	2442 (253)	o
ɑ	548 (52)	1226 (59)	2564 (155)	ɑ
ʌ	509 (41)	1540 (64)	2570 (53)	ʌ
ɹ	396 (13)	1439 (54)	1725 (55)	ɹ

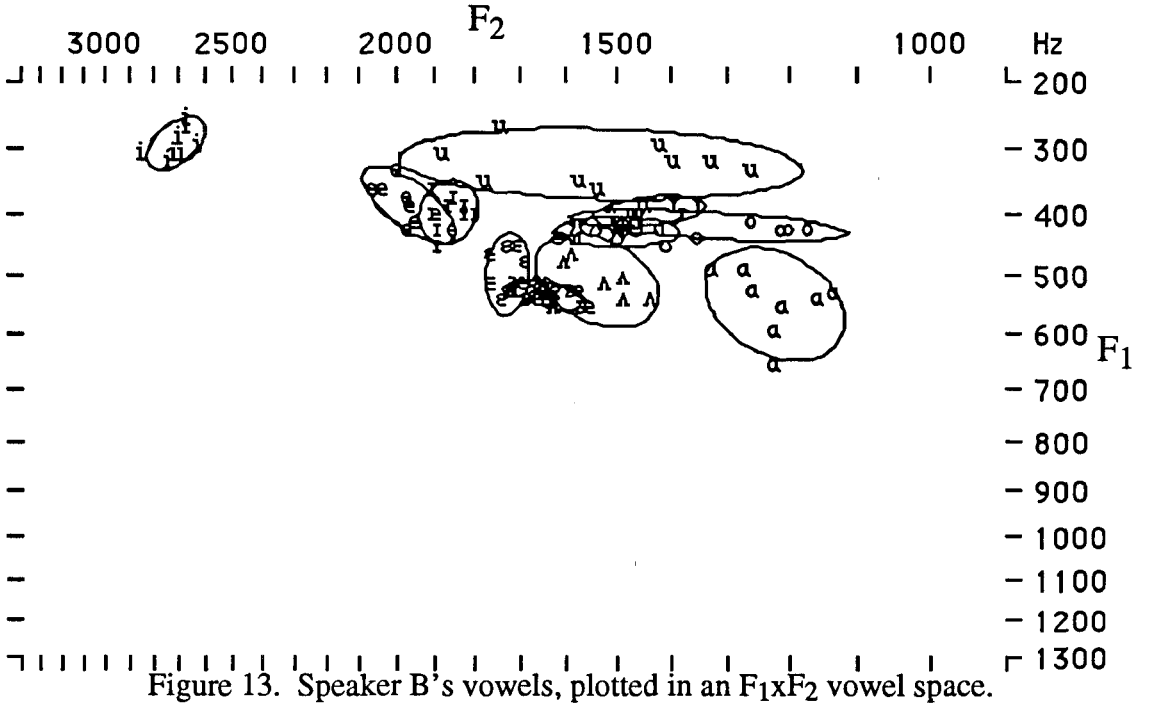


Figure 13. Speaker B's vowels, plotted in an F_1xF_2 vowel space.

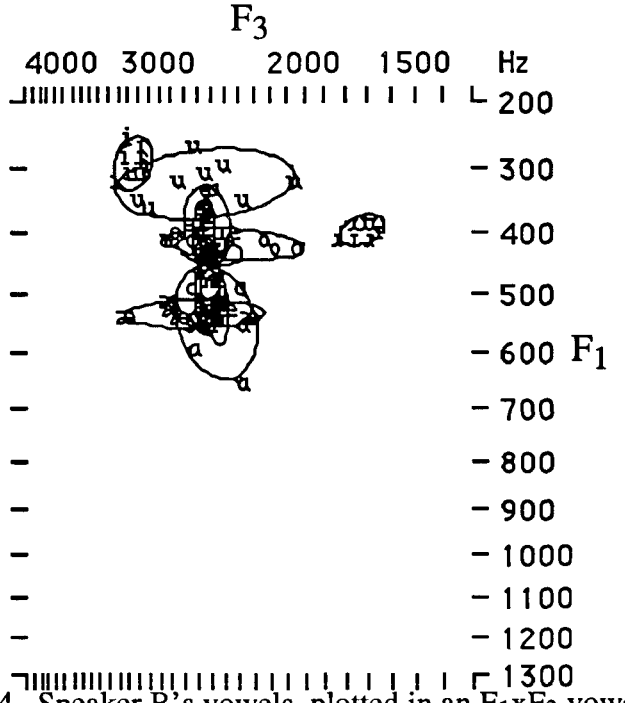


Figure 14. Speaker B's vowels, plotted in an F_1xF_3 vowel space.

Speaker C

Speaker C's formant averages are given in Table IX. His F_1xF_2 vowel space is illustrated in Figure 15; his F_1xF_3 vowel space is illustrated in Figure 16.

Table IX. Formant averages for Speaker C. Units are Hertz; standard deviations are in parentheses; n=9 for each vowel.

	F_1	F_2	F_3	
i	296 (30)	2216 (42)	2858 (76)	i
ɪ	406 (19)	1829 (82)	2670 (103)	ɪ
e	412 (37)	2007 (91)	2884 (235)	e
ɛ	530 (48)	1676 (51)	2632 (142)	ɛ
æ	644 (63)	1639 (65)	2761 (432)	æ
u	336 (21)	1412 (243)	2409 (163)	u
ʊ	437 (22)	1385 (132)	2548 (228)	ʊ
o	435 (25)	1128 (51)	2629 (357)	o
ɑ	719 (31)	1235 (62)	2369 (313)	ɑ
ʌ	551 (60)	1399 (50)	2673 (399)	ʌ
ɹ	429 (41)	1335 (61)	1656 (22)	ɹ

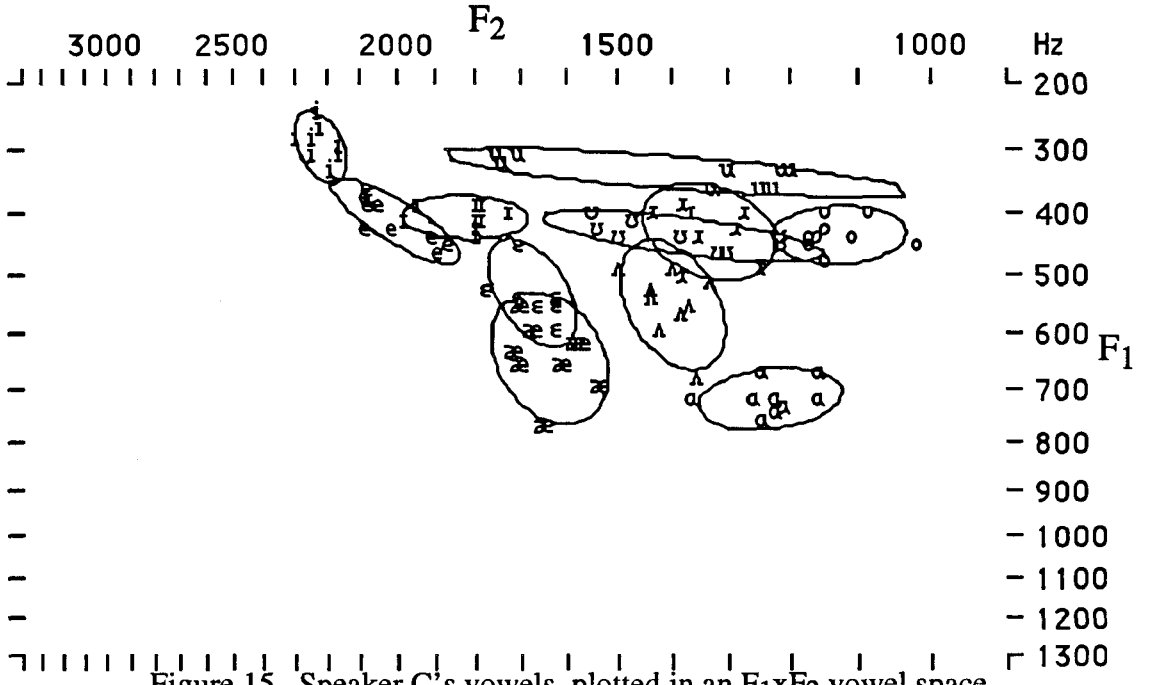


Figure 15. Speaker C's vowels, plotted in an F_1xF_2 vowel space.

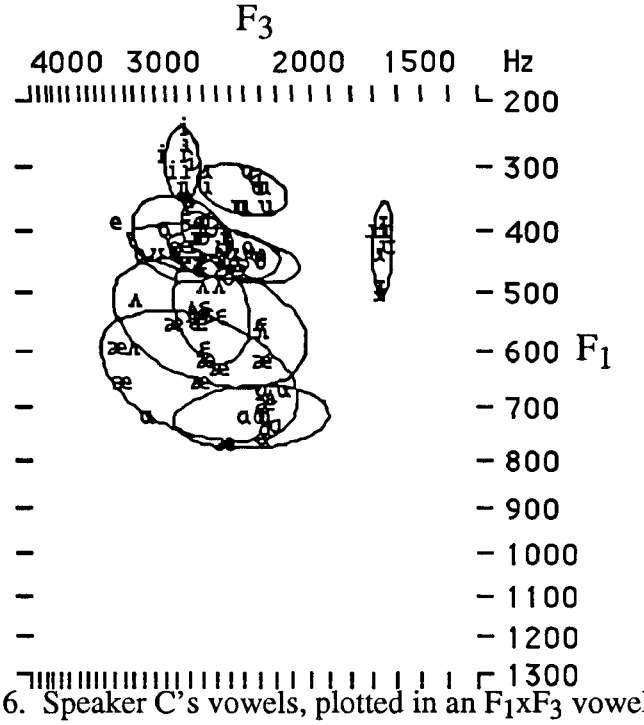


Figure 16. Speaker C's vowels, plotted in an F_1xF_3 vowel space.

With the exception of the fronted /u/ tokens and the variation in F_2 of /ʊ/, there does not appear to be anything particularly odd or remarkable in the vowels of Speaker C. Note in Table IX and Figure 16, however, the extremely narrow distribution of F_3 values for /ɹ/ as compared with the other vowels.

Speaker D

Speaker D's formant averages are given in Table X. His F_1xF_2 vowel space is illustrated in Figure 17; his F_1xF_3 vowel space is illustrated in Figure 18.

Table X. Formant averages for Speaker D. Units are Hertz; standard deviations are in parentheses; n=9 for each vowel.

	F_1	F_2	F_3	
i	273 (9)	2160 (31)	2736 (105)	**i**
I	450 (18)	1765 (86)	2518 (63)	**I**
e	432 (32)	1972 (111)	2571 (73)	**e**
E	522 (57)	1633 (50)	2457 (74)	**E**
æ	704 (55)	1563 (42)	2428 (59)	**æ**
u	318 (21)	1366 (211)	2256 (59)	**u**
U	447 (6)	1401 (66)	2339 (87)	**U**
o	455 (38)	1170 (58)	2314 (47)	**o**
A	754 (34)	1269 (35)	2362 (83)	**A**
ʌ	601 (44)	1445 (41)	2401 (92)	**ʌ**
ɹ	457 (22)	1448 (63)	1788 (88)	**ɹ**

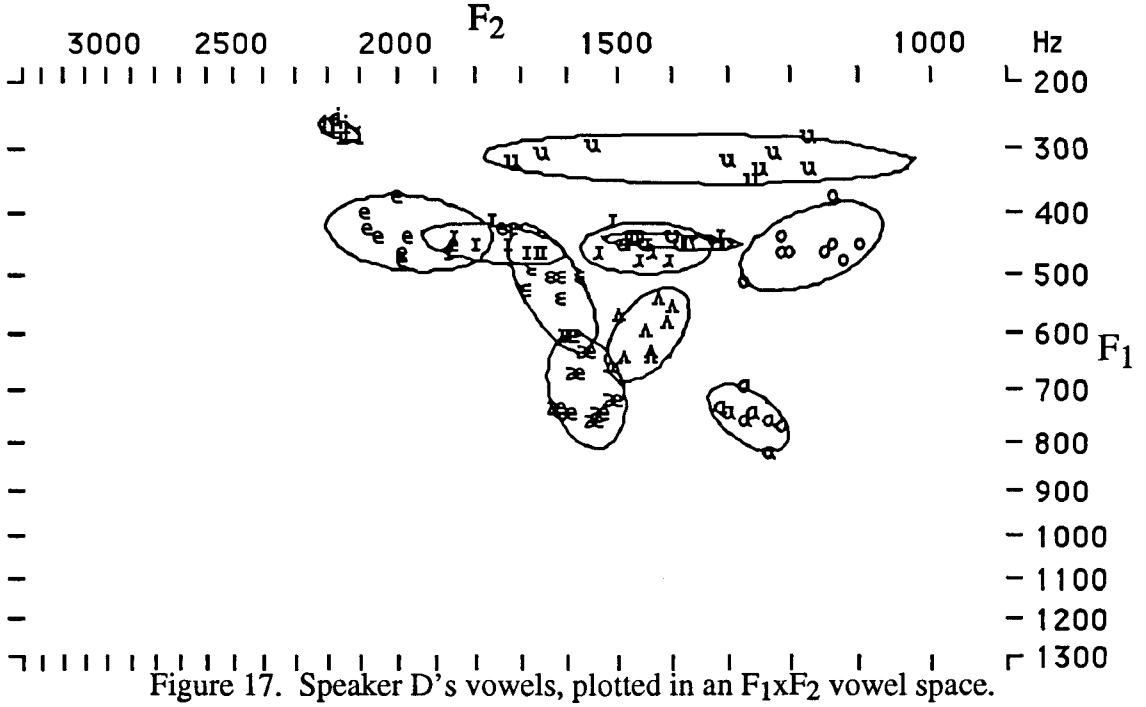


Figure 17. Speaker D's vowels, plotted in an F_1xF_2 vowel space.

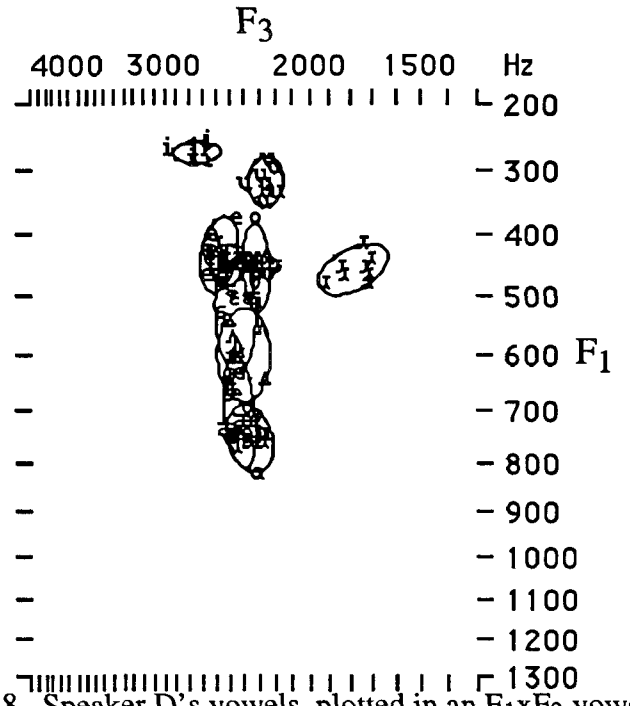


Figure 18. Speaker D's vowels, plotted in an F_1xF_3 vowel space.

Speaker D's vowel space is generally similar to Sc's.

Speaker E

Speaker E's formant averages are given in Table XI. His F_1xF_2 vowel space is illustrated in Figure 19; his F_1xF_3 vowel space is illustrated in Figure 20.

Table XI. Formant averages for Speaker E. Units are Hertz; standard deviations are in parentheses; n=9 for each vowel.

	F_1	F_2	F_3	
i	333 (22)	2346 (118)	2846 (269)	i
ɪ	468 (25)	1709 (63)	2571 (163)	ɪ
e	442 (48)	2077 (151)	2676 (54)	e
ɛ	594 (56)	1602 (55)	2509 (156)	ɛ
æ	782 (25)	1557 (41)	2425 (70)	æ
u	357 (9)	1364 (170)	2500 (52)	u
ʊ	502 (45)	1298 (85)	2493 (131)	ʊ
o	487 (28)	1194 (63)	2496 (77)	o
ɑ	727 (126)	1202 (116)	2430 (56)	ɑ
ʌ	670 (44)	1355 (62)	2428 (87)	ʌ
ɪ	477 (41)	1306 (41)	1689 (48)	ɪ

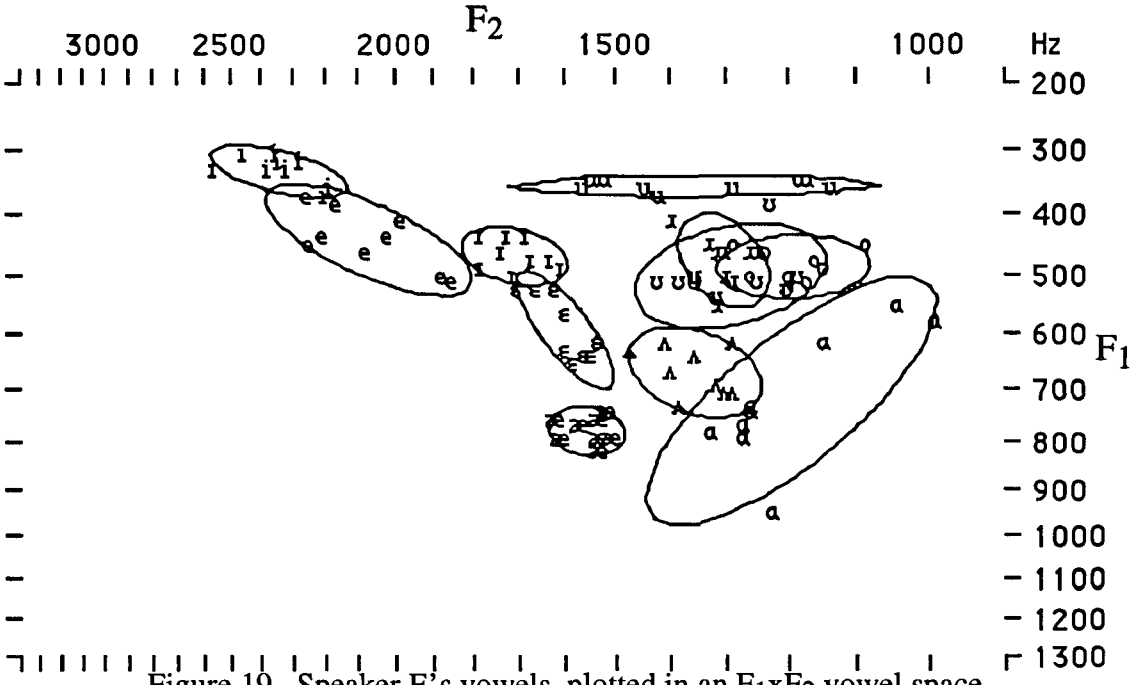


Figure 19. Speaker E's vowels, plotted in an F_1xF_2 vowel space.

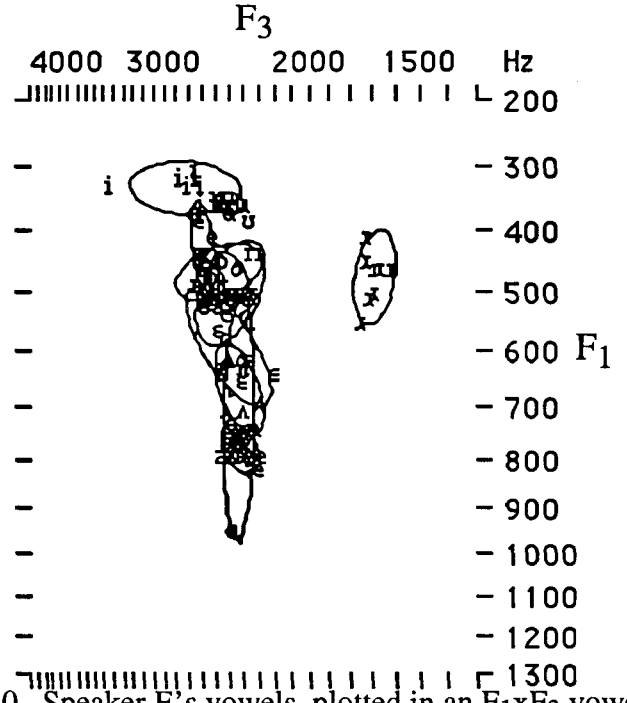


Figure 20. Speaker E's vowels, plotted in an F_1xF_3 vowel space.

Speaker E's /e/ tokens are fairly broadly scattered, unlike this vowel for the other men (Sad). This does not appear to be particularly the result of phonological context. Like female Speaker 4, Se does not appear to clearly differentiate fronted /u/ in "duke" from the other, more back /u/ used in "hoot" and "boot". Se is the only speaker in this study who maintains a consistent distinction between the /ɑ/ of "tock" and "hod" and the vowel in "bought", which for Se should be transcribed as [ɔ]. It may be misleading to collapse Se's /ɑ/ and /ɔ/ categories, but no more so than to discard his productions of [ɔ].

Se shows wider covariation in F_1 and F_2 of /e/ than many of the other speakers, but this variation occurs for the most part outside the areas bounded by other vowel categories. That is, it does not lead to any overlap with other categories.

Summary

The following table summarizes the formant frequency measurements (averages and standard deviations) for the speakers in this study. In each column, the average for female speakers is given on the left, and for males on the right.

Table XII. Formant frequencies of American vowels produced by women (left) and men (right). All values are rounded to the nearest whole number. Standard deviations are in parentheses. Units are Hertz.

	F_1		F_2		F_3		
	W	*M*	*W*	*M*	*W*	*M*	
i	358 (38)	291 (33)	2847 (195)	2371 (205)	3557 (236)	2939 (236)	i
ɪ	489 (59)	421 (39)	2428 (123)	1800 (87)	3262 (321)	2583 (118)	ɪ
e	454 (47)	406 (46)	2681 (141)	2048 (144)	3316 (342)	2681 (167)	e
ɛ	883 (148)	543 (65)	2179 (186)	1666 (65)	3087 (346)	2531 (136)	ɛ
æ	1055 (75)	689 (110)	1802 (145)	1607 (62)	2793 (277)	2558 (282)	æ
u	398 (52)	324 (31)	1762 (397)	1413 (213)	2867 (272)	2442 (220)	u
ʊ	541 (113)	443 (41)	1682 (179)	1378 (113)	2894 (331)	2470 (171)	ʊ
o	561 (144)	439 (38)	1447 (257)	1182 (114)	2892 (310)	2466 (217)	o
ɑ	991 (80)	706 (105)	1372 (91)	1230 (70)	2699 (200)	2423 (175)	ɑ
ʌ	887 (102)	588 (76)	1791 (161)	1431 (88)	2956 (319)	2489 (217)	ʌ
ɝ	508 (83)	435 (41)	1570 (194)	1370 (82)	2041 (395)	1703 (72)	ɝ

The first and second formant values given above in Table XII are plotted in a standard F_1xF_2 vowel space in Figure 21 (for the women) and Figure 22 (for the men).

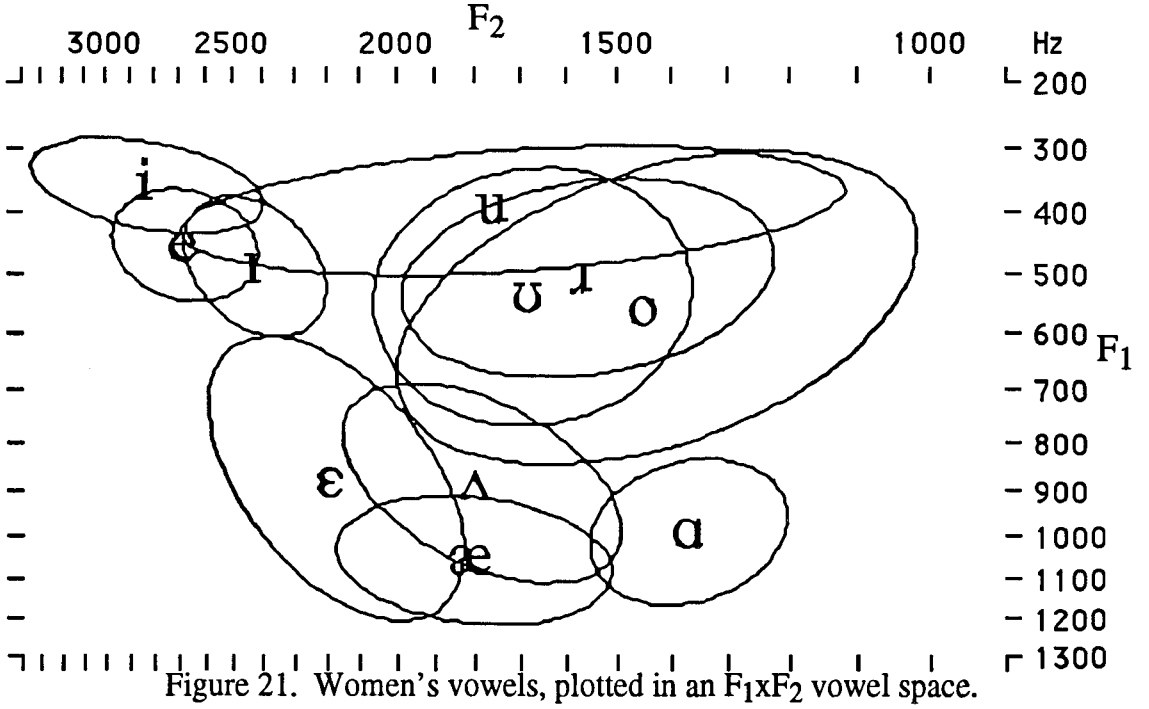


Figure 21. Women's vowels, plotted in an F_1xF_2 vowel space.

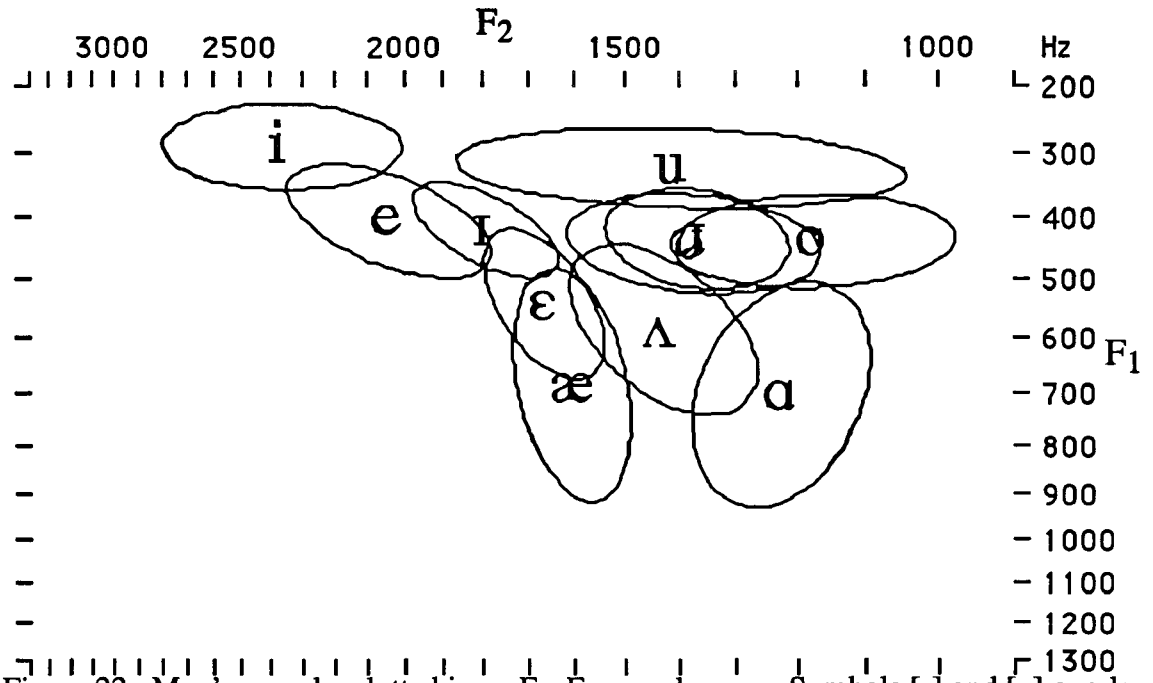


Figure 22. Men's vowels, plotted in an F_1xF_2 vowel space. Symbols [ɪ] and [u] overlap.

In Figures 23 and 24, F_3 (along the horizontal axis) is plotted against F_1 (vertical axis). Again, ellipses enclose areas defined by two standard deviations from the mean. Because of the tremendous overlap in F_3 values, only the means for [ɪ] are labelled.

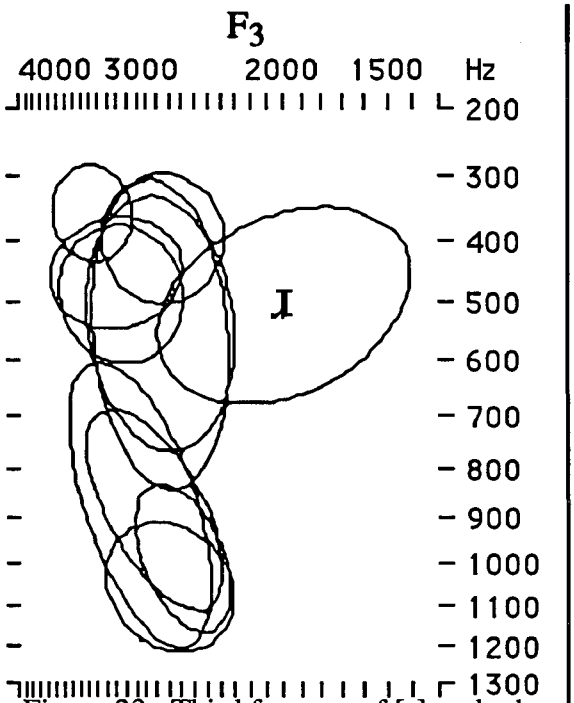


Figure 23. Third formant of [ɪ] and other vowels for women.

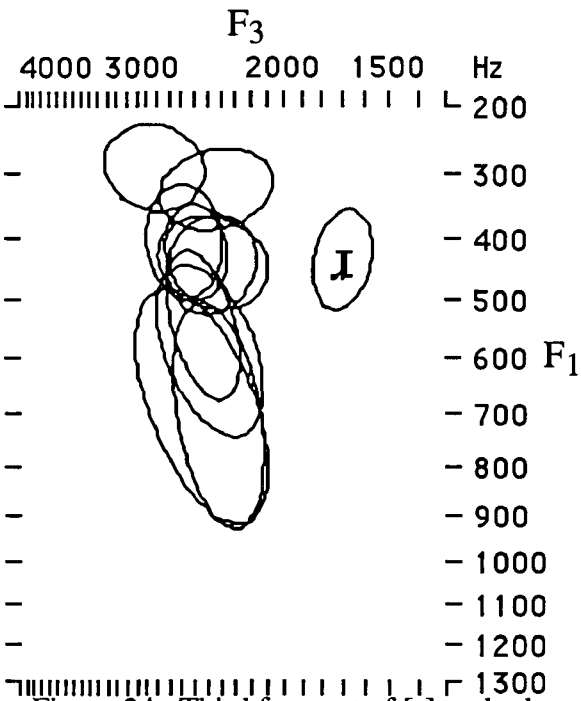


Figure 24. Third formant of [ɪ] and other vowels for men.

4. General discussion

Sex differencs in between and within-category variance

As can be seen in Figures 21 and 22, the women in this study appear to vary a great deal more than the men. However, much of that impression is the result of the large standard deviations associated with the /ɛ/ and /o/ categories, and other vowels. The comparatively large variations in /ɛ/ and /o/ are more important, reflecting as they do real variation within and across speakers. Speaker 4 in particular seemed not to have an /ɛ/ target, in the way she clearly had one for /i/.

Among the men, however, the disagreements appear to be a matter of individual detail among specific targets. True, some men varied one formant of one vowel while others varied another, but this kind of disagreement among speakers was relatively small. Each of the male speakers appears to have settled, more or less, on some kind of acoustic target for each vowel, and the variation among them is mostly a matter of individual differences. However, in the /ɛ/ and /o/ categories many of the women appear *not* to have settled on a particular target, resulting in wider variation overall for /ɛ/ and /o/. The female speakers (S1-5) seemed to disagree even on the relative stability of particular formants. Recall Speaker 3, whose F_2 of /ɛ/ varied more than her F_1, where Speakers 1 and 2 had more variation in F_1 than F_2. Like the men (Sa-e), the female speakers *have* selected individual targets for /i/, which vary little from one speaker to the next, resulting in relatively smaller standard deviations.

Table XII summarizes the results of F-tests of the difference in variance between the women's and men's data in this study. Because each vowel category contains the same number of measurements in both the male and female populations (45), the Degrees of Freedom in each case is the same (44). The F-values are equivalent to the variance ratios (i.e, the ratio of the squared standard deviations). Where the ratio (F-value) is greater than one, the women's variance is greater. When the F-value is less than one, the men's variance is greater.

Table XIII. Results of F-tests of men's and women's formant frequencies (significant F-values in bold; F-values less than one indicated by *).

	F_1		F_2		F_3		
	F-value	*P-value*	*F-value*	*P-Value*	*F-value*	*P-Value*	
i	1.290	.4022	.900*	.7293	1.007	.9828	i
ɪ	**2.300**	**.0067**	**2.014**	**.0222**	**7.388**	**<.0001**	ɪ
e	1.050	.8726	.963*	.9011	**4.187**	**<.0001**	e
ɛ	**5.239**	**<.0001**	**8.087**	**<.0001**	**6.431**	**<.0001**	ɛ
æ	**.470***	**.0139**	**5.549**	**<.0001**	.968*	.9134	æ
u	**2.773**	**.0010**	**3.465**	**<.0001**	1.534	.1595	u
ʊ	**7.570**	**<.0001**	**2.525**	**.0027**	**3.774**	**<.0001**	ʊ
o	**14.112**	**<.0001**	**5.057**	**<.0001**	**2.045**	**.0195**	o
ɑ	.584*	.0782	1.703	.0808	1.303	.3834	ɑ
ʌ	**1.834**	**.0472**	**3.357**	**.0001**	**2.175**	**.0114**	ʌ
ɹ	**4.123**	**<.0001**	**5.634**	**<.0001**	**30.310**	**<.0001**	ɹ

Some interesting patterns emerge from the F-tests which are not readily apparent from the Bark-scaled plots in Figures 21-24. In six of the eleven vowel categories (/ɪ,ɛ,ʊ,o,ʌ,ɹ/), significant between-sex differences were found in the variance of all three formants, with the women varying more. Only in five of the thirty-three possible cases did the F-value suggest greater variance among the men's tokens than the women's. These were F_1 of /æ/ and /ɑ/, F_2 of /i/ and /ɛ/, and F_3 of /æ/. Of these, only the difference in F_1 of /æ/ was significant.

The F_1 of /æ/ did show significantly greater variance in the men's tokens than the women's; this is reflected in Figure 22, where the two-standard deviation ellipse for /æ/ is longer in the vertical (F_1) dimension. The /æ/ ellipse for the women in Figure 1 has its longer axis in the

horizontal (F_2) dimension; the F_2 of /æ/ showed significantly greater variance among the women's tokens than the men's. The variance in the /æ/ category did not show a significant difference in F_3. The overall variance in the /æ/ a category is not overwhelmingly greater in the men's tokens than the women's.

In sum, all three formants in the women's tokens varied significantly more than the men's for six of the 11 vowel categories tested (/ɪ,ɛ,ʊ,o,ʌ,ɻ/). Of the remaining five vowels, /i/ and /ɑ/, the most peripheral vowels in the F1xF2 space, never show significant differences in the variance in any formant.

In ten of the 33 cases (3 formants of 11 vowels), no significant difference in the variance was found between the men and thewomen. In one case (F_1 of /æ/), the men showed significantly more variance than the women. In the remaining 22 cases, the women's formants varied more than men. On the balance, then, it can be concluded that women's formant frequencies generally exhibit more within-category variance than men. As stated above, some of this variation is the result of speaker-to-speaker differences in phonetic detail. In other cases, such as /o/ and /ɛ/, greater variation appears to be less principled.

Dissimilar distribution of women's and men's mean vowel positions

Comparing the means in Table XII, the women's formant frequencies are always higher than the men's. The mean differences range from a low of 47 Hz (F_1 of /e/) to a high of 679 Hz (F_3 of /ɪ/). All three formants of every vowel category showed significant differences as a function of speaker sex. Table XIV summarizes the results of unpaired t-tests to which the data were subjected. For every t-value, the Degrees of Freedom=88 and $p \leq .0001$.

Table XIV. Mean differences between women's and men's formants (F_W-F_M). Frequency values are rounded to the nearest whole number; units are Hertz.

	F_1		F_2		F_3		
	Mean Diff.	t-Value	Mean Diff.	t-Value	Mean Diff.	t-Value	
i	66	8.897	476	11.270	616	12.414	i
ɪ	67	6.450	628	28.035	679	13.306	ɪ
e	47	4.832	633	21.092	635	11.203	e
ɛ	339	14.116	513	17.472	555	10.030	ɛ
æ	367	18.496	195	8.315	235	3.985	æ
u	74	8.247	349	5.195	425	8.151	u
ʊ	98	5.458	304	9.649	424	7.637	ʊ
o	122	5.511	265	6.314	427	7.557	o
ɑ	285	14.484	142	8.322	276	6.968	ɑ
ʌ	298	15.710	361	13.168	467	8.120	ʌ
ɻ	73	5.253	200	6.370	338	5.656	ɻ

In general, the /ɛ/ category is quite a bit lower in the women's vowel space than in the men's. While the men's front vowels (excluding either /e/ or /ɪ/) seem more or less evenly distributed in the height (F_1) dimension, the women's vowels seem to be unevenly distributed. The mean F_1 of /o/ is 561 Hz; the mean F_1 of /ɛ/ is 883 Hz. That is, where the men exhibit a relatively even distribution of vowel categories in the height dimension, the women seem to divide the vowels into a higher group and a lower group, with a gap of about 300 Hz between the two.

More striking in the F_1xF_2 vowel space is the distribution of the lower vowels in the F_2 dimension. Note that for both men and women /ɑ/ is quite low and back. /æ/ has a higher F_2 (i.e., is more front) than /ɑ/. However, in the men's vowel space (Figure 22), /æ/ is clearly a front vowel, having F_2 values similar to the /ɛ/ category and being considerably more front than /ʌ/. This is not the case for the women's vowel space in Figure 21, where /æ/, /ʌ/, and /ʊ/ have simlar

F_2s, making /æ/ acoustically a central vowel. Note also that the highly variable F_1s of /ɛ/ seem to be expanding into the low front space 'vacated' by /æ/, as well as into the sparsely populated middle frequency region.

For both the women and men, the relatively central position of /u/ (i.e., its relatively high F_2) is not particularly surprising. As noted earlier, at least part of the variation in the F_2 of this vowel comes from the 'duke' tokens. In addition to this alternation, however, is a more general change affecting this segment, as well as the other back vowels. In southern California, the back vowels are often unround. Without lip rounding, the formants, particularly F_2, may be higher than expected if the vowel were fully rounded. In general, the back vowels in this dialect are much more central than has been reported in other dialects. This point will be illustrated again later.

The dissimilarity of the men's and women's vowel spaces is further exemplified in Figures 24 and 25. Figure 24 is the men's and women's F_1 and F_2 means overlaid in the same plot. Figure 25 is the men's and women's F_1 and F_2 means from Peterson & Barney (1952), plotted to the same scale. For better visual comparisons, the /e/ means were removed from the plot in Figure 5. Peterson & Barney (1952) did not include /e/ or /o/ in their data. Recall that for the men, the /ɨ/ and /ʊ/ category means overlapped. In Figure 24, note the overlap of the men's /ɛ/ with the women's /ʊ/.

The back vowels of the speakers in the present study are acoustically quite central compared with those of Peterson & Barney's speakers. The speakers in this study seem to avoid F_2 frequencies below 1100 Hz, where Peterson & Barney's female speakers appear quite happy to produce F2 frequencies lower than this.

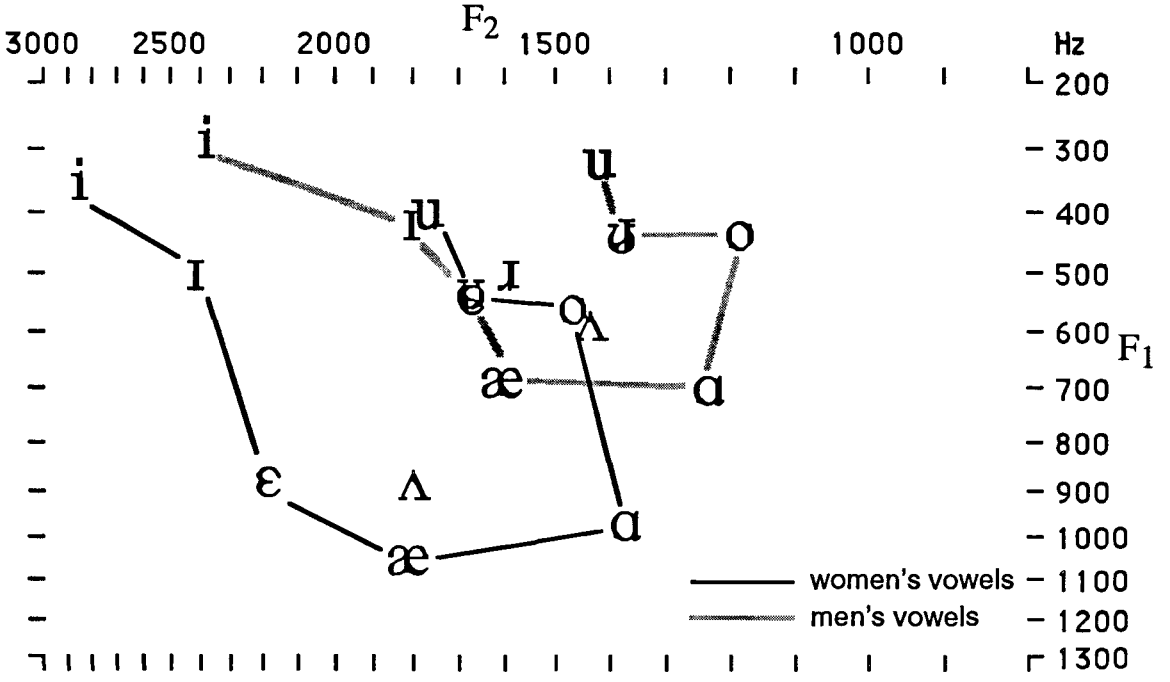


Figure 24. Men's and women's vowel means from the present study.

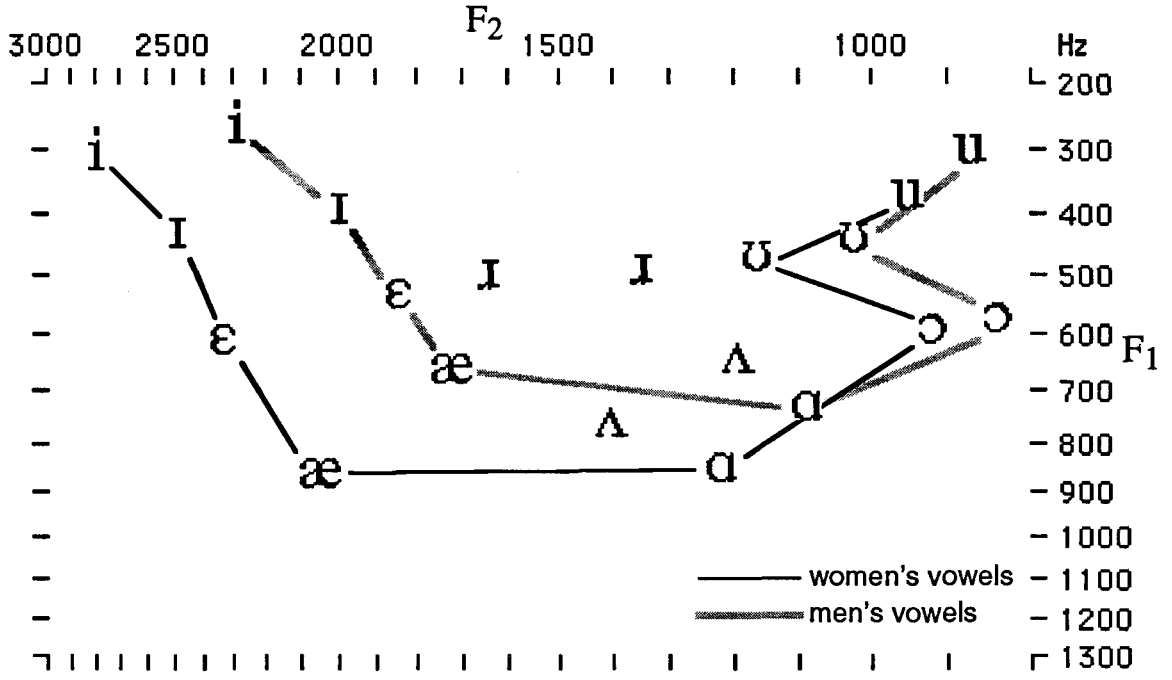


Figure 25. Men's and women's vowel means from Peterson & Barney (1952).

The distribution of mean vowel positions in the present study is quite different for men than for women; in contrast, Peterson & Barney's data look more like the same linear relationships being expressed over a larger area. That is, 'normalizing' for speaker sex in Peterson & Barney's data set is a matter of scaling. However, the present data are more difficult for the problem of normalization.

Fant (1973) has discussed male-to-female vowel scaling, noting that, while in general, women's formants are approximately 20% higher than men's,

> "A range of typical and substantial deviations from this rule is concealed if the data are averaged over the whole vowel system.... [T]he female-to-male relations are typically different in the three groups of (1) rounded back vowels, (2) very open unrounded vowels, and (3) close front vowels." (Fant, 1973:84).

The observation that male-to-female vowel formant scaling is not a single procedure but one that differs vowel to vowel is well taken here. However, Fant sought a physiological explanation as to why the three vowel groups he mentioned in the quoted passage should behave differently. It is not at all clear if a physiological explanation will suffice to explain the present data.

The asymmetry found in the distribution of vowel categories within the men's and the women's vowel spaces should not come as particular surprise. The present data are at least impressionistically comparable to Disner's (1983) findings that phonologically similar vowel spaces in different languages can differ in idiosyncratic ways, although certain points of similarity, notably at the 'point vowels' /i,ɑ,u/, are likely. Disner found this to be true in different languages, but it is not unexpected that such differences should appear across dialects of the same language. The relative stability of /i/ and /ɑ/ in this dialect of American English for both sexes is probably related to their peripherality in the F_1 and F_2 dimensions. The observed differences in vowel mean distributions might be parts of a more general trend, for instance, the relatively lower settings for the low vowels in the women's dialect. They might also be more random, as the relative positions of the /ɛ/ and /æ/ categories. It remains to be shown if any of these differences are the result of social or historical pressures, or if they can somehow be derived from the different physiological

shape of men's and women's vocal tracts. The the problem of vowel identification appears not to be merely a matter of scaling or normalization, but the location of a given set of formant frequencies relative to the vowel system of the language or dialect in question--and men and women may have quite different dialects.

A closer look at syllabic [ɻ̩]

Syllabic /ɻ̩/ in this dialect overlaps the /ʊ/ category almost completely in the F1xF2 vowel space. In the present data, the women's [ɻ̩] has a lower F_2 value than for the vowel /ʊ/; the men's [ɻ̩] has an F_2 is almost identical to that of /ʊ/.

While Peterson & Barney's (1952) study included syllabic /ɻ̩/ tokens of only one type ('herd'), Lehiste's (1962) landmark study of semivowels included many more tokens of syllabic /ɻ̩/ in different contexts from male speakers. Some comparison of the present data to these classic studies is called for.

Table XV presents the men's and women's averages for the formants of syllabic [ɻ̩] in the present study, along with the analogous averages from Peterson & Barney (for both male and female speakers) and Lehiste (male speakers only).

Table XV. Comparison of the present data with data from two classic studies

	F_1	F_2	F_3	source
Female speakers				
Speakers 1-5	508	1570	2041	present study, all
Speakers 1-5	537	1615	2055	present study, herd' tokens only (n=15)
P&B(f)	**500**	1640	1960	Peterson & Barney's average ('herd')
Male speakers				
Speakers A-E	435	1370	1703	present study, all
Speakers A-E	420	1404	1701	present study, 'herd' tokens only (n=15)
P&B(m)	**490**	1350	1690	Peterson & Barney's average ('herd')
Lehiste(m)	435	1253	1550	Lehiste's classes III and IV

For the female speakers, the average F_1 of syllabic [ɻ̩] in the present study is very similar to Peterson & Barney's female population. F_2 is is about 70 Hz lower in the present data, and F_3 is about 80 Hz higher. Restricting the present data to the 'herd' tokens, F_2 moves much closer to Peterson & Barney'ys average F_2, but F_3 raises, moving further from Peterson & Barney's average.

For the male speakers, the present data's F_1 average is quite close to Lehiste's average F_1 for her male speakers. The F_1 average is also lower than Peterson & Barney's, both overall and in the 'herd' context only. The present data's F_2 and F_3 averages, again both overall and in the 'herd' context only, are higher than in Peterson & Barney's male population and much higher than in Lehiste's. Interestingly, the average F_2 of the present study's 'herd' tokens is higher than that of the overall average F_2, but unlike for the women's data, this higher F2 in 'herd' is much further away from Peterson & Barney's male F_2 average.

Figure 26 is composed of bar charts representing the number of measurements of the first formant of [ɻ̩] that fell within a particular range of frequencies (50 Hz bins for F_1, 100 Hz bins for F_2 and F_3). The women's tokens are counted on the left, the men's on the right. Recall that the mean F_1 of [ɻ̩] for women was 508 Hz; for men, it was 435 Hz.

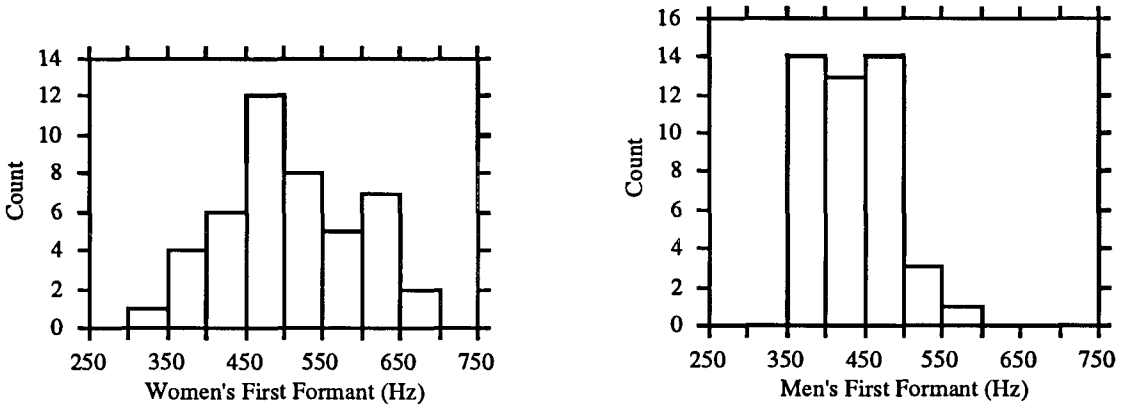


Figure 26. Number of F_1 measurements for syllabic [ɹ̩] at a given frequency.

Figure 27 is similar to Figure 26, except that it represents measurements of F_2. Recall that the mean F_2 of [ɹ̩] for women was 1570 Hz; for men, it was 1370 Hz.

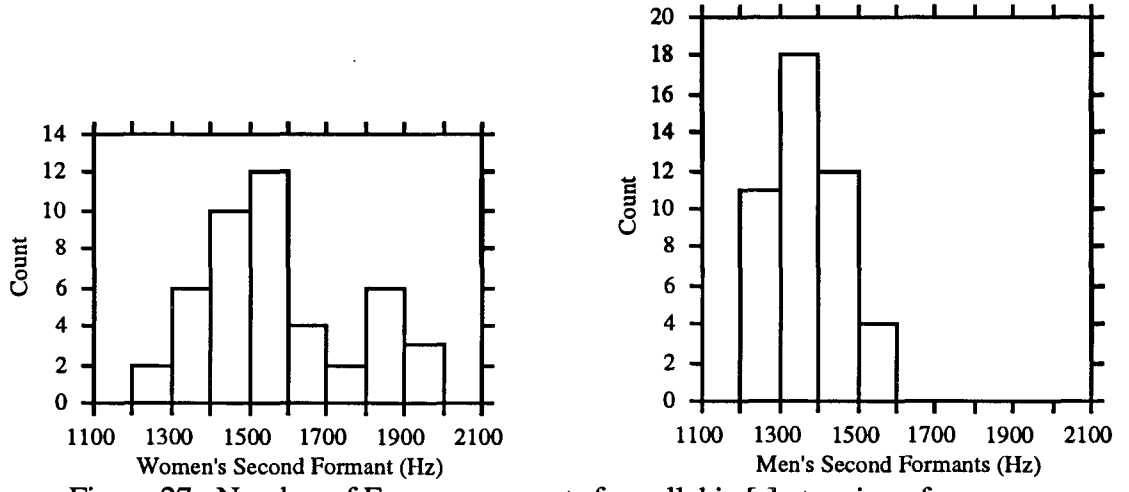


Figure 27. Number of F_2 measurements for syllabic [ɹ̩] at a given frequency.

Figure 28 represents counts of F_3 tokens at a given frequency. The mean F_3 of syllabic [ɹ̩] was 2041 for women, and 1703 for men.

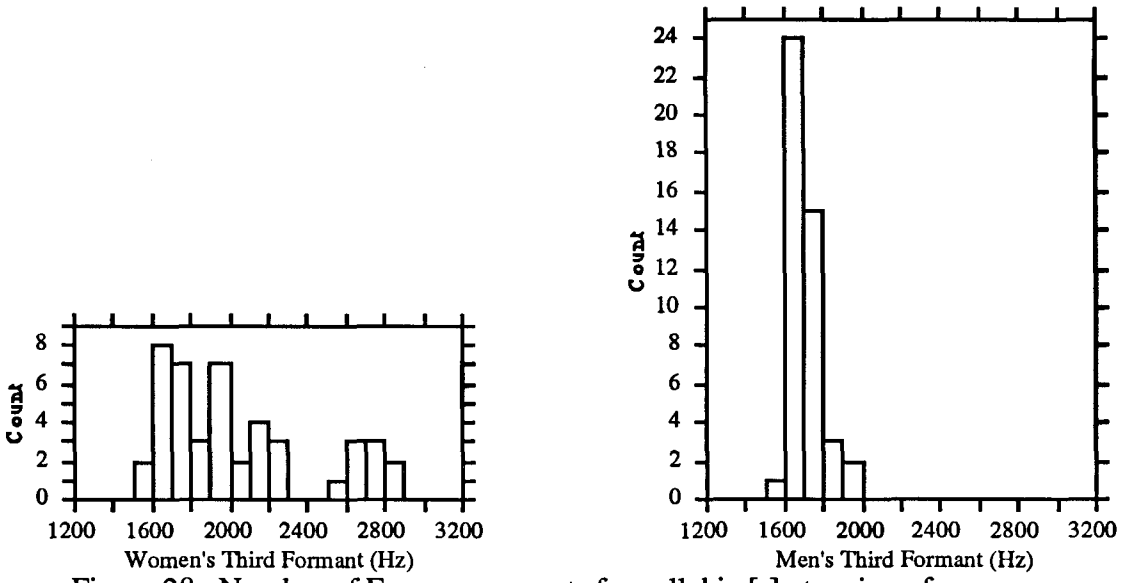


Figure 28. Number of F_3 measurements for syllabic [ɻ] at a given frequency.

Except in F3, where the extremely high-frequency measurments among the women are attributable to a particular speaker (S5), the apparently bimodal distribution of values among the women appears to be accidental.

In all three of these figures, it seems as if the men more consistently place their formants at some agreed-upon frequency, where the women have a much broader range of acceptable formant values. Put another way, if Johnson, Ladefoged and Lindau (1993) are correct in suggesting that vowel targets are auditorily controlled, it appears that the preferred target for men is a particular frequency. For F_3, that frequency is around 1700 Hz. Once again, the women in this study do not appear to prefer a particular frequency, but a range of frequencies; for F_3, between 1900 and 2100 Hz. These figures are arbitrary, representing rounded values of the mean (2040.978) plus and minus one standard error (58.864). The analogous range for men is much narrower: 1690-1710 Hz (rounded from 1702.578±10.692).

American /r/ is rarely described as having formants *at* particular frequencies. Rather, it is described as having a third formant usually *below* 2000 Hz (e.g., Boyce & Espy-Wilson, 1993). In the face of the present evidence, it is possible to refine that figure.

First, it is necessary to define 'usually'. Interpreting 'usually' as something less than 'always' (more than 95% of the time, given a standard 5% acceptable error rate), but something more than a simple majority of cases, the present data suggest that F_3 of syllabic [ɻ] 'usually' drops below 2440 Hz for women and 1775 Hz for men. These frequencies are rounded from one standard deviation above the mean, or over 80% of the time. Using the mean plus one standard deviation as a metric is, of course, an arbitrary decision, but one based on real, quantifiable observations. This can only be an improvement over a very vague figure such as 2000 Hz.

Whether this is the best definition of a cut-off frequency for American [ɻ] is, of course, open to discussion. It was simply a convenient formula, easily calculated from the present data and easy to interpret. Formulae using percentiles, standard errors, and other methods of describing the distribution of measurements might be equally useful. The purpose of presenting these figures is not to suggest that they are the result of the best possible method of describing the 'critical' frequency, but merely to replace the vague 2000 Hz figure with something more empirically defensible. Further, refining this 'critical' frequency in this way provides a way of describing how much higher women's formant can be, compared to men's.

5. Conclusion

This study has yielded evidence that significant differences exist between the vowel spaces of women and men. Women's formant frequences are generally higher than men's, as predicted

by their relatively smaller vocal tracts, but the women also showed greater within category variance, as well as different distribution of categories within their vowel space. Thus, women and men appear to have vowel spaces which differ arbitrarily as a function of gender-specific dialect. Both these dialects are similar, however, in their centralization of back vowels compared to other reports of English. This centralization results in almost complete overlap between the /ɻ̩/ and /ʊ/ categories in the F_1 and F_2 dimensions, though [ɻ̩] can be distinguished from [ʊ] and the other vowels generally by its relatively lower third formant.

The greater variance associated with women's vowel productions suggests that the notion of an acoustic phonetic target for a vowel is not a particular set of frequencies, but a set of frequencies within some range, particularly for women. For men, these ranges are quite small, giving the impression of targets at specific frequencies.

Syllabic [ɻ̩] is a particularly good example of this, in that the men in this study appear to place the formants of [ɻ̩], particularly F_3, at specific frequencies, in a way that women do not. F_3 of /ɻ̩/ for men usually drops below 1775 Hz. For women, F_3 usually drops below 2440 Hz. However, while men tend to place the third formant of [ɻ̩] in the range of 1600 to 1800 Hz, the women in this study do not appear to overwhelmingly prefer any specific frequency within a much broader range.

In short, sex-specific differences were noted in the domain of vowel formant frequencies, and these were mostly compatible with a model which accounts for such variation as a function of vocal tract size. However, this model alone cannot account for all the differences noted; even in the domain of formant frequency, some sex differences are the result of learned speech behaviors.

References

Boyce, Suzanne & Carol Espy-Wilson (1993). "Coarticulatory stability in American English /r/". In *Speech Communications Group Working Papers* **9**:80-94, Research Laboratory of Electronics, MIT.

Disner, Sandra F. (1983). *Vowel quality: the relation between universal and language-specific factors.* Ph.D. dissertation, UCLA. Also *UCLA Working Papers in Phonetics*, 58.

Fant, Gunnar (1973). *Speech sounds and features.* Cambridge, MA: The MIT Press.

Hagiwara, Robert (1994). "Speaker sex and formant frequencies of American [ɹ]". 127th Meeting of the Acoustical Society of America, 7 June 1994, Cambridge, MA.

Hagiwara, Robert (in preparation). *Acoustic realizations of American /r/ as produced by women and men.* Ph.D. dissertation, UCLA.

Johnson, Keith, Peter Ladefoged & Mona Lindau (1993). "Individual differences in vowel production". *Journal of the Acoustical Society of America*, **94**(2):701-714.

Lehiste, Ilse (1962). "Acoustical characterisitcs of selected english consonants". *Communication Sciences Laboratory Report*, **9**.

Peterson, Gordon E. & Harold L. Barney (1952). "Control methods used in a study of the vowels". *Journal of the Acoustical Society of America*, **24**(2), 175-184.

A Manual for Phonetic Transcription: Segmentation and Labeling of Words in Spontaneous Speech

Pat Keating, Peggy MacEachern, Aaron Shryock, Sylvia Dominguez

1. Introduction

This manual describes a set of conventions for the segmental phonetic transcription of the kind of speech that is found in spontaneous conversation. These conventions were developed for a project in which particular target words were segmented and labeled from the conversational utterances in the Switchboard corpus. The Switchboard corpus, collected by Texas Instruments and now available from the Linguistic Data Consortium at the University of Pennsylvania, comprises 3 million words of telephone conversations by 550 speakers. It has been completely transcribed in standard orthography by TI, but has not been phonetically transcribed. In our project we marked begin and end timepoints of selected words and provided a segmental phonetic transcription for each. We did not time-align the individual segments within the words, but the guidelines contained here should be sufficient for that task as well. However, due to the nature of this project, the transcription conventions we developed have little to say about the treatment of pauses, hesitations, or disfluencies.

The starting point for the transcription system we developed was the one associated with the TIMIT corpus of read speech. The symbol set presented here will be called the UCLABET. It is an extension of the TIMITBET, the symbol set used by transcribers at MIT for the TIMIT corpus, which is itself an extension of the ARPABET. The UCLABET adds to the TIMITBET new symbols that allow a narrower transcription, in particular, additional places of articulation for stop and fricative consonants, and three diacritics. Throughout this manual we note the correspondence between our use of the UCLABET, and what is seen in the TIMIT database transcriptions.

In terms of the phonetic categories added to the symbol set, we follow the IPA system of Place by Manner distinctions for consonants, along with various diacritics. However, some arbitrary mapping between IPA and ASCII symbols is required for maximally useful machine-based transcriptions. We chose a TIMIT style of symbol set over two other, more extensive, symbols sets, namely PhonASCII (Allen 1988) and Worldbet (Hieronymous ms., n.d.) because of what we judged to be easier learning and keyboarding of the individual symbols. Our experience is that it is preferable for symbols to consist largely of lower-case letters, not numbers or non-alphabet characters, and for symbols to be formed of these characters in systematic ways. We believe that the ease of use of such a symbol set offsets the price it exacts, namely, symbols of variable length that must be separated by spaces. In any event, our symbol set could be automatically translated into one of these others.

A symbol set is only the starting point for phonetic transcription. A set of conventions for the use of those symbols is also necessary, most notably because the transcription imposes the two idealizations of **discrete segments** in a **linear order** on the speech signal. It is not always clear, especially in fluent speech, how the signal is to be segmented into phoneme-sized units, no matter how narrowly defined they may be. Other transcription difficulties arise from the limited number of symbols (in any finite symbol set), and from differences in the phonemic systems of the various speakers and transcribers. Therefore any phonetic transcription project must incorporate guidelines for transcribers. In general, the literature is of little practical help in this regard. For example, the TIMIT database is accompanied by only sketchy explanations of transcription practice

(Seneff and Zue 1988, Garofolo et al. 1993). PhonASCII (Allen 1988), being a symbol set rather than a transcription method, offers no explicit conventions.

Most prior transcription projects have apparently not yielded publicly-available materials. Even textbooks are notably vague on the actual practice of anything but the broadest transcription, and in any event they rely entirely on the listening ability of the transcriber. One naturally looks for conventions, like the TIMIT ones, that refer to the acoustic signal as well as to the trained listener's percept. We found most useful the conventions for computer-corrected transcription by Henton and Bladon (1987), and a draft of the OGI conventions for transcription of telephone speech by the Spoken Language group at the Oregon Graduate Institute (Metzler and Nathman 1993), kindly made available by Ron Cole. It is for this reason that we consider it worthwhile to circulate our own conventions: that there be something in the public domain that at least aims at completeness and which can serve as a starting point for debate and development.

Our transcription was carried out using the Labeler facility in Entropics' *xwaves*. The screen display consisted of a time-aligned waveform, wideband spectrogram and orthographic transcription for the portion of the signal containing the word to be transcribed. Examples of this display will be seen throughout this manual (except as otherwise noted). The phonetic transcription was typed into an additional window, in which prior transcriptions of the same token (e.g. by other transcribers) could also be displayed.

2. Consonants

2.1. Presence vs. absence of a consonant. Though we did not perform word-internal temporal segmentation, we tried not to transcribe any segments which had no plausible segmentation in the target word. Sometimes a consonant is heard but (especially with flaps or other weak sonorants) no clear interval that could be uniquely associated with the consonant is seen in the signal. If there is a brief local amplitude drop in the signal which could be assigned to the consonant, the consonant is transcribed. If there is no amplitude drop in the signal and/or no interval in which that consonant seems to predominate, then no consonant is transcribed. An example of this latter kind is that in the word "mavericks", it seems that the /v/ and /r/ often overlap and it can be impossible to distinguish any piece of the signal that might be attributed to the /v/ alone. It would be possible to devise a form of segmental transcription in which nominally-adjacent segments are allowed to overlap, but we have not done so.

Occasionally a fricative, e.g. [s], is clearly heard, but only a gap is seen in the acoustic displays. In these cases the fricative is transcribed.

2.2. Stop closure symbols
2.2.1. TIMIT *pcl,tcl,kcl,bcl,dcl,gcl*. We follow the TIMIT practice of dividing stop consonants into separately-transcribed acoustic closure and release portions. The closure symbols are formed from the basic stop character followed by "cl". For example, the full transcription for a labial stop would be *pcl p:* a labial closure followed by a labial release. A stop is expected to have a closure interval unless it is the second stop in a cluster (section 2.2.3.4 below). The onset of stop closure is associated with a sudden sharp drop in amplitude and loss of energy in formants above F1. Voicing offset is not a criterion for onset of closure.

A closure may contain some noise and still be transcribed as a stop: most signals are somewhat noisy, so only noise beyond the background level can be taken as a clear indication of loss of stop closure. As a result, some instances of weak frication in a stop

consonant will be ignored. Figure 1 ("limited") shows a token with two instances of *dcl* which are clearly defined even though they do contain some weak energy. Figure 2 ("but") shows a token in which stop closures *bcl* and *tcl* are transcribed against a high level of background noise. However, if frication noise or sonorant energy during a phonemic stop constriction is strong enough that a true fricative or an approximant is heard and seen, then it is generally transcribed as a fricative. The symbols available for spirantized or sonorized stops are *ph, bh* for the bilabials, *tfr, dfr* (these are non-sibilant alveolar fricatives) or *dx* (flap) for the alveolars and *x, gh* for the velars. For more discussion of flaps, see below. Figure 3 ("baby-sitter") shows a pronuciation typical for this word, in which the first /b/ has a closure interval and release (*bcl b*) but the second /b/ is a voiced sonorant *bh*. One other context in which stops appear as fricatives is next to another fricative. In this case the two fricative segments must be distinct in order to be labeled separately. Figure 4 ("chips") shows a /ps/ sequence in which the /p/ has no stop closure and yet is clearly distinct from the /s/; it is transcribed as *ph*. Figure 5 (also "chips") shows the initial affricate phoneme with a spirantized closure, distinct from the fricative part of the affricate. If this interval were a stop, it would be transcribed as *tcl*, but because it is not a stop yet does not sound sibilant (that is, does not sound like [s]) it is transcribed as *tfr*. If a stop is spirantized or sonorized for some but not all of its constriction, then it is transcribed as a stop. For example, in a /ks/ sequence, the /k/ often exhibits some noise during part of its closure, but it is not always fricated enough to be heard as *x*, so it would be transcribed *kcl* despite the noise. At the same time, such weak frication may have as a consequence that the stop has no burst, so the whole sequence would be *kcl s*, and therefore distinct from *kcl k s*.

In a few instances a speaker had a clearly nasal closure for /b/, i.e. a prenasalized stop. This was not due to the context but appeared to be a consistent characteristic of the speaker. We transcribe this as *m b*, without *bcl*. Figure 6 ("bear") shows one such instance.

These stop closure symbols are also used for phonemic fricatives which are produced with a clear stop closure interval. In practice this means *tcl, dcl* as the other fricatives are covered by the extra stop symbols given immediately below. A stopped fricative will usually have a release, which can be transcribed with the fricative symbol (e.g. *tcl s*). The same transcription would be used for a more extended frication interval after a stop closure without a clear burst. Note that stop symbols are not used for short bits of silence adjacent to fricatives. The silent interval must actually sound like a stop; otherwise, the silence is not noted. If there is doubt about whether a fricative is stopped, the fricative is transcribed.

This discussion applies to a so-called epenthetic stop adjacent to a fricative. If it sounds like a stop followed by a fricative, it is transcribed as such, with or without a burst depending on the signal. Figure 7 ("since", displayed using CSL, not xwaves) shows an epenthetic stop with a closure *tcl* and release burst *t*. Sometimes there may be an ambiguity between an epenthetic stop and a single pitch period of glottalization at the end of the previous segment. In the case of glottalization, the frequency band of the pulse should match that of the previous segment. The frequency band of a burst for an epenthetic stop, however, will not match the frequency of the previous segment. Figure 8 shows another token of "since" which clearly contains glottalization (and nasalization) of the vowel, not an epenthetic stop.

Figure 1. Examples of stop closures.

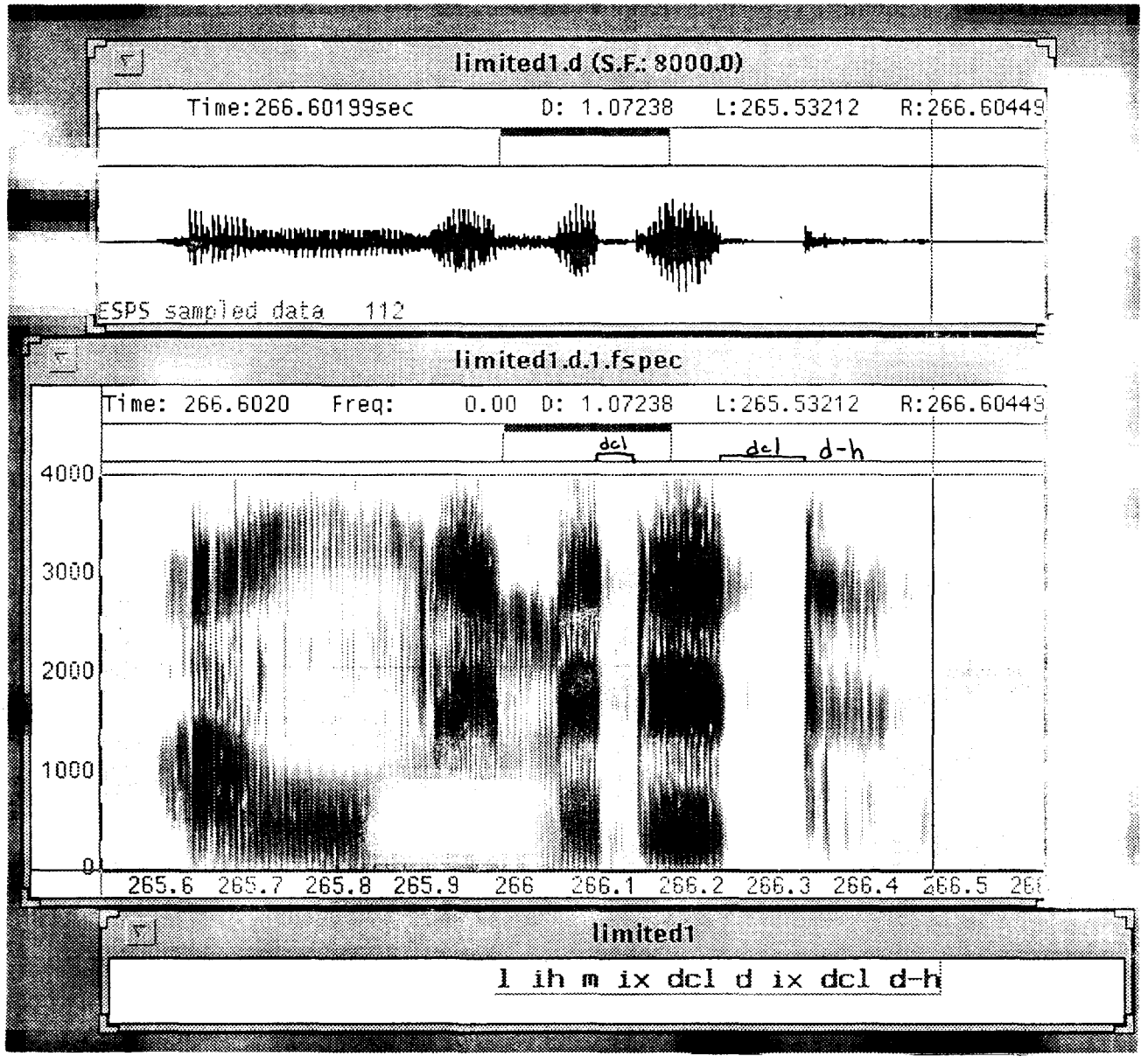


Figure 2. Examples of stop closures.

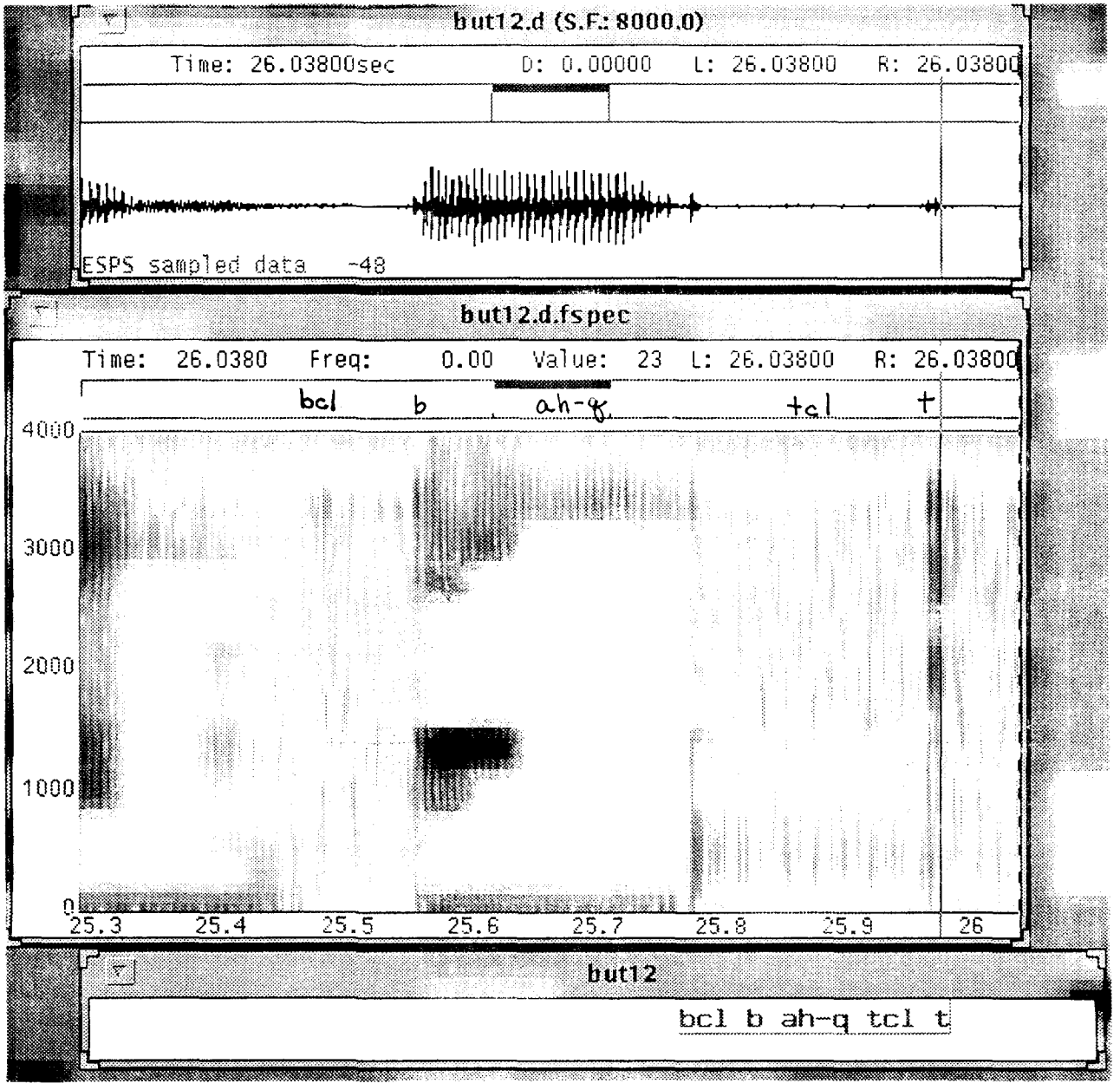


Figure 3. Example of *bh*.

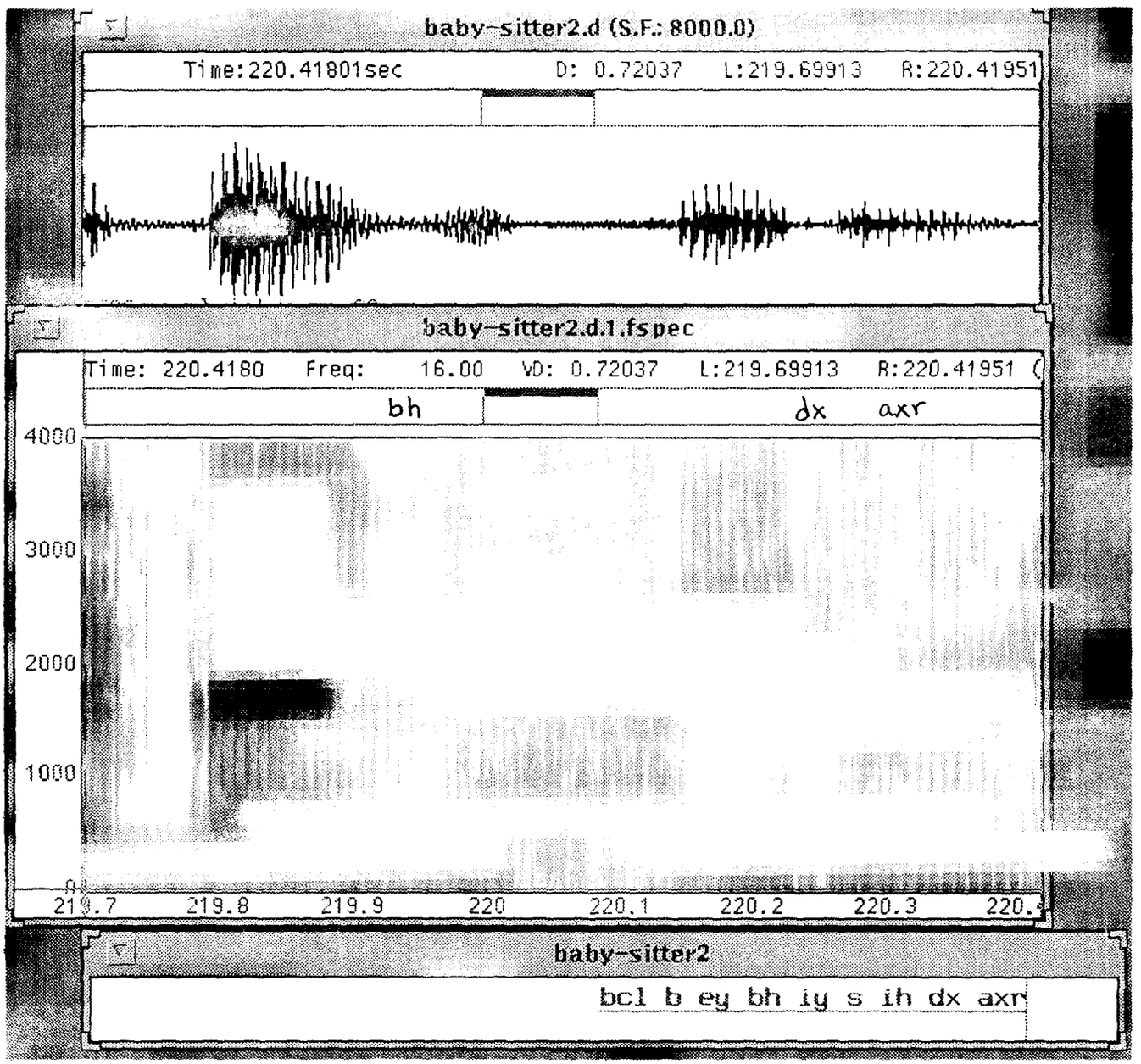


Figure 4. Example of *ph*.

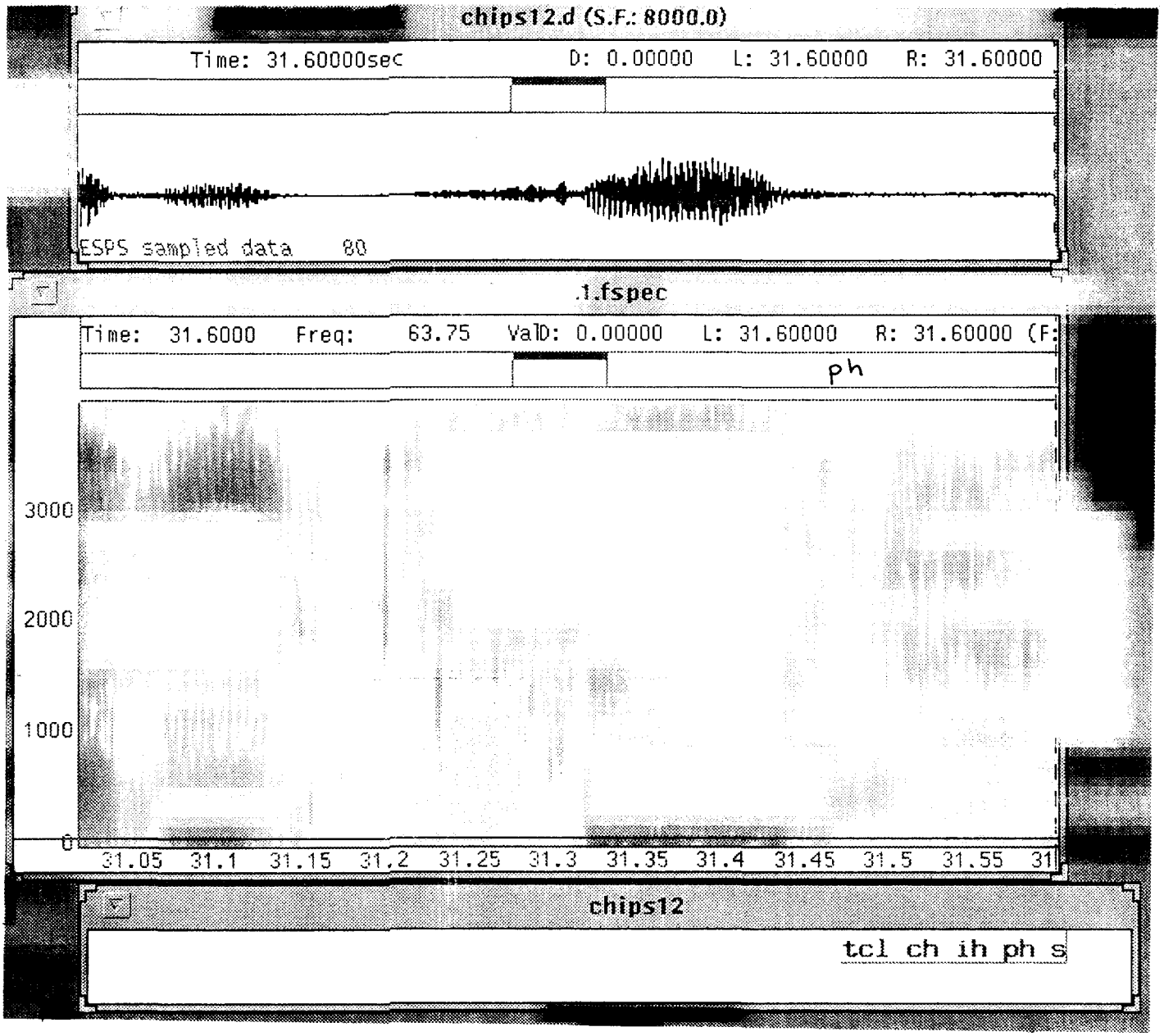


Figure 5. Example of *tfr*.

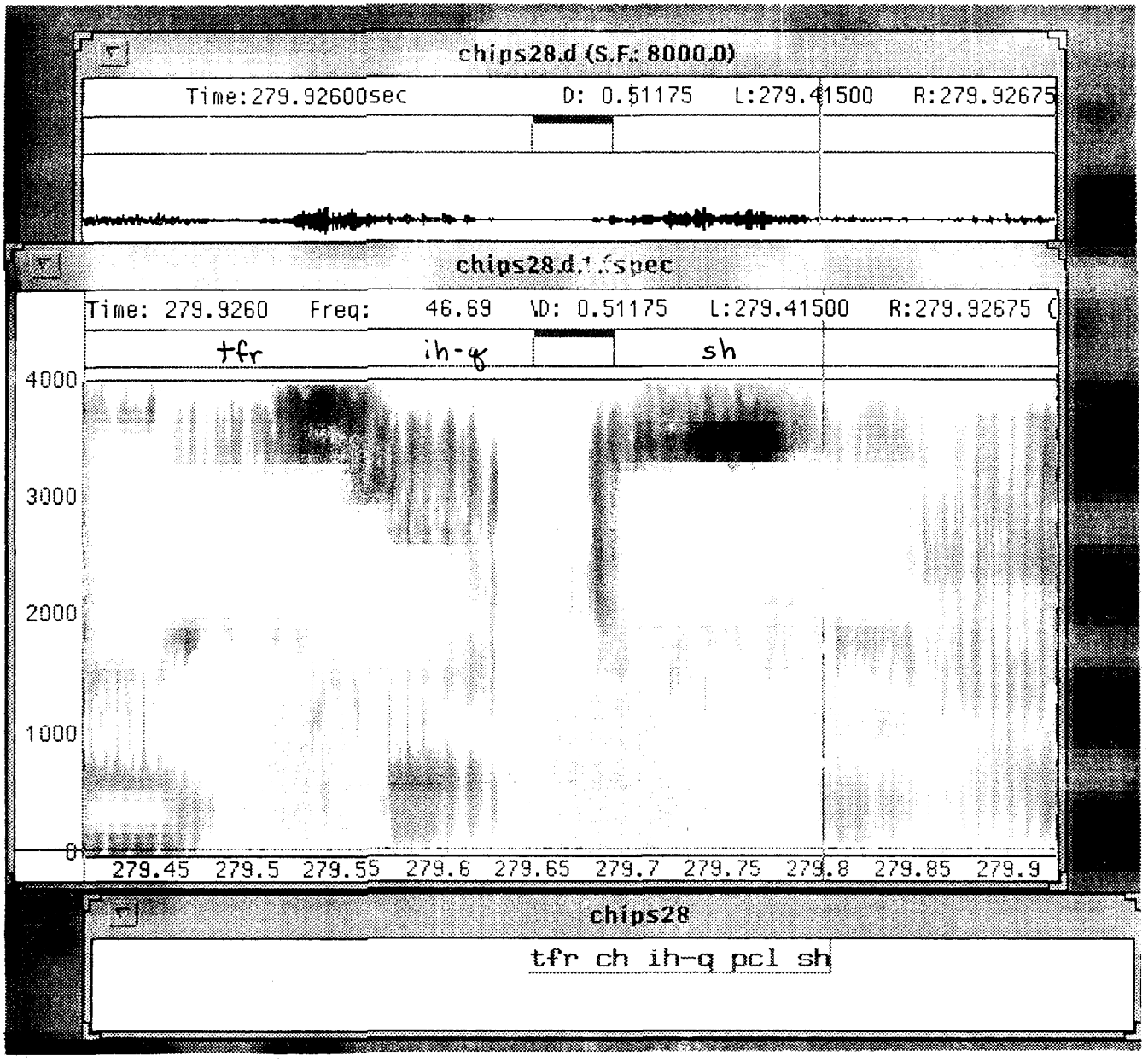


Figure 6. Example of prenasalized stop.

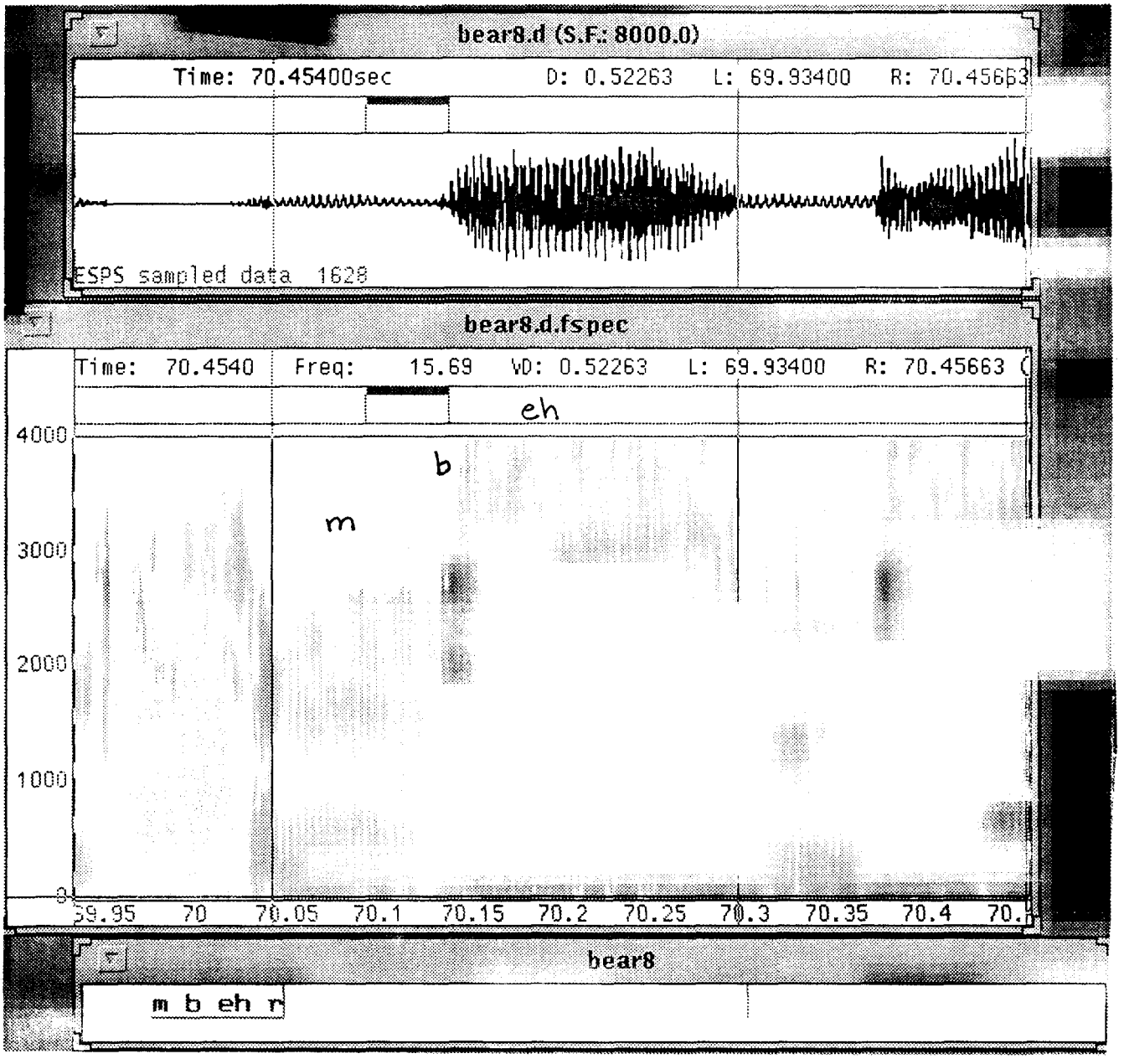


Figure 7. Example of epenthetic stop.

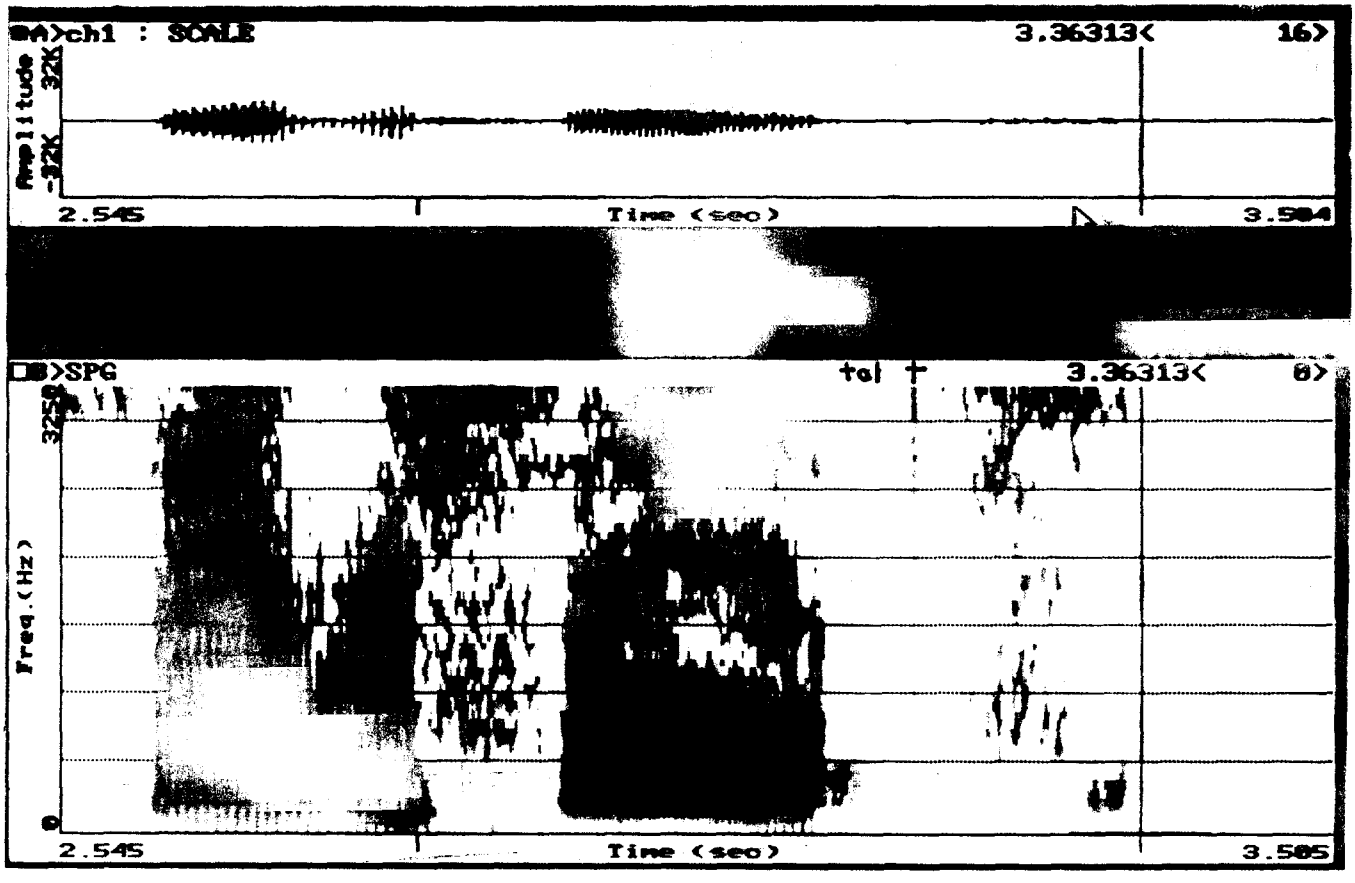
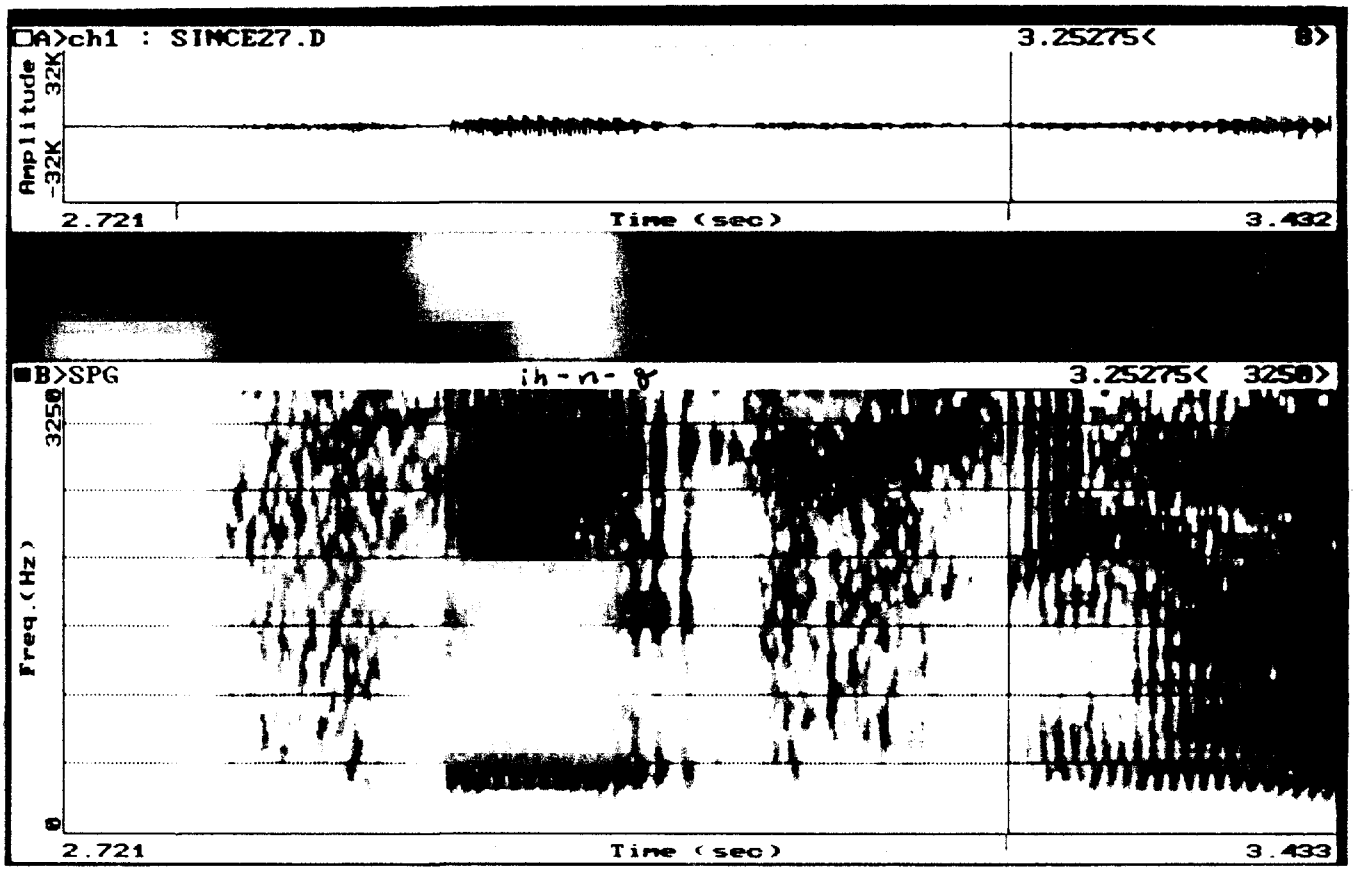


Figure 8. Example of glottalization.



2.2.2. Additional stop closure symbols *fcl, vcl, thcl, dhcl.* Sometimes labiodental and dental fricative phonemes are pronounced as stops or affricates, with stop closures. TIMITBET provides no symbols for stops at these places of articulation, so we have added symbols for the closure intervals. The releases of these stop closures are transcribed with the corresponding fricative symbols *f, v, th, dh.*. For example, a phonemic /v/ pronounced as either a stop or an affricate would be transcribed *vcl v.* No distinction is made between stops and affricates in this transcription.

It should be noted that these dental stop symbols have been used only for stopped variants of the dental fricatives. They have not been used to transcribe dental variants of phonemic alveolar stops. We would expect contextual assimilations in phrases such as "about the", but the /t/ is then in a cluster and so is typically treated as unreleased. Judging whether the closure itself is assimilated in place is too difficult a decision, and so such cases are transcribed simply as *tcl dh.*

TIMIT correspondences: None of these symbols are in the TIMITBET. It seems that *fcl* and *vcl* should be mapped into TIMIT *f,v.* thcl corresponds variably to *th, tcl,* and *epi;* presumably *dhcl* similarly. The apparent inconsistency of TIMIT practice in dealing with stopped fricatives is one reason we have added these symbols.

2.2.3. Other aspects of transcribing stop closures.
2.2.3.1. Adjacent to pause: Voiceless closure adjacent to a silent pause cannot be accurately segmented because either its beginning or its ending point is not known. With TIMIT no principled basis is given for transcribing such a voiceless closure. As a result, in TIMIT voiceless stops after pause are consistently transcribed without any closure, in the way that second stops in stops clusters are transcribed without closures. On the other hand, in TIMIT voiceless stops before pause are treated differently. If their closure were not transcribed, then they would not be recorded in the transcription at all, since closure and release are the only stop elements transcribed -- formant transitions, no matter how audible, are not recorded. In TIMIT, such closure intervals are transcribed so that the stops are recorded, but their segmented duration in the signal seems to be random. Our approach to this difficulty is the following. We consistently transcribe closures for stops adjacent to pause, giving them the arbitrary duration of 200 msec. This number was chosen to be longer than any observed delimited closure, so that these closures would stand out in any distribution of measured closure durations in a corpus. (This criterion is meant to apply generally. In our actual transcription project, no durations were assigned to individual segments in words. The 200-msec closures do however determine the location of word boundaries.) Note that this provision applies only to silent pauses. If a breath is heard as part of a pause, the closure will not extend (forward or backward) into the breath, and the closure will therefore have a duration less than 200 msec.

2.2.3.2. Voicing decision: Whether there is acoustic voicing during the closure is not the most important thing in determining the closure voicing label. A voiceless phoneme can have some voicing during closure and a voiced phoneme can have none. We determine voicing by listening to the segment before the stop, including the transitions, or to the stop release and following segment if there is no preceding segment. In ambiguous cases, the phonemic voicing of the segment is preferred in labeling.

If both a closure and release are present for a **single** phonemic stop segment, they must agree in transcribed voicing. For example, with a final stop, if the closure sounds voiced but the release is clearly voiceless, then this is a case of partial devoicing, which we do not transcribe (just as when a fricative is partially devoiced). Therefore in this case the whole stop is transcribed as voiced. Only if the whole stop, in its context, sounds voiceless would it be transcribed as such.

2.2.3.3. Place decision: Place of articulation is determined by the formant transitions into the stop. Assimilation of place is transcribed if it is both seen and heard. As with voicing, a closure and its release must agree in their transcribed place.

2.2.3.4. Clusters: If the first of two stops is unreleased and the closure interval is otherwise indivisible, then only one closure and one release are transcribed. Generally the first phonemic stop's place and voicing are used for the closure, and the second phonemic stop's place and voicing are used for the release. This convention is followed both within words and across word boundaries. Across word boundaries this convention can be confusing, as part of the stop is not recorded within a given word. For a word-initial stop, there is in fact little ambiguity. If the stop has no closure, then it must be part of a cluster with a stop in the preceding word (except as described in the next section). That is because, as described in 2.2.3.1, even a closure in a silent pause is transcribed, and as described in 2.2.1, a spirantized stop is transcribed as a fricative. On the other hand, a word-final stop with no release may or may not be followed by another stop.

2.2.3.5. After nasals: The oral closure of a stop may or may not be present after a nasal, i.e. it is possible to have "*n d*" instead of "*n dcl d*" -- no non-nasal closure visible before the release. The acoustic convention here is that a complete vertical band of "white space" above the baseline is enough to count as a closure, or a clear sharp drop in amplitude after the nasal before the release -- so usually some closure will be seen. Figure 9 ("conditions") is an example in which the nasalization was judged to extend up to the stop release. Similarly, prenasalized voiced stops as in Figure 6.

TIMIT correspondences: To the extent determinable from the TIMIT documentation, use of these closure symbols follows TIMIT except in three respects: (1) we standardize silent closures, not delimited by a breath, to 200 msec, and use these postpausally as well as prepausally; (2) because we transcribe closures for stopped fricatives, it is possible that some instances of *tcl, dcl* used by us for stopping of alveolar fricatives would not appear in TIMIT transcriptions; (3) because we do not use *epi* we use stop closure symbols where TIMIT would sometimes use *epi* .

2.3. Stop release symbols.
2.3.1. Oral stop releases.
2.3.1.1. TIMIT *p,t,k,b,d,g*. Transcription of a stop release is associated with an acoustic burst, a sudden sharp increase in amplitude. The beginning of a release is the left edge of this burst (on the waveform) and the end of a release into a sonorant is the Voice Onset, as seen at higher formants. Before a voiceless segment or silence, the end of release is the end of the full-frequency range of noise seen in the release. Sometimes there is no clear release. The default transcription is none, that is, positive evidence of a release is required for one to be transcribed.

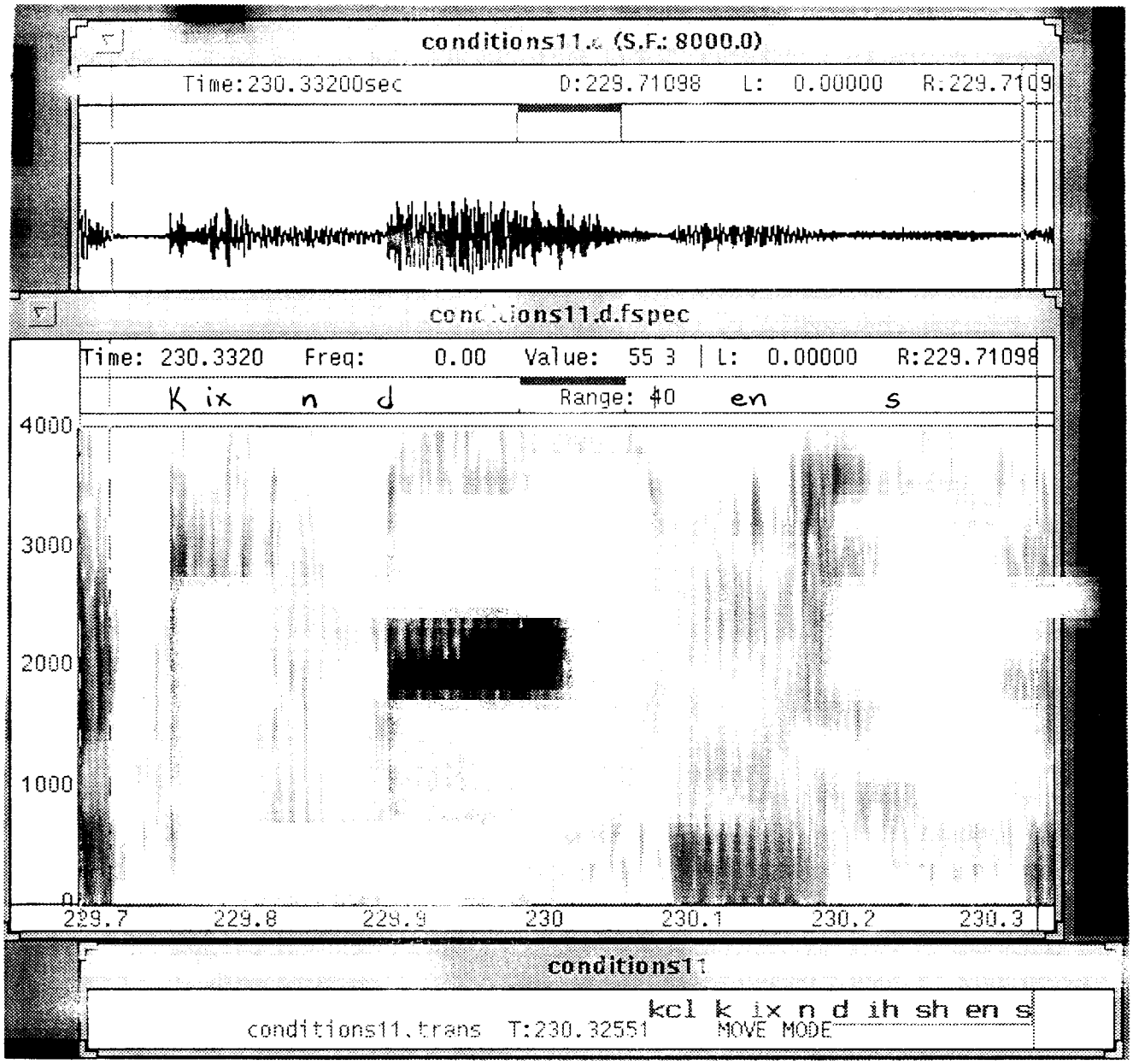
For affrication of stop releases, see section 2.8 on affricates below.

TIMIT correspondences : Seems the same.

2.3.1.2. Aspiration. Transcription of stop consonant aspiration (-h) is described in 4.2. Of relevance here is that a final stop (voiced or voiceless) with a long voiceless release is coded as aspirated, e.g. *p-h* or *b-h*. Figure 1 ("limited") shows an example of a final /d/ transcribed in this way.

2.3.1.3. Other stop releases. As described above in section 2.2.2, the fricative symbols *f,v,th,dh* are also used to transcribe releases of labio-dental and dental stops.

Figure 9. Example of nasal plus oral stop sequence.



2.3.2. Glottal stop. Our treatment of glottal stop and laryngealization differs from that in TIMIT. Parallel to our treatment of other stops, "glottal stop" includes glottal stop closure *qcl* and glottal stop release *q*. We try to distinguish such a full glottal stop (a more pronounced closure/release of pressure) from mere laryngealization (creaky voicing) of a voiced segment. With laryngealization, discussed in section 4.3 below, the formant structure of the oral articulation is preserved. In practice, full glottal stop is expected to be less common than laryngealization. Glottal stop can be found (1) phrase-initially (glottal attack); (2) when substituted for an oral voiceless consonant (e.g. /t/), as verified by sound and by lack of formant transitions. Figure 10 shows a token with "about" followed by "the", in which neither a /t/ nor a /ð/ is heard. There is simply a glottal closure without audible release (against a background of noise).

For consistency with the treatment of oral stops, *qcl* adjacent to a pause is given a duration of 200 msec.

TIMIT correspondences: Our *q* probably should be merged with our *qcl* for TIMIT correspondence. These together correspond to (one use of) TIMIT *q*. *qcl* with 200 msec duration has no TIMIT correspondence (was not segmented in TIMIT).

2.4. Flaps (taps)
2.4.1. Oral flap *dx*. A flap is prototypically a very short voiced closure with no release burst. However, flaps can be of medium duration, and they can also have bursts, but they should not sound like stops. The closure of flaps may contain frication; in practice, flap is transcribed where a fricative would be more accurate, but where the brief duration of the sound leads the transcriber to hear a flap. Figure 3 ("baby-sitter") shows an instance of such a flap. Flaps are not determined by duration per se, but by sound, except for the following guideline: if the closure is less than 20 msec, then we use the flap symbol even if there is a burst (except after a nasal, where a very short *dcl* could be found), but if there is no burst, then decide by sound. Figure 1 ("limited") contains a phonemic /d/ in a flapping context transcribed as a stop, not a flap.

TIMIT correspondences: Seems the same.

2.4.2. Nasal flap *nx*. These are hard to distinguish from short /n/. One criterion in listening is to segment out the nasal plus a following vowel: if the nasal sounds like a plausible syllable onset for the following vowel, then /n/ is preferred.

TIMIT correspondences: Seems the same.

2.5. Nasals *m,n,ng*. Place assimilation of nasals is transcribed. Figure 11 ("explain") shows a final /n/ transcribed as [m]; it occurs before the word "more" (the transcribed [m] thus belongs to both words). Sometimes there is no interval in the signal that corresponds to a nasal stop: there is a nasalized vowel followed by an oral stop. The TIMIT convention is to arbitrarily mark the last one or two pitch periods of the vowel as the nasal consonant in these cases; instead we use a nasal diacritic (see below) followed by the oral stop.

TIMIT correspondences: Seems the same, except for TIMIT's use of *n* as a nasal diacritic as noted above. These very short *n* 's which result from the diacritic use of /n/ in TIMIT do not appear in our transcriptions.

Figure 10. Example of *qcl*.

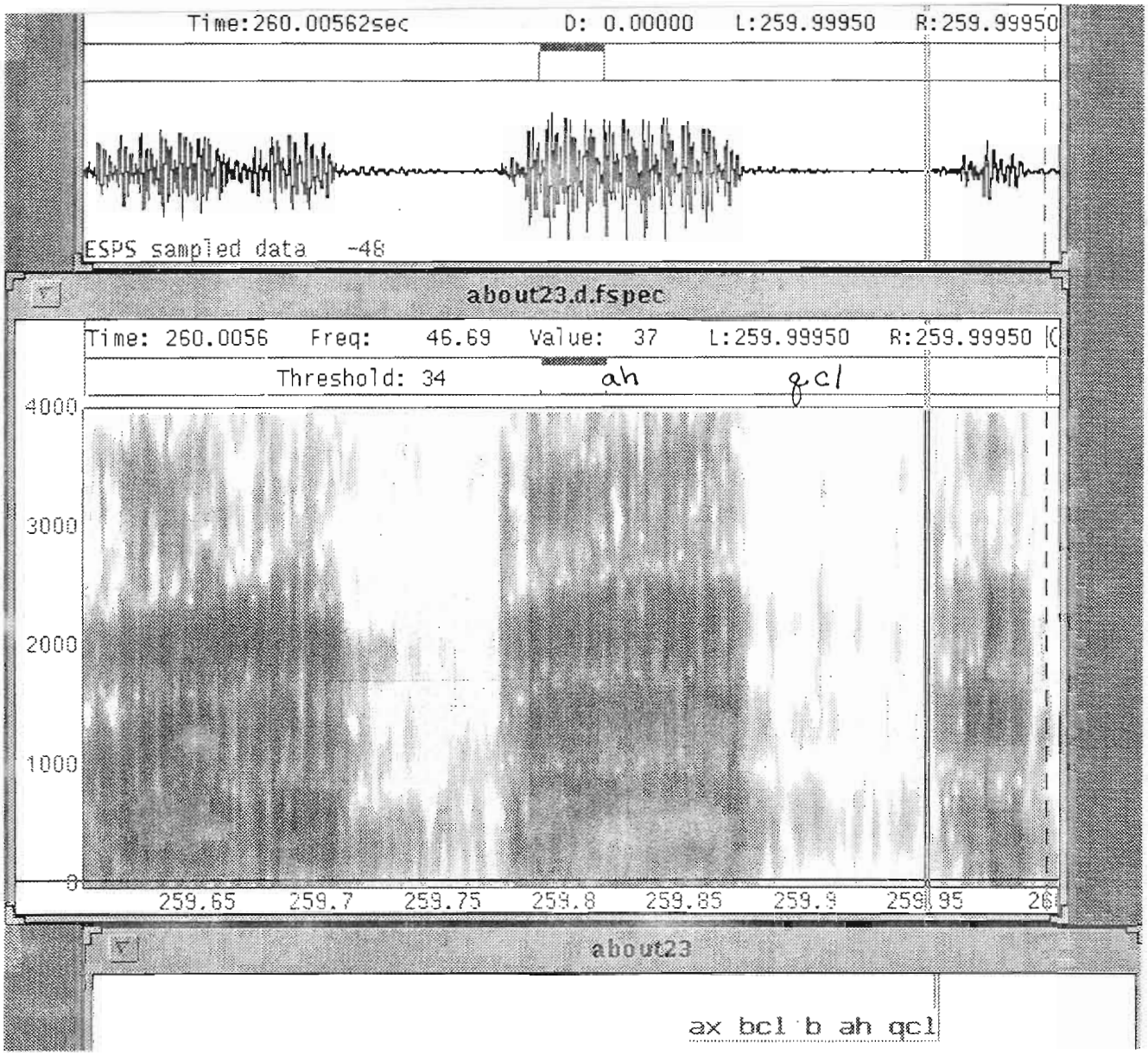
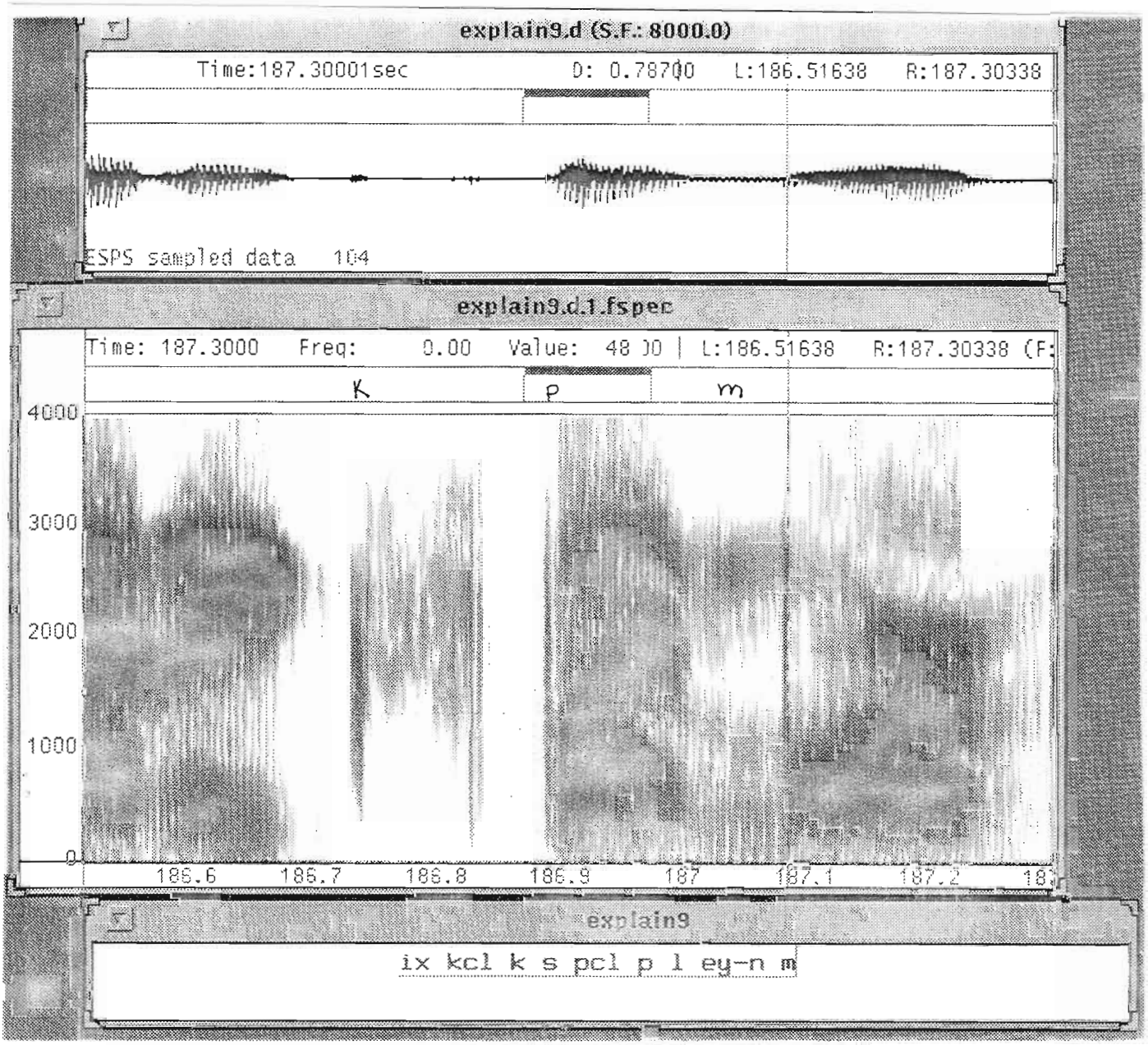


Figure 11. Example of place assimilation.



2.6. Syllabic sonorants *em,en,eng,el,er,axr*. Syllabic sonorants are used when /m n ŋ l r/ appear as syllabic nuclei, without a supporting /ə/ or other vowel that could be segmented out. All of these except *er* are expected only when stressless. As the default, transcription of a vowel is preferred, except for *er* and *axr* where the syllabic consonant is the default. Figure 3 ("baby-sitter") contains an example of syllabic stressless *axr*. Figure 9 ("conditions") contains an example of stressless syllabic *en*; compare the first syllable, with a vowel-nasal sequence, to the last syllable, with a syllabic nasal. Figure 12 ("well") shows a syllabic /l/ *el* in a monosyllabic function word.

We differ from TIMIT in our treatment of coda /r/. Unlike TIMIT, we use *r* for codas as well as onsets and reserve *axr* for syllabic-r. As a result, we do sometimes have to make difficult decisions about syllabicity when /r/ follows a high vowel or glide. Since we have to make difficult decisions about syllabic vs. onset /r/ in any case, it seems preferable to make these decisions more generally and keep *axr* consistent with the other syllabic sonorants.

Such a decision about syllabicity is needed for /r/ in words like "mavericks" or "mystery". To justify the syllabic variant, we look for amplitude dips before and after the sonorant which set it off as a nucleus, and look for the sonorant to have locally high rather than low amplitude. For example, with /r/, if the part of the signal having the lowest F3 value is very low in amplitude, this suggests the /r/ is an onset consonant, whereas if that part has a higher amplitude, this suggests it is a nucleus. In listening to the word, we segment the portion in question and try to judge that part alone as being one or two syllables. We must stress that these are very difficult judgments, however. Figures 13 and 14 show two tokens of "mystery", the first transcribed as two syllables (onset *r)*, and the second transcribed as three syllables (syllabic *axr*). The onset-*r* shows low amplitude throughout the low-F3 region, while the syllabic-*axr* shows a higher energy portion before the minimum, especially in F1.

No additional onset consonant is transcribed after a syllabic consonant, e.g. there is no *r* after *axr* in "mystery". This convention is from Henton & Bladon (1987). However, it would be possible and perhaps desirable to develop the criteria for when a syllabic consonant extends into onset position, so that each token could be transcribed accordingly.

TIMIT correspondences: Mostly the same as their description (but we hope we are more consistent in following it), except that we use *r* for nonsyllabic postvocalic /r/ where TIMIT uses *axr* .

2.7. Fricatives *ph,bh,f,v,th,dh,tfr,dfr,s,z,sh,zh,x,gh*. Fricative symbols are used for phonemic fricatives produced as such, and also for phonemic stops produced as fricatives or the corresponding approximants. They are also used for the releases of affricates.

Fricative symbols *ph,bh,tfr,dfr,x,gh* are used for spirantized stops, as described in section 2.2.1 above. Use of stop closure symbols for stopped fricatives is also described in section 2.2.2 above.

Devoicing of fricatives is transcribed only when it clearly dominates the percept. It must be judged in the context of the preceding vowel (e.g. something that in isolation sounds like /s/ might still sound like /z/ after a long vowel). A small amount of voicing at the onset of the fricative, or a relatively short fricative after a long vowel, will generally make a fricative sound voiced. Figure 9 ("conditions (that)") shows a devoiced /z/; note its s-like duration.

Figure 12. Example of *el*.

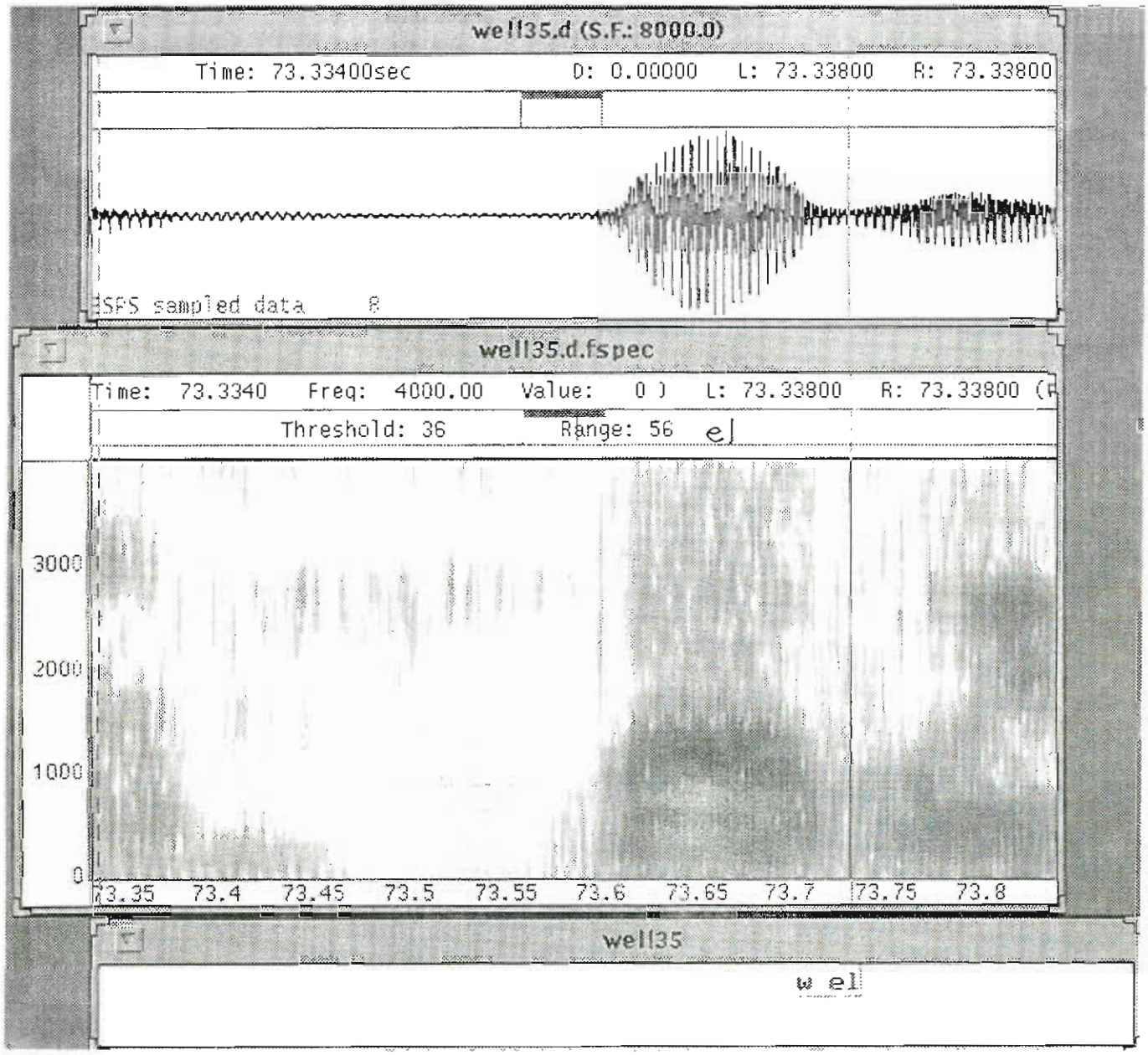


Figure 13. Example of onset *r*.

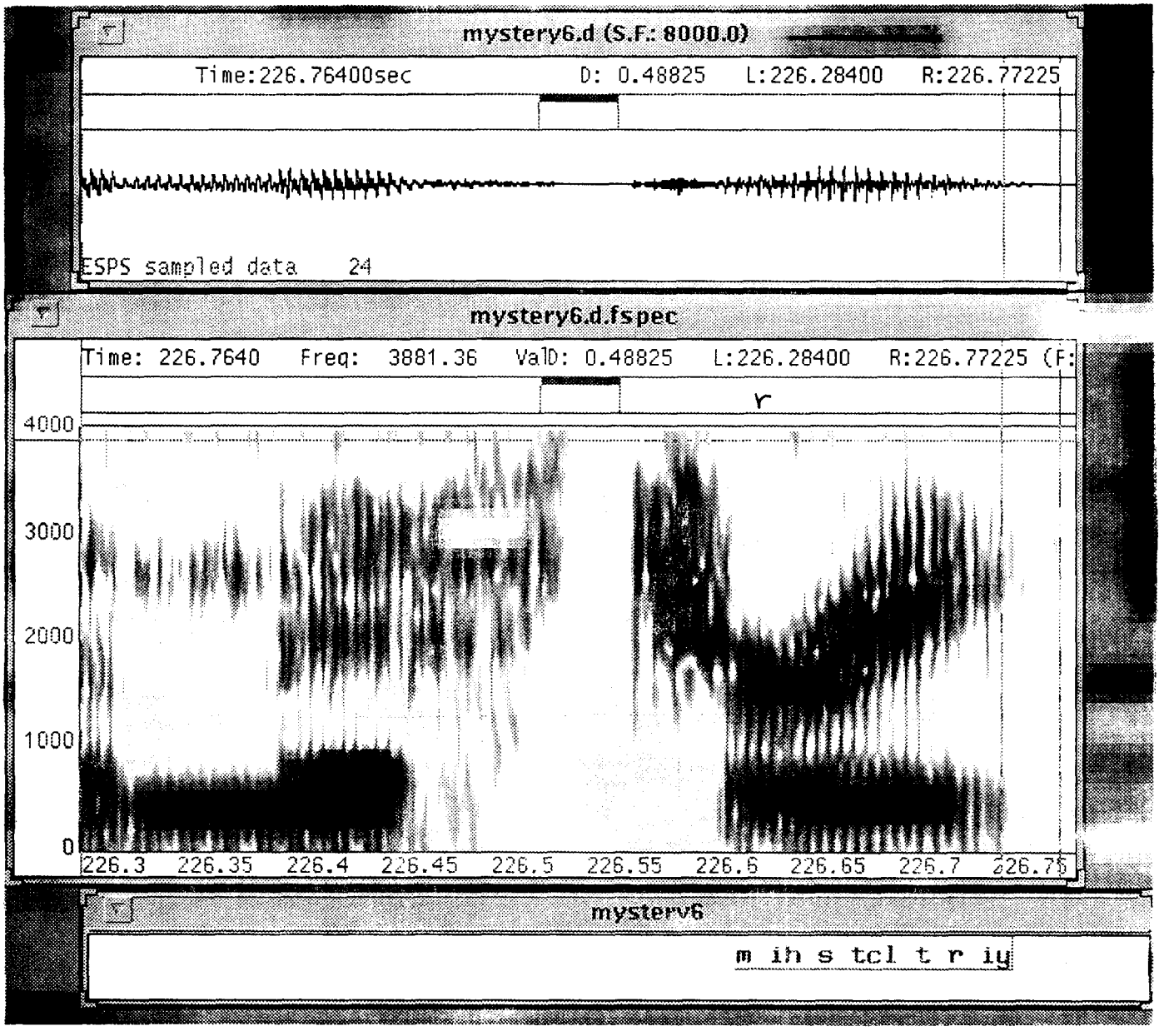
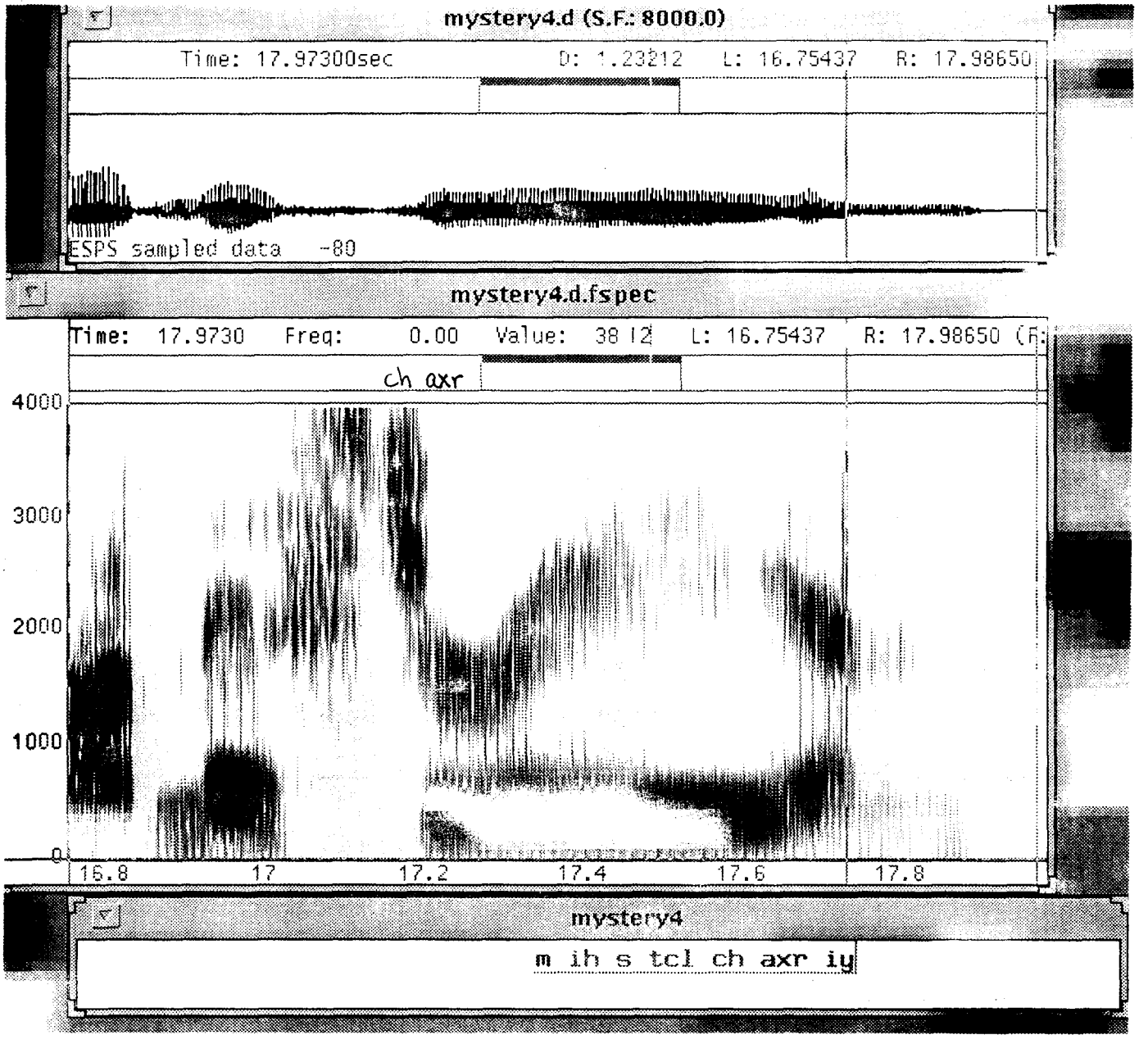


Figure 14. Example of syllabic *axr*.



Place of articulation assimilation is transcribed if it is complete. Partial assimilation is not transcribed, e.g. /s/ before /ʃ/ with lower frequency noise than is usual in /s/, but still distinct from /ʃ/, is transcribed *s*. Figure 5 ("chips") shows a token in which the frequency-lowering is deemed extreme enough to yield a place of articulation change; whether it is an assimilation (due to the previous affricate, or to the following glide *y*) or not need not be decided by the transcriber.

TIMIT correspondences : We have added new fricative symbols *ph,bh,x,gh*. In TIMIT these would presumably appear as *pcl, bcl, kcl, gcl.*

2.8. Affricates. These have *tcl* and *dcl* as their closures and *ch* and *jh* as their releases and are therefore distinguished from most stop-fricative sequences, which would be *tcl t sh* or *dcl d zh.*

Affrication of /t,d/ before approximants (e.g. *dcl jh r* for /dr/ as in "driver's") is transcribed only if the affricate percept is clear and neutralizing, with substantial noise after the release (not just a labialized or retracted stop release). See Figure 14.

TIMIT correspondences: Seems the same.

2.9. Liquids *r, l.* As discussed in section 2.6 above, our use of *r* differs in one respect from TIMIT. As in TIMIT, syllable-initial or intervocalic /ɹ/ is *r,* but we use *r* for coda /ɹ/, whereas TIMIT uses *axr.* That is, we use *r* for all non-syllabic /ɹ/, just as we use *l* for all non-syllabic /l/.

A very dark /l/ might sound like [w]. The criterion for labeling is that if F3 is high, it is *l*, while if F3 falls, it is *w*. Figure 15 ("probably") shows an example of this: note the fall of F3 between the flanking vowels.

2.10. Glides *y,w*. These are not used for off-glides of diphthongs, even if prolonged, since unit diphthong symbols are available. They are also not used as onsets for syllables which follow a diphthong; the diphthong is segmented to include any such material.

In a reduced syllable like /pju/ in "reputation", the glide might be missing altogether, or it might be manifested only as some influence on the surrounding segments. If there is no interval in the signal that might be segmented as the glide, then it is not transcribed. Figure 16 ("reputation") shows a token in which the /ju/ sequence appears as a steady fronted vowel *ux* .

As mentioned in 2.9, *w* can be transcribed for a very dark /l/.

We do not systematically distinguish phonemic voiceless-/w/ (IPA /ʍ/) for those speakers who preserve this as a phoneme distinct from /w/. If a /w/ appears completely voiceless, for any reason, it would be transcribed *w-h.*

TIMIT correspondences: Seems the same.

2.11. Voiced and voiceless /h/. This distinction is determined by voicing in the acoustic signal. Voiced *hv* is quite common.

TIMIT correspondences: Seems the same.

Figure 15. Example of *w* for /l/.

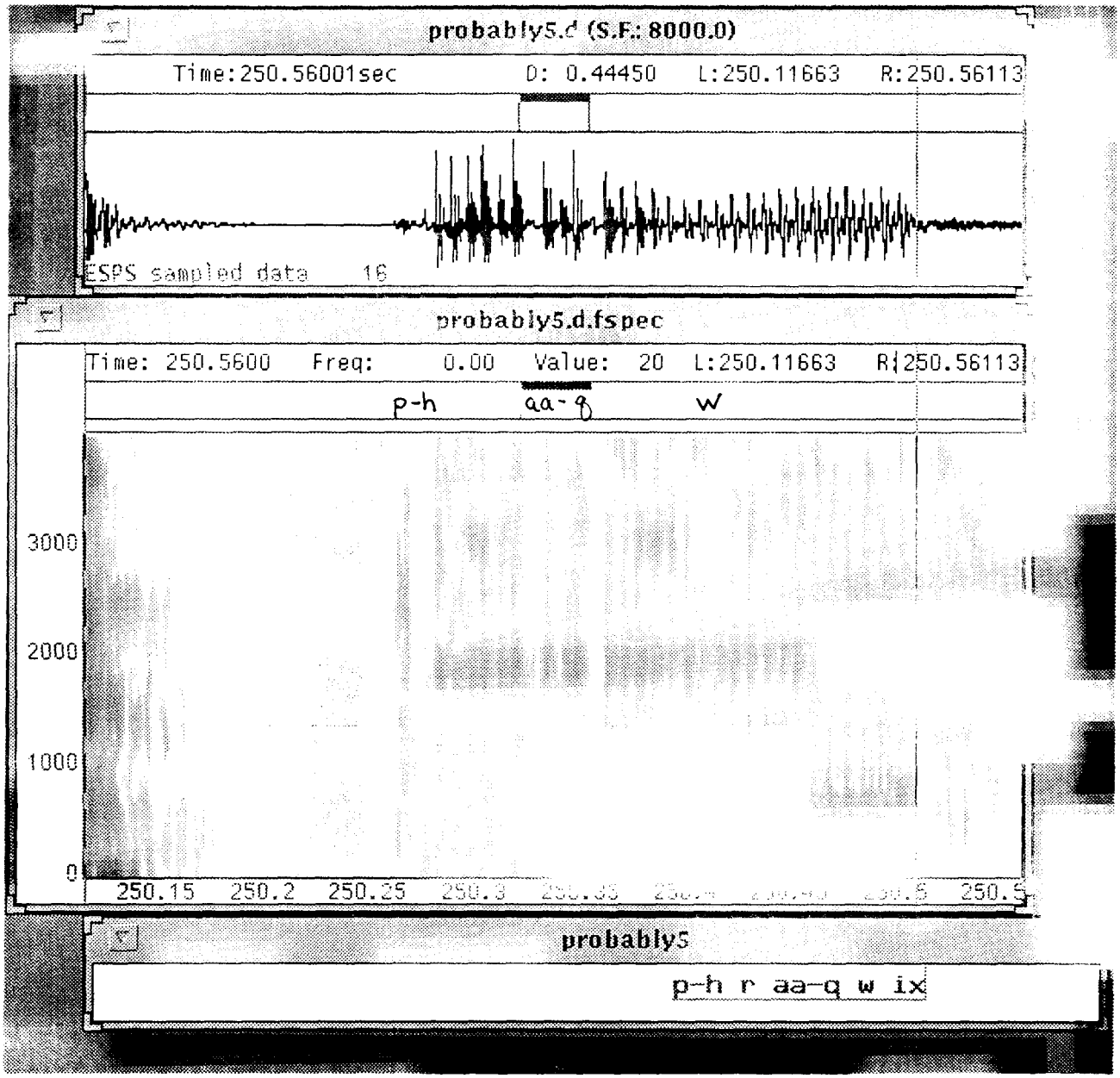
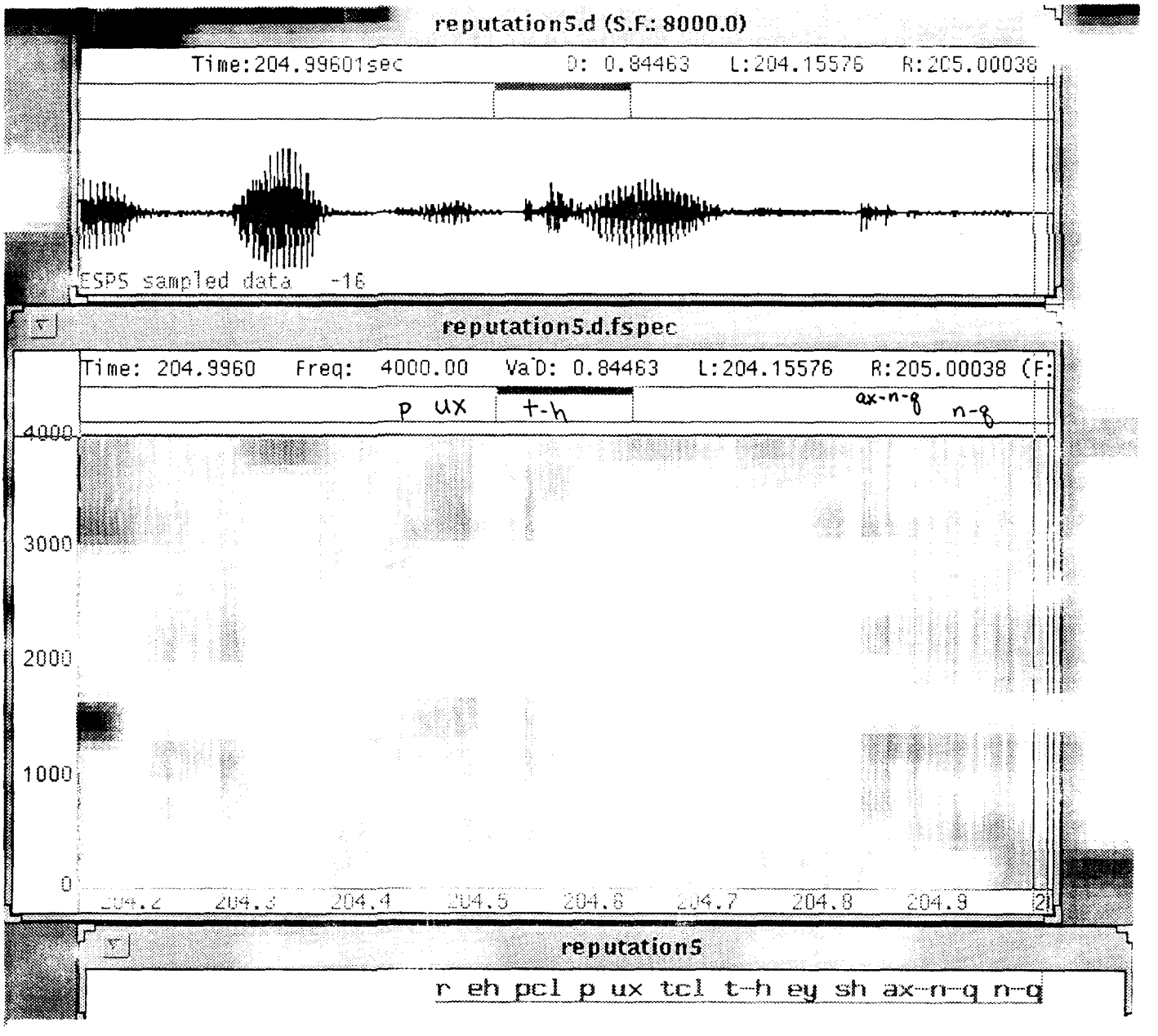


Figure 16. Example of glide-vowel coalescence.



3. Vowels

We use the TIMITBET symbols for vowels. Each symbol, whether monophthong or diphthong, contains two characters. In general, tense vowels end in *y*, lax vowels end in *h*, and centralized vowels end in *x*, except for the low vowels.

The strong preference is for one vowel segment&label per phonemic vowel. If vowel quality changes over the course of a phonemic unit, it is one segment and the dominant quality determines the label.

TIMIT correspondences: Seems the same.

3.1. Diphthongs. These are units, not segmented. No additional glides are transcribed as onsets after these, even if they are quite extreme (e.g. no *y* after *ay* or *ey* even if it sounds like [j] rather than [ɪ]). Similarly, if a vowel before a glide (or any other consonant) sounds diphthongal by anticipation, that gliding is not transcribed.

Monophthongization of diphthongs is rarely transcribed. The formants must look quite steady in the spectrogram (except of course for consonantal transitions at the margins). Most variation that sounds like monophthongization is more accurately greatly-reduced diphthongal movement, and that is something we make no effort to transcribe. Figure 10 ("about") shows one of the few diphthongs transcribed as a monophthong.

TIMIT correspondences : Seems the same.

3.2. *ax, ix, axr.* Following TIMIT, one of these is used for any short and/or reduced vowel, and they are the default set if the vowel is phonologically stressless. In practice, the full/reduced distinction is a very difficult one. We look for vowels which have 6 or fewer pitch periods, except that next to an approximant the longer transitions must be factored out. Following TIMIT, the distinction between *axr* and the others depends on F3, with *axr* having a low F3, while the distinction between *ax* and *ix* depends on the relative frequency of F2 at its strongest point. If the F2 is closer to F3, we use *ix*; if closer to F1, we use *ax*. Note that /i/ and /o/ also occur as stressless vowels in English, but these will be transcribed only if the quality is clear. One difficult distinction is *ix* vs. *iy* for short vowels; when in doubt, we prefer the label corresponding to the phoneme. Another difficult decision arises with vowels that are not short and may have some degree of stress, which sound between *ih* and *ix* (e.g. final syllable in "limited").

TIMIT correspondences: Basically the same, but we are not satisfied with critera for when to use a reduced vowel symbol over a full one.

3.3. *uw* vs. *ux*. If F2 is closer to F3, we use *ux*; if closer to F1, we use *uw*. Fronted and backed versions of *ow* and *uh* are not distinguished.

TIMIT correspondences: Seems the same.

3.4. r-colored vowels. The two vowels *er* and *axr* are to be distinguished by stress (*er* is the regular vowel, *axr* is reduced).

It should be recognized that vowels before /r/ have different qualities than the same vowels elsewhere. We have chosen not to transcribe contextual r-coloring with a diacritic or other symbols. For vowels before /r/, we follow the TIMIT convention which is to prefer the lax vowel symbols over the tense ones: *ih* and *eh* (over *iy* and *ey*), as in Figure 6

("bear"). We use iy, *ey* only if there is a distinct rising movement in F2 and F3, distinct from consonant transitions, before the fall into /r/.

TIMIT correspondences: Seems the same as what they say.

3.5. Voiceless vowels. TIMIT has only one voiceless vowel, ax-h. Because we have a general *-h* diacritic, described below, we can indicate any vowel quality as voiceless. In practice, this means that we distinguish *ax-h* from *ix-h;* these two are distinguished according to F2, as for the voiced counterparts. Figure 17 ("reputation") contains 2 voiceless vowels.

TIMIT transcribes a voiceless vowel even for vowels with one or two pitch periods, but we use it only for completely voiceless vowels.

As also described in 4.2 below, if a voiceless consonant is followed by a voiceless vowel, such that there is a choice between transcribing aspiration on the consonant and devoicing of the vowel, we transcribe the voiceless vowel but not aspiration on the consonant. (This is an arbitrary solution to a logical difficulty in segmental transcription.)

TIMIT correspondences: We are more conservative than TIMIT in what counts as voiceless, and in what counts as *ax*. Our *ix-h* should be folded into our *ax-h* for correspondence with TIMIT.

4. Diacritics

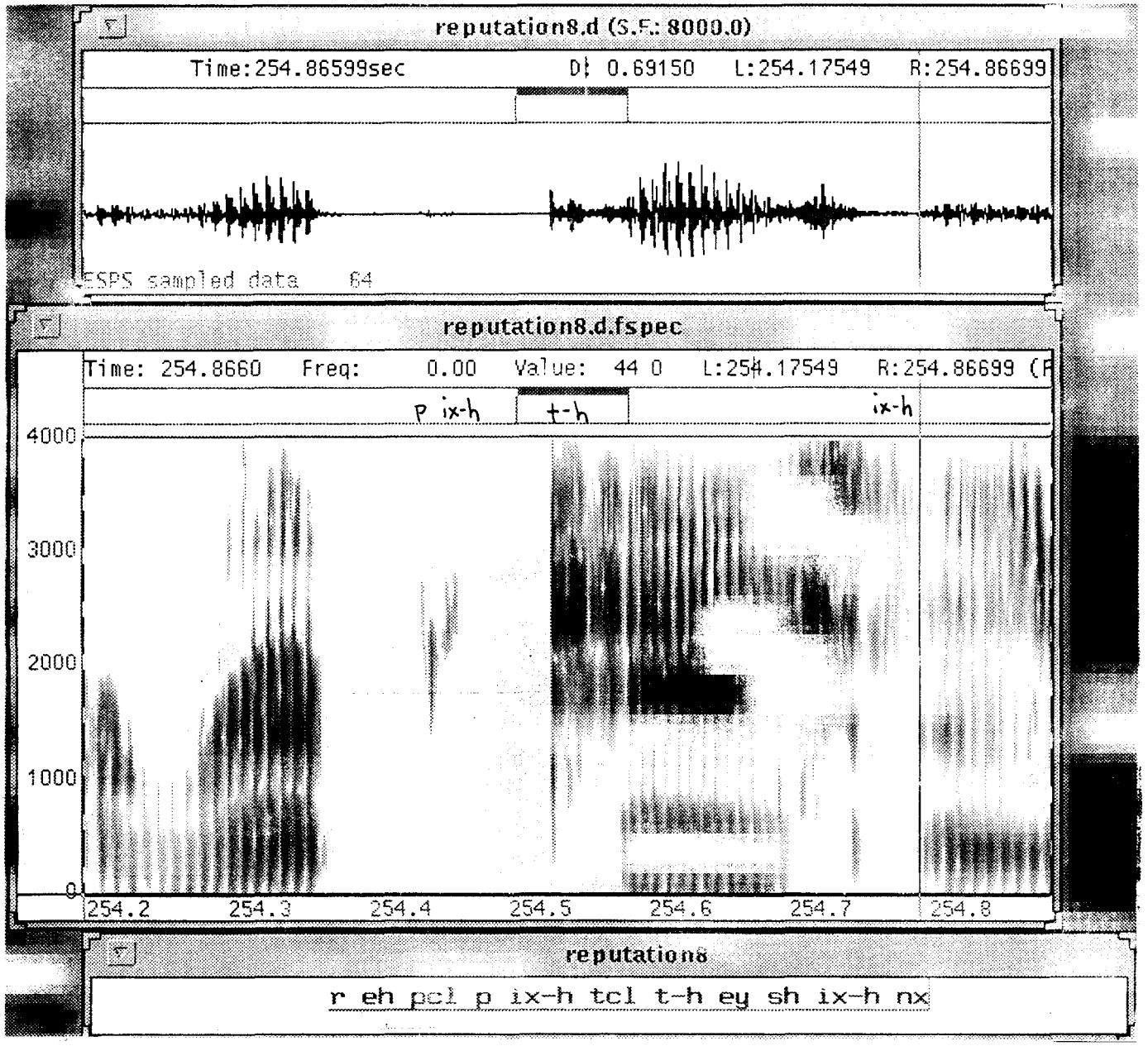
A major difference between our symbol set and the TIMITBET is that we provide three diacritics which can be freely added to other segments. Multiple diacritics on a single segment are listed in alphabetical order (for example Figure 16). No attempt is made to render any temporal ordering of the features involved.

4.1. *-n.* This is a nasalization diacritic applied to an entire vowel to indicate audible nasalization over any part of the vowel. Most instances involve nasalization of a vowel near a nasal consonant, but a few instances are of vowels which are heavily nasalized seemingly independent of the context. No effort is made to segment a nasal vs. non-nasal portion of a vowel, and the criterion for use of *-n* is auditory rather than acoustic: we listen to the vowel segment out of context, especially without the couple of pitch periods adjacent to any contextual nasal consonant, and if it sounds nasal, we use the diacritic. Various figures above show vowels adjacent to nasals with and without *-n*.

Nasalized vowels often show a quality shift compared to their non-nasalized counterparts. We generally do not transcribe such differences. The use of *-n* indicates the possibility of a quality difference, as well as nasalization.

4.2. *-h.* This is for sounds made with the vocal cords audibly spread apart, that is, two kinds of sounds: devoiced versions of voiced phonemes, and aspirated versions of voiceless stops. With respect to devoiced allophones of voiced phonemes, one use is for reduced vowels which are entirely voiceless--*ax-h, axr-h,* or *ix-h*, using the same F2/F3 frequency criterion as for voiced versions. Another use is for utterance-final devoicing of any sonorant, such as nasals. Partial devoicing of voiced phonemes is not generally transcribed.

Figure 17. Two examples of voiceless vowels.



With stops as base symbols *-h* indicates aspiration, e.g. *p-h* means an aspirated /p/. The rule of thumb used in transcribing aspiration is that labials and alveolars must have a VOT of 21 or more msec, and velars must have a VOT of 35 or more msec, always measured from the beginning of the **last** release burst. Figures 15, 16, 17 contain aspirated voiceless stops while Figures 2, 7, 9, 11 contain unaspirated voiceless stops. Figure 1 illustrates that this diacritic can be used even with a "voiced" stop as a base, e.g. *d-h*, because the stop has a voiceless release and clear aspiration beyond the required VOT value. In principle the affricate release *ch* could be aspirated as well.

In a consonant cluster beginning with a voiceless consonant, as in Figure 15, the "aspiration" is seen in effect as partial devoicing of the following consonant. We label this simply as *-h* on the stop, and do not also mark the following consonant as voiceless. This is for consistency with aspiration before vowels -- we never indicate that part of the following vowel is voiceless. Thus *-h* is interpreted as some devoicing of whatever segment follows the aspirated stop. The exception to this general rule is if the following segment is completely voiceless: in that case, the complete voicelessness takes precedence in labeling and we label for example the vowel as voiceless, but no aspiration on the consonant. For example, if the initial stop in "conditions" were aspirated, we could have either *k-h ax* or *k ax-h*, depending on whether the vowel is partially or completely voiceless after the /k/. Thus devoicing is labeled only once in a given sequence.

4.3. *-q*. This is a laryngealization (or "glottalization" or "creaky voice") diacritic for voiced sonorants. It is used for a segment which is partially or completely laryngealized; the entire segment is given the diacritic no matter how much or how little of the segment is laryngealized, and the diacritic follows the base symbol regardless of where in the segment the laryngealization occurs. Examples include vowels in hiatus (surprisingly rare); sonorants before glottalized coda consonants (Figures 2, 5); sonorants adjacent to *q* itself; intonational creak (Figures 15,16). However, if every segment in a speech sample is laryngealized, i.e. that is simply the speaker's usual voice quality, then it is not transcribed.

The difference between *-q* and the glottal stop (*qcl q*) is that with laryngealization, a sonorant maintains its particular F-pattern; that is, along with the constricted glottis there is a clearly identifiable primary supraglottal articulation.

TIMIT correspondences: TIMIT has no such general diacritics. (1) Our vowels with *-n* should be mapped into the vowels without diacritics, and at the same time the last pitch period (or 10 msec) of each such vowel should be segmented out and labeled *n,* to correspond to TIMIT. (2) Our i*x-h* should be mapped into *ax-h*, and any of our vowels with *-h* that follow the same vowel symbol without *-h* should be merged into that symbol. *-h* on any consonant symbol should be omitted. (3) Vowel symbols with *-q* should be mapped to TIMIT *q*. Other symbols with *-q* should omit the *-q* to correspond to TIMIT.

5. Segmentation/time-alignment

No word-internal temporal segmentation was performed, but a time alignment for the target word as a whole was provided. The segmental transcription and time alignment include only the target word. In general this segmentation was minimal, allotting to the target word only material that unambiguously belonged to that word. No segments from adjacent words and no adjacent pauses are included in the transcription of the target word or are otherwise transcribed, except that shared segments are presumed to belong in both words' transcriptions (5.3 below). Thus, if a segment is seen in a transcription, it was part of the pronunciation of that token. For example, if "bear" is transcribed as beginning with [mb], that means that there was no neighboring /m/ but that instead the speaker produced a

prenasalized /b/. Other conventions for segmentation at word boundaries include the following:

5.1. Boundary between two vowels. If there is an abrupt frequency or amplitude change suggesting an acoustic boundary, that is used as the boundary. If there is no clear boundary, then any transition between two vowels is not included in our target word.

5.2. Boundary between two consonants. Between two stops, the general TIMIT convention for stop clusters is followed: If the first of two stops is unreleased and the closure interval is otherwise indivisible, and if the transitions into the closure are consistent with the first stop (as judged from listening), then the entire closure is assigned to the first stop and thus the first word; the release is assigned to the second stop and thus the second word. If the transitions into the closure are consistent with the second stop, then the closure is assigned to the second stop and thus the second word (the first stop is treated as deleted). The default option is to give the whole closure to the first stop in this fashion. Distinct positive evidence can however be the basis for dividing the closure between the two stops: a release of the first stop in mid-closure, or a change in voicing consistent with the phonemic sequence.

Between consonants, especially approximants, where transitions are seen, the transition is divided at an acoustic boundary given by an abrupt change in amplitude or frequency.

5.3. Boundary between a shared consonant. Following TIMIT, if two identical consonants are adjacent across a word boundary and are realized as a single long or short consonant, then it is given one phonetic symbol but assigned to both of the words. That is, the first word ends at the end of this segment, and the second word begins at the beginning of this segment. Because we transcribe only single words we do not have to deal with the technical problem of notating that a consonant belongs to two words at once. (It is also possible to have two separate but identical consonants across a word boundary, if some boundary between them can be located, e.g. if there is an amplitude drop or glottal stop between them.)

5.4. Boundary between consonant and vowel. Between a consonant and a vowel there can be extensive formant transitions. In our segmentation these transitions always go into the vowel, even when they are very long (as when the consonant is *w, y, r* or *l)*. The consonant is therefore segmented as just a steady state, and the vowel is segmented to include all formant transitions as well as any steady-state portion it may have.

Note that diphthongs are transcribed as unit segments. That is, off-glides are not distinguished from peaks, and so no segmentation between these portions needs to be determined in our transcriptions.

6. Silences

epi. We chose not to use this TIMIT symbol at all, as its use in TIMIT transcriptions was not entirely clear to us.

pau. We transcribed only words that had been orthographically transcribed as complete, so that there were no word-internal pauses.

TIMIT correspondences: Some instances of TIMIT *epi* correspond to some of our use of stop closures in our transcriptions, in a way that cannot be automatically converted in either direction.

ACKNOWLEDGEMENT

This work was supported by a contract from the Linguistic Data Consortium at the University of Pennsylvania.

REFERENCES

Allen, G.D. (1988). The PhonASCII system. *Journal of the IPA 18*(1): 9-25.

Henton, C. and Bladon, A. (1987). Developing computerized transcription exercises for American English. *Journal of the IPA 17*(2): 72-82.

Hieronymous, J.L. (n.d.). ASCII Phonetic Symbols for the World's Languages: Worldbet. Ms., AT&T Bell Laboratories.

Garofolo, J.; L. Lamel; W. Fisher; J. Fiscus; D. Pallett; N. Dahlgren (1993). DARPA TIMIT. Distributed with TIMIT on CD-ROM, second (full) release, 1990. Section 5.2, Notes on Checking the Phonetic Transcriptions, by L. Lamel.

Metzler, T. and T. Nathman (1993). Labeling conventions. Ms., Center for Spoken Language Understanding, Oregon Graduate Institute.

Seneff, S. and V. Zue (1988). Transcription and alignment of the TIMIT Database, Distributed with TIMIT on CD-ROM, first (prototype) release.

Review of the Oxford Acoustic Phonetic Database on compact disc by J.B. Pickering and B.S. Rosner

Patricia Keating

Oxford University Press, 1993, 2 compact discs and 196 pp. ISBN 0-19-268086-2. Price £ 250 ($412.74 including tax and shipping to USA).

The *Oxford Acoustic Phonetic Database* (henceforth *OAPD*) is a recording of readings of wordlists in seven languages. The basic design of the speech materials is to focus on variability in vowels, in terms of language, speaker, and phonetic context effects. For each language, four males and four female native speakers were recorded. The languages, and the dialect of each chosen for the recording, are English (two dialects: northern New Jersey American and RP British), French (Paris and Loire Valley), German (Lower Saxony "Hochdeutsch"), Hungarian (Budapest), Italian (northern -- mostly Milan), Japanese (Tokyo), and Spanish (Madrid area). The speakers were chosen to be homogeneous in age (18-30 years) and background (middle-class), and fluent readers. The context variables are both prosodic and segmental. Within each language, the wordlist was designed to provide vowels in accented and unaccented syllables, and combined with all consonants, at least to the extent possible using only real words in each language. Monophthong vowels were also recorded in isolation. Each speaker provided two tokens of each word.

This design gives 704 items for American English, 806 for British English, 576 for French, 754 for German, 971 for Hungarian, 449 for Italian, 489 for Japanese, and 382 for Spanish. With eight speakers and two repetitions for each item, that yields a total of 82,096 utterances, an impressive amount of speech. (Compare TIMIT at 6300 utterances and about 54,391 word tokens.) Each word is in a separate headerless digital file in DOS format, and the entire set of recordings is distributed on two CD-ROM discs. Unlike with TIMIT, no text files are included on the CD.

The *OAPD* is thus more extensive than the kind of recording typically made in phonetics, that is, a recording designed for a particular experiment. At the same time, it is more like that kind of recording than it is like some general-purpose speech databases, such as TIMIT or Switchboard, because in the *OAPD* a few speakers each produce the entire list of items, and each item is a single word selected to represent some particular combination of interest.

The speech was recorded directly to computer disk, using a directional microphone positioned 1 meter from the speaker, in a sound treated room. Each speaker started and ended the recording of each word by the computer, and spoke at a self-selected rate and loudness. When we received the *OAPD,* we found we had made an incorrect and naive assumption. We thought that because the database is distributed on CD-ROM discs, that it was therefore recorded at the CD standard of sound quality, that is, 44kHz sampling rate and 16 bit quantization rate. Instead, it is recorded at an older standard, namely 10kHz sampling rate (filtered to 4.7kHz) and 12 bit quantization rate. While this is certainly adequate for vowels, which the database is designed to study, it is not ideal for obstruent consonants.

It must be stressed that the speech materials are organized around the topic of variability of vowels. The documentation of the *OAPD* is devoted largely to discussion of which factors that potentially determine vowel variability are varied in the database design,

and which factors are kept constant. For example, the only speaker variation is in sex; the manual discusses this and other effects of speaker differences, and describes efforts to obtain a homogeneous set of speakers for each language -- not just in terms of age, background, and linguistic dialect, but even to the extent that the male and female speakers within each language are of generally similar build. Other factors which were not varied include speaking rate and speech style, as is typical of most speech corpora. Finally, the database is not designed to study consonant variation. Although all consonants of each language are represented in the wordlist, combined with all vowels, the rest of the consonantal context is not controlled. Some questions about consonants can be studied, to an extent, but the potential user should be keenly aware that this is entirely accidental in terms of database design.

The important point to be made about this is that the database contains few minimal pairs. Figure 1 reproduces the first page of the listing of the database, the words illustrating /i/ in American English. These words provide a nearly complete set of preceding and following consonant contexts in both stressed and unstressed syllables. However, neither the effect of consonant context nor that of stress can be studied directly because so much else varies randomly together with these conditions. For example, the following /s/ versus /ʃ/ contexts in the stressed syllable condition cannot be compared directly because the first context occurs in a disyllable with no preceding consonant (*Easter*) , whereas the latter occurs in a monosyllable with a preceding /l/ (*leash*). This paucity of minimal pairs seems to reflect a strategy to have one word provide more than one target context wherever possible. For example, for /i/ *pediatric* provides both the unstressed CV after /p/ and the unstressed VC before /d/, but this precludes a controlled comparison of unstressed /pi/ with unstressed /pI/ (in *picturesque*), /pɛ/ (*apex*), /pæ/ (*pacification*), etc. The same criticism can be made of the wordlists for the other languages.

Some more standard measures that cannot be made in the *OAPD* are:
(1) Inherent vowel duration (discussed for English, German, and Spanish in Lehiste 1970): for English, most stressed vowels are given in a /p_p/ frame, but for the other languages this level of consistency is not reached, though several minimal pairs may be found. Note that number of syllables per word needs to be matched for any such duration comparison. The isolated monophthong tokens might be used for such a study, but sometimes the boundary between speech and silence is quite hard to locate. (2) Fundamental frequency measurements: because the utterance-level intonation contours are variable both within and across speakers, any FO measurements require extra care. (3) Vowel-to-vowel coarticulation: not only can this not be studied in the database, but it is a potential confound in any measurement of vowel quality, since the nonadjacent context is not controlled. On the other hand, Voice Onset Time (discussed for English, Hungarian, and Spanish by Lisker and Abramson 1964, and for some of the other languages by others since) is perhaps the most plausible consonant measurement, since the stressed CV sequences are almost always word-initial, and the effect of other variables is likely to be limited. The effects of place of articulation and following vowel quality and (phonemic) length on VOT could be determined in most languages; the Japanese CV combinations are the spottiest.

In some cases, the number of tokens of a combination will be large enough that other variables might not matter. For example, intrinsic FO of vowels could be assessed by collapsing all consonant contexts and trusting that any intonational effects will not be systematic. This is the method used in the study of large databases such as TIMIT or collections of interviews as recorded by sociolinguists: tokens may come from all kinds of contexts, both segmental and prosodic, but the effects of those contextual differences should be randomly distributed in the measurements. The point to be made here is that the

American English (New Jersey) · Classified List Vowel /i/

	Stressed VC		Stressed CV, CVC	Unstressed VC	Unstressed CV
Plosive					
/p/	heap 071		pea 169 peep 112	prepacked 227	pediatric 270 happy 271
/b/			bee 321 beet 371	debug 322	automobile 323 baby 372
/t/	eat 171		tea 226	athlete 225	ovaltine 170 duty 480
/d/	Eden 072		'd' 168 deed 111	pediatric 270	seedy 481
/k/	cke 482		key 483	decompose 268	Cherokee 069 leaky 070
/g/	eagle 269		geese 423	colleague 073	leggy 001
Fricative					
/f/			fee 324 fief 167	defenestration 369	coffee 110
/v/	eve 370		'v' 272	evangelic 484	gravy 485
/θ/	ether 541		thief 479	rethink 542	frothy 113
/ð/			thee 543 these 228		
/s/	Easter 544		see 486	desegregation 566	axes 662
/z/	ease 661		'z' 664	odysseys 114	benzene 273
/ʃ/	leash 422		she 424	reshaping 540	banshee 546 flashy 166
/ʒ/					
/h/			he 320 heed 002	prehensile 568	hedonic 601
Affricate					
/tʃ/	each 172		cheat 068		preachy 224
/dʒ/	aegis 325		'g' 539		geocentric 663
Approximant					
/l/	eel 165		lea 074	airfield 602	acetylene 421
/ɹ/			spree 616 real 223	prerequisite 274	cherry 173 reunion 327
/w/			we 267 wheel 646		kiwi 326 dewy 487
/j/			ye 115 yield 003		
Nasal					
/m/	seem 478		me 425	demist 547	antihistamine 569 creamy 229
/n/	mean 108		knee 368	denotation 420	tiny 538

30

Figure 1. First page of the listing of the database: vowel /i/ in American English

careful construction of the wordlists in the *OAPD* buys you little beyond what a large random sample would.

Some potential users may be disappointed with the range of languages chosen for the database. The lack of even a single tone language is particularly striking. If language diversity is a prime concern, consider instead the *Phonetic Database* that Kay Elemetrics Corp. sells. It contains recordings of 35 languages, many of them relatively little-studied. (Only Japanese and Hungarian are in common between *OAPD* and the Kay database.) Unlike the *OAPD*, the Kay database has only one speaker per language and a very limited language sample -- usually about 50 words plus some sentences. The words are usually not in minimal pairs and seem to have been chosen to illustrate points of phonetic interest, though this is not explained and is not always obvious from the wordlists.

Another disappointment is the lack of any discussion of prior acoustic phonetic work on each of the languages in the database. This would have made it easy to see what is already known about each language, and to compare any measurements made from this database with earlier ones. Similarly, it would be helpful to have basic references on the phonology of each language. The set of phonemes for each language is given before the wordlist, but the only language references are to dictionaries for two of the included languages. Even the correspondence between phonemes and standard orthography has to be figured out by the user from the examples in the lists, since the words are not transcribed phonemically and no key to spelling is given. Also, no English glosses are given for the non-English words. In comparison, the Kay Elemetrics *Phonetic Database* gives a phonetic and phonemic transcription and a gloss in addition to an orthographic form of each utterance, and also basic information, phoneme charts, and at least one reference for each language.

The publicity for the *OAPD* makes clear that a waveform editor is needed to display and hear the speech files. Unlike TIMIT, however, the *OAPD* does not come with any routines for converting to other file formats. The manual includes some hints on converting the files for use with the CSRE or Signalize editors, but users should be aware that basically they must deal with this issue themselves. I can add two notes on this point. the first is that the "cget" command in Paul Milenkovic's CSpeech system for PCs seems to work quite satisfactorily. The second is that the "data import/generic data" menu function in Kay Elemetric's CSL for PCs produces weak (albeit adequate) signals when used with its default settings. CSL expects 16 bit data, while the *OAPD* is only 12 bit; therefore the signal appears weak in CSL. (CSpeech, in contrast, expects 12 bit data.) Kay Elemetrics advises CSL users with 12 bit data to use the "scale" command to boost the gain by a factor of about 16, into the normal range for CSL.

In summary, the idea of the *OAPD* is a good one, but two deficiencies in its execution must be emphasized. First, the quality of the sound files is not ideal. Considering the price of the database, it would be reasonable to expect higher sampling and quantization rates, even though that would require more discs. Second and perhaps more importantly, the design of the materials does not permit the user to ask many kinds of standard questions. In part this is unavoidable -- no database can be truly general-purpose -- but the near-absence of minimal pairs severely limits the kinds of controlled comparisons that can be made. For some users, the database might prove valuable if only for instructional or illustrative purposes: quite limited sets of minimal pairs can be found without too much effort. In the end, potential buyers should consider carefully whether the database will meet their needs well enough to justify the expense. For those whose needs it meets, even if marginally, it is a valuable resource.

References.

DARPA TIMIT CD-ROM. NIST Speech Disc 1-1.1, October 1990. Distributed by National Technical Information Service (No. PB91-505065) and by the Linguistic Data Consortium.

Kay Elemetrics Corp. and University of Victoria/Speech Technology Research Ltd. (1991). *Phonetic Database.* CD-ROM distributed by Kay Elemetrics Corp.

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

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

Modeling Coarticulation and Place of Articulation Using Locus Equations [1]

Court S. Crowther
UCLA Phonetics Laboratory

INTRODUCTION

Most consonants in consonant-vowel (CV) syllables have multiple acoustic correlates in the speech signal, and researchers have had difficulty isolating an acoustic correlate that distinguishes each consonant and does not depend upon vowel context. Although the acoustic realization of a particular consonant seems to be context dependent, listeners categorize the context-varying acoustic signals as instances of the same phoneme. A number of different approaches have been proposed to resolve this "non-invariance problem," many of which have been shown to fail to hold under some circumstance. In this paper I examine the locus equation approach (Nearey and Shammass, 1987; Sussman, 1989; Sussman, Hoemeke, & Ahmed, 1993; Sussman, McCaffrey, & Matthews, 1991) to the non-invariance problem for stop consonant identification in CV syllables. The main goal of this study was to determine how well locus equations resolve /b/, /d/, and /g/ when degree of coarticulation is varied in naturally produced speech. /b/, /d/, and /g/ are found to be differentiated better when two locus equations are derived for each stop, one for Front and one for Back Unrounded vowel contexts. Effects of place of articulation and degree of coarticulation are not dissociated in locus equation parameters. The relationship between Klatt's (1987) locus equation and the locus equation utilized by Sussman and colleagues is examined, and the relative merits of these equations are discussed.

Several approaches to the non-invariance problem have sought to identify in the speech signal spectral properties of the release burst and points thereafter that do not differ with vowel context. According to these accounts, which will be referred to as "acoustic signature" accounts, each stop consonant is thought to have an acoustic specifier (Fowler, 1994) that does not depend upon the following vowel. Stevens and Blumstein (Blumstein and Stevens, 1979; Stevens and Blumstein, 1978) have proposed that gross shape of release burst spectra distinguishes /b/, /d/, and /g/ and is invariant across vowel context. Kewley-Port's (1983) approach to the problem is similar to that of Stevens and Blumstein, but is more dynamic in that it takes account of spectral shape not only at release, but at several successive time points up to 40 ms after release onset. Lahiri, Gewirth, and Blumstein (1984) proposed that relative changes in the distribution of spectral energy at low and high frequencies from burst onset to voicing onset, rather than spectral shape *per se*, specify place of articulation. An extensive review of these proposals can be found in Sussman, *et al.* (1991).

Forrest, Weismer, Milenkovic, and Dougall (1988) provided an approach that is related to the three accounts discussed above. They quantified the invariant spectral shape notion by examining the first four moments (mean, variance, skewness, and kurtosis, respectively) of burst spectra for /p,t,k/ in initial position in CVC's. Their spectral moments technique yielded place of articulation information that allowed impressive (approximately 80%) place classification with spectral moments computed for a time slice encompassing release burst onset to burst offset plus 10 ms. When moments from a second spectral cross-section (encompassing the offset of the previous cross-section plus 10 ms) were included in the analysis, classification accuracy improved substantially for the male speakers' tokens, but less so for the female speakers' tokens.

Two of the more popular speech perception theories do not have difficulty handling a lack of acoustic invariance, and in fact do not assume that invariance is necessary for

[1]Portions of this paper were presented at the 127th meeting of the Acoustical Society of America, Cambridge, MA, Spring, 1994 (Crowther, 1994).

speech perception. For example, motor theorists (Liberman and Mattingly, 1985) take acoustic non-invariance as evidence that the speech signal is heavily encoded. That listeners nevertheless are able to decode the signal is, in their view, support for the claim that listeners possess tacit knowledge of speech production. The Fuzzy Logic Model of Perception (Massaro and Oden, 1980; Oden and Massaro, 1978), or FLMP, assumes that listeners match properties of speech signals to syllable sized prototypes stored in long term memory. To illustrate, consider the likely nature of the prototypes for the syllables "ki" and "ku". The F2 transitions for these syllables tend to be dissimilar, so the F2 transition information in the "ki" prototype would be different from that in the "ku" prototype. From the perspective of FLMP, F2 transition information is but one piece of information among many that is included in the syllable prototypes. Because FLMP assumes that the syllable is the perceptual unit, and that listeners' decisions are based upon many, integrated sources of information, rather than upon one source of information (i.e., an invariant cue), listeners do not require acoustic invariance.

The locus equation approach to the non-invariance problem represents a conceptual departure from the acoustic signature accounts. Rather than seeking an acoustic signature for each consonant that is independent of vowel context, the approach exploits the systematic nature of the context dependence that exists between the second formant frequency at onset (*F2on*) and at its target value (*F2t*). In CV syllables, *F2on* varies according to place of articulation; however, there is no measurable, invariant *F2on* that characterizes each place, because *F2on* depends upon *F2t*. The consonant is "coarticulated" with the vowel, and the dependence of *F2on* on *F2t* is the acoustic consequence of C's coarticulation with V. For a given stop, the dependence between *F2on* and *F2t* is quite consistent, and is captured well with linear, locus equations

Lindblom (1963) introduced locus equations: for CV syllables, where C is a particular stop, and V ranges over the speaker's vowel inventory, when *F2t* is plotted on the horizontal axis and *F2on* on the vertical axis, the points fall on an approximately straight line (hereafter, the "locus line"). Locus equations fit a straight line to the *F2on*, *F2t* data points:

$$F2on = k*(F2t) + c\,, \qquad (1)$$

where k, a constant, is the slope of the line and c, in units of Hertz, is its intercept. Equation 1 predicts *F2on* on the basis of knowledge of *F2t*, and the fact that it usually fits the data well for CVC syllables suggests that the relationship between *F2on* and *F2t* is linear (see Sussman, 1994). Currently, locus equations are utilized to characterize *place of articulation* (Nearey and Shammass, 1987; Sussman, 1989; Sussman, *et al.*, 1991, 1993) and also *degree of coarticulation* (Krull, 1987, 1988) of C_1 in C_1VC_2 syllables.

Sussman *et al.* (1991) applied discriminant analysis to locus parameters derived from *F2on*, *F2t* measurements of /bVt/, /dVt/ and /gVt/ produced by 20 native English-speaking subjects. The analysis indicated that the slope and intercept parameters define a slope-intercept space that distinguishes the stops perfectly (and see Nearey and Shammass, 1987). Estimated parameters for locus equations reported by Sussman and colleagues typically have:

large slopes & small intercepts for /b/

intermediate slopes & intermediate intercepts for /g/

small slopes & large intercepts for /d/

Larger slopes are associated with greater degrees of coarticulation (Krull, 1987). Slope and intercept parameters change in a systematic way with place of articulation. Sussman, et al.

(1991) report a slope-intercept correlation of -0.94, and Fowler (1994) reports a somewhat smaller correlation of -0.9, both studies suggesting that the parameters provide redundant information. One goal of this study was to determine whether slope and intercept parameters would be correlated when degree of coarticulation was an independent variable. A smaller correlation may indicate that locus parameters dissociate the effects of coarticulation and place of articulation.

Sussman (1989; Sussman, *et al.*, 1991) proposes that human listeners may identify stop consonants in the face on non-invariant acoustic information (*F2on*) in a manner analogous to the barn owl's computation of sound source azimuth given non-invariant phase difference information. The barn owl is thought to use interaural time difference (ITD) to determine sound source azimuth (e.g., Wagner, Takahashi, & Konishi, 1987). It is clear that ITD is potentially informative for object location: a sound arriving at an owl's left ear before its right ear (the difference in arrival time being the ITD) indicates that the sounding object is to the left of the owl. Individual neurons in the central nucleus of the barn owl's inferior colliculus are narrowly tuned to frequency, but, if sensitive to ITD, respond maximally to ITD's differing by integer multiples of the period of a source tone. The relative firing rate of these neurons indicates interaural phase difference. Because many different ITD's are consistent with a given interaural phase difference, these neurons can not by themselves indicate a unique ITD. However, experimental evidence suggests that unique ITD's emerge (or are derived) through the computational efforts of arrays of neurons (including the phase ambiguous neurons), each of which forms "a physiological and anatomical unit" (Wagner, *et al.*, 1987, p. 3105). This allows the barn owl to derive a computational map (Knudsen, du Lac, & Esterly, 1987) of azimuth in the external nucleus. Neurons in this nucleus respond uniquely to ITD, and are associated with specific azimuthal locations.

Sussman proposes that, analogous to the barn owl's neural arrays that compute azimuth on the basis of ITD's, human listeners may possess something like a neural realization of locus equations that compute unique stop consonants on the basis of ambiguous acoustic information. Presumably, the listener samples *F2on* and *F2t* in the speech signal and derives stop consonant identity by passing these (ambiguous) values through a locus circuit that computes a unique stop (see Sussman, 1989, for details). Of course, as Sussman (1989) notes, the situation for human speech perception differs from that for sound source localization in the barn owl:

> "Unlike the barn owl situation in which the many frequencies contained in the complex stimulus input are extracted at the moment of processing, the on-line speech stimulus (e.g., a given stop + vowel) contains only the auditory properties of the phonemic segments currently under analysis. Ultimately, the language learner must develop a processing mechanism capable of coactivation of stored vowel-context representations" (p. 638).

Fowler (1994) has argued that the slopes of locus equations specify coarticulation directly, but place of articulation only indirectly. The first part of her argument involves understanding why one should expect the slope to vary with place of articulation. She notes that production of /b/ does not involve the tongue body as a main articulator, and therefore the tongue is free to coarticulate with the vowel. For consonants like /d/, however, the tongue is the main articulator for both the stop and the vowel, so, "intuitively," such consonants "do not permit the vowel to pull the tongue very far away from the consonant's characteristic locus of constriction" (Fowler, 1994, p. 600). /d/ has more "coarticulatory resistance" than /b/. When there is less coarticulatory resistance, *F2on* and *F2t* values should be more similar, hence the slopes of locus equations corresponding to /b/ should be expected to be larger than those for /d/. Likewise, as compared to stop production, fricative production may require more precise control over tongue position, and therefore fricatives

may be expected to coarticulate less with vowels. By this reasoning, Fowler predicted that, place held constant, locus equations for fricatives should have smaller slopes than those for stops. She derived locus equations for both /d/ and /z/ and found that, as predicted, the slopes corresponding to /d/ were statistically larger than slopes for /z/. These findings suggest that locus equation slopes are able to capture place of articulation only because coarticulation resistance differs with place. Further, locus equation slopes can indicate coarticulation resistance for consonants sharing the same place of articulation (i.e., /d/ vs. /z/). These considerations strongly support her conclusion that locus equation slopes convey coarticulation information directly, but place information indirectly.

Experiment 1 is an empirical study of coarticulation. *F2on* and *F2t* were measured in natural CV tokens produced under three different coarticulation conditions. Two coarticulation measures that involve distributional properties of *F2on* and *F2t* are introduced. An articulatory measure of coarticulation based on EPG data is introduced, and the results are used together with the acoustic coarticulation measures to verify that different degrees of coarticulation were induced by the experimental manipulation. Then, Equation 1 is used to model the acoustic data obtained in Experiment 1 to determine whether locus equation representations distinguish /b/, /d/, and /g/ when coarticulation is a variable. Equation 1 is compared with a similar locus equation that was introduced earlier by Klatt (1987). Finally, Experiment 2 is a pilot experiment that tests predictions derived from locus theory for listeners' identification of tokens collected in Experiment 1.

Experiment 1

To generate acoustic data for the locus equation analysis, a female native English speaker produced CV syllables with three different degrees of coarticulation. *F2on* and *F2t* were measured and the distributional properties of these parameters were examined. The first acoustic coarticulation measure is dispersion (variance) of *F2on*. When there is little coarticulation in CV tokens, the variance of *F2on* computed over all 10 vowels for each stop should be small compared to the same measure applied to more highly coarticulated CV's. The second measure is perhaps more pertinent to the assumptions implicit in the locus equation technique, in that it involves the frequency difference between *F2on* and *F2t*. Here the mean *absolute* frequency difference (in Hertz) between *F2on* and *F2t* for each stop is computed over all target vowels. In this paper, "degree of coarticulation" is taken to mean, in the articulatory domain, the extent to which lingual-palatal contact is similar at *F2on* and *F2t*.. For low coarticulation, lingual-palatal contact at *F2on* and *F2t* is different, so we should expect the absolute difference between *F2on* and *F2t* to be large when there is less coarticulation. In other words, the absolute frequency difference should be inversely proportional to the degree of coarticulation.

Also introduced is an EPG measure of coarticulation that is based on lingual-palatal contact patterns at the times when *F2on* and *F2t* were measured. The acoustic coarticulation measures are verified by comparing their results with the EPG coarticulation measure.

Method

Speech Materials

The stop consonants, /b,d,g/, and the vowels, /i, I, eI, ɛ, æ, a, o^{w}, ʌ, ɔ, u/, were the same as those used in Sussman, *et al* (1991). Syllable structure was CV.

Subject

The subject was an adult female native English speaker with training in phonetics.

Procedure

Five tokens of each CV were recorded for each condition. In the **High** coarticulation condition (hereafter, the "High condition"), a sustained target vowel was produced and interrupted by articulation of the target stop consonant. The vowel was sustained over approximately 3 s during which five stop consonants were produced. In the **Medium** coarticulation condition, CV's were produced within the carrier phrase "Say _____ again." This carrier phrase is used often in locus equation studies, so the derived locus parameters may be compared validly with those from other studies employing the same carrier phrase. In the **Low** coarticulation condition, the goal was to minimize the influence of V on the preceding C. This was accomplished by producing the target C and V in the syllable /íCjV/, with C palatalized and stress on the initial vowel. This manipulation should have helped to block the coarticulatory influence of the target V. That is, producing stressed /i/ before the target C, and palatalizing C offers good control over lingual position for C, in that lingual-palatal contact for a given C should be similar over all target vowels.

Acoustic Measurements

Acoustic measurements were made from spectrograms produced using the Kay Computer Speech Lab (CSL) system. Following Sussman, *et al.,* (1991), *F2on* was measured at the first glottal pulse after closure release. *F2t* was measured at the vowel midpoint when the vocalic segment formant structure was relatively steady (unchanging). In some tokens, F2 was u-shaped, and in others it was shaped like an inverted u. In these cases, *F2t* was measured at the point at which F2 attained, respectively, a minimum or maximum value.

Articulatory (EPG) Measurements

To verify that the experimental manipulation resulted in different degrees of coarticulation, EPG was used to measure lingual-palatal contact at the points where *F2on* and *F2t* were measured. EPG measurements were made using the Kay Palatometer system (# 6300) with a 100 Hz sampling rate. This system employs a pseudo palate, custom made for each speaker, embedded with 96 electrodes. When the tongue touches an electrode in the pseudo palate with pressure exceeding some threshold value, the electrode is activated. Degree of coarticulation was indexed by comparing the pattern of contact at the point when *F2on* was measured with the pattern at the point when *F2t* was measured. A poor match between contact patterns at *F2on* and at *F2t* was taken to imply minimal coarticulation.

Results and Discussion

I. *Acoustic Measures of Coarticulation*

i. Variance of *F2on*

Coarticulation was indexed for each stop according to dispersion (variance) of the distribution of *F2on* values over the target vowel set. Relatively small variance implies less dependence of *F2on* on the following vowel (*F2t*), hence less coarticulation. Means and standard deviations (SD) of *F2on* for each stop and each coarticulation condition are given (in Hertz) in Table 1. For the Low condition, standard deviations are substantially smaller than for the other conditions. For all three stops, standard deviations for the Medium condition are smaller than those for the High condition.

Stop	Coarticulation Degree	mean *F2on*	SD *F2on*
"b"	Low	2537	59
	Medium	1729	444
	High	1654	552
"d"	Low	2429	79
	Medium	2058	165
	High	1963	243
"g"	Low	2515	61
	Medium	2082	470
	High	1915	559

Table 1. Means and standard deviations (SD) of *F2on* for each stop (/b,d,g/), computed over all 10 vowels, as a function of degree of coarticulation.

ii. Absolute difference between *F2on* and *F2t*

The second acoustic coarticulation measure is based on the reasoning that, if degree of coarticulation is reflected in differences between *F2on* and *F2t*, then the absolute difference between *F2on* and *F2t* (i.e., |*F2on* - *F2t*|) should be smaller when there is more coarticulation and larger when there is less coarticulation. To test this idea, the mean absolute difference between *F2on* and *F2t* was computed over all 10 vowels for each stop in each condition. Figure 1 shows the mean absolute difference (in Hertz) between *F2on* and *F2t* for each stop as a function of degree of coarticulation.

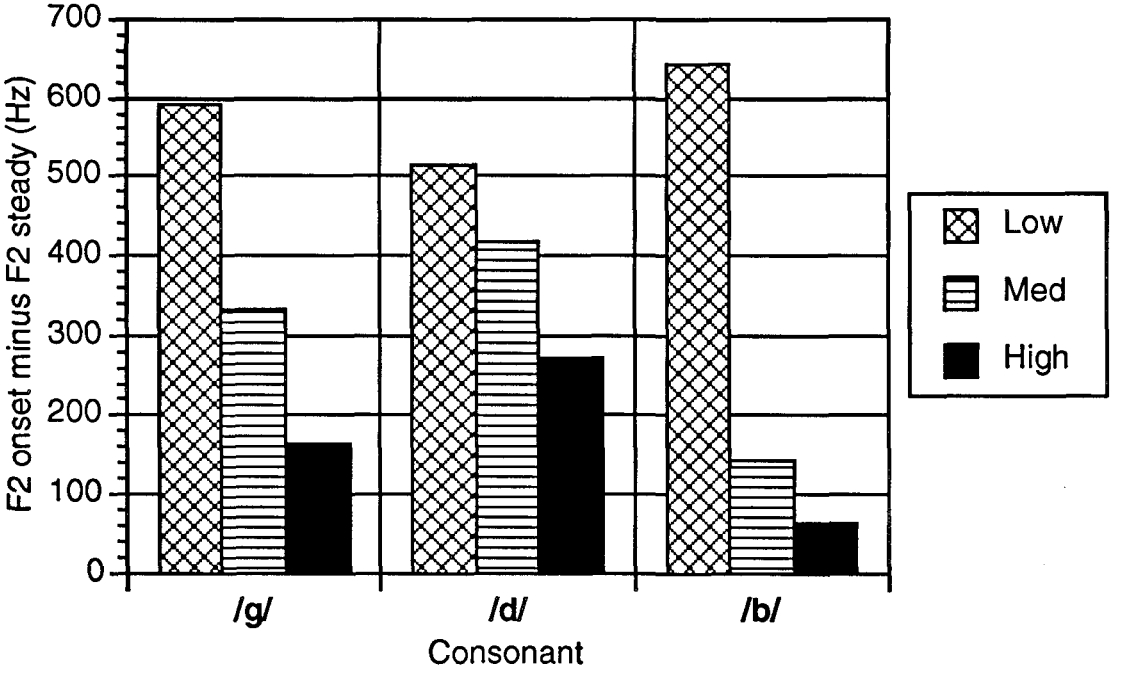


Figure 1. Mean absolute difference between *F2on* and *F2t* (i.e., |*F2on* - *F2t*|) for each stop (/b,d,g/) computed over all 10 vowels, as a function of degree of coarticulation (Low, Medium, or High). Larger differences imply less coarticulation.

As predicted, for /b/, /d/, and /g/, the mean absolute difference between *F2on* and *F2t* is largest for the Low condition, intermediate for the Medium condition, and smallest for the High condition. For each condition, the difference between *F2on* and *F2t* appears to differ for the stops. With the exception of the Low condition, the difference for /d/ appears

to be largest, suggesting that there is less coarticulation overall for /d/ as compared to /b/ and /g/.

The coarticulation measures proposed in (i) and (ii) both suggest that the manipulations employed to induce different degrees of coarticulation worked as intended. It is desirable to have independent, articulatory evidence that there were different degrees of coarticulation in the three conditions.

II *Articulatory Measures of Coarticulation*

Poorness of match of lingual-palatal contact at *F2on* and *F2t* was the articulatory measure of coarticulation. First, templates of lingual-palatal contact at *F2t* and *F2on* were produced, for each token in each coarticulation condition. The templates essentially are a record of the state ("on" or "off") of each electrode in the pseudo palate at *F2t* and *F2on*. A "poorness of match" score was computed by comparing the pattern of electrode activity at *F2t* and *F2on*. If a particular electrode was either "on" or "off" in *both* templates (call this a "match"), then a zero was scored. If a particular electrode was "on" in one template, but "off" in the other (call this a "mismatch"), then a one was scored. Overall poorness of match for each token, P, was the total number mismatches computed over all of the potentially matching electrodes in the pseudo palate (96):

$$P = \sum_{i=1}^{96} X_i, \quad X_i = \begin{cases} 0 & \text{if electrode i is on or off in both templates} \\ 1 & \text{otherwise} \end{cases},$$

where i indexes the pseudo palate electrodes, $1 \leq i \leq 96$.

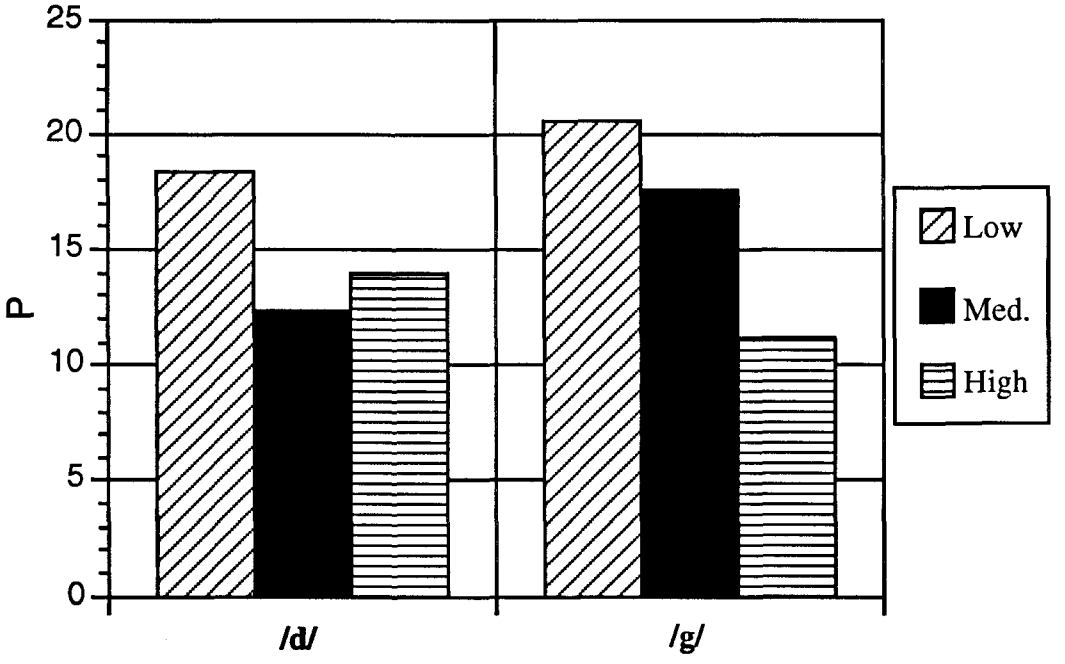


Figure 2. Mean number of non-matching electrodes (P) at *F2t* and *F2on* for /d/ and /g/ tokens (front vowel context). Larger P values imply less coarticulation.

A relatively large P shows that lingual-palatal contact is dissimilar at *F2t* and *F2on*, suggesting that coarticulation is relatively low. EPG electrodes do not extend far enough

back on the palate to record contact for back vowels, so only front vowels were considered. For most of the vowels, lingual-palatal contact for /b/ was not measurable at *F2on*. On these grounds, the coarticulation measures proposed above might be misleading for /b/, and therefore were not applied to /b/. Figures 2 shows the number of mismatches, P, in /d/ and /g/ tokens, averaged over the five front vowels, for each condition. P tended to be larger in the Low condition than in the higher coarticulation conditions. However, for /d/, P was larger in the High condition than in the Low condition. This was due to the very large P for /dæ/, which greatly elevated P when averaged over the five vowels. If /dæ/ is not included in the average, P is smaller in the High condition than in the Medium condition. The acoustic measures of coarticulation proposed in (i) and (ii) are consistent with the articulatory data -- the manipulations induced different degrees of coarticulation.

III. Modeling the *F2t* and *F2on* Data with Locus Equations

Using locus equations to model place of articulation and coarticulation involves the assumption that place and coarticulation information is conveyed in the relationship between *F2t* and *F2on*, and the acoustic and articulatory data from Experiment 1 are consistent with this assumption. Based on earlier findings (e.g., Krull, 1987), we should expect to find that, for a given stop, slope increases with increasing coarticulation. Figure 3 shows the locus equations (Equation 1) that give the best fits (according to the least squares criterion) to the *F2t* and *F2on* data for the target stops /b/, /d/, and /g/ for the Medium condition, and Table 2 gives the same information for all three conditions.

Coarticulation		**/b/**	**/d/**	**/g/**
Low	*c*	2418	2204.9	2344.5
	k	[0.062]	[0.116]	[0.088]
Medium	*c*	438.8	1609.0	627.3
	k	[0.77]	[0.258]	[0.823]
High	*c*	-59.5	1125.6	191.6
	k	[1.00]	[0.48]	[0.98]

Table 2. Intercept (*c*, in Hertz) and slope (*k*) values computed using Equation 1 for /b/, /d/, and /g/ in the Low, Medium, and High conditions. (Slope values are in brackets).

As predicted, for all three stops, the slope, *k*, increases with increasing coarticulation. Thus, *k* may be used as a coarticulation index for a given speaker for a given stop, although at this point, it is not clear whether such an index could be used for inter-speaker comparisons.

At this point, it should be noted that, for the Medium condition, the slope is greater for /g/ than for /b/, which is consistent with locus slopes derived for the native Swedish speaker in Lindblom (1963), and for the Canadian English speakers in Nearey and Shammass (1987). However, this ordering occurred for only two of the subjects in Sussman, et al. (1991). Apparently, there may be language-based differences as well as individual differences in the slopes for /g/ and /b/.

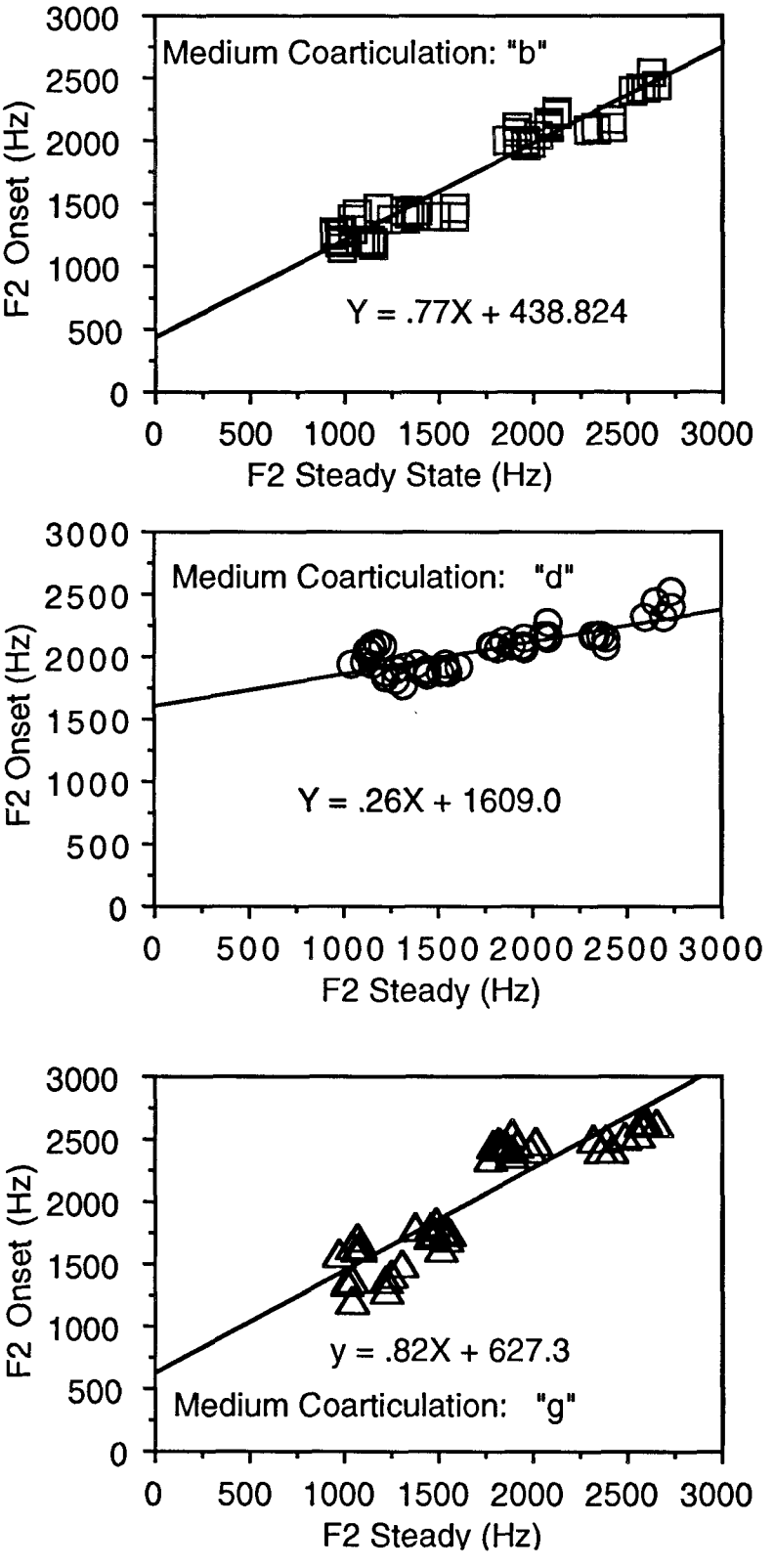


Figure 3. Locus equation plots for /b/ (top panel), /d/ (middle panel), and /g/ (bottom panel) in the Medium condition. *F2on* is on the y-axis, and *F2t* (F2 Steady) on the x-axis.

The intercept, *c*, decreases with increasing coarticulation. As noted in the Introduction, Sussman, *et al.* (1991) reported a slope-intercept correlation of -.94, and Fowler (1994) reported a somewhat smaller correlation of -.90. The correlation between parameters in Table 2 is -.99. The correlation obtained in the current study must be treated with some caution, however, as it is based on only nine data points. Because *k* and *c* are highly correlated (-0.99), the parameters of Equation 1 do not dissociate the effects of place of articulation and coarticulation, even when manner of articulation is held constant. That is, both parameters change as a function of place of articulation and coarticulation.

For Sussman's neurobiological model, perhaps what is most at stake is not whether locus equations can disentangle effects of place of articulation and coarticulation, or whether they convey place directly or indirectly. Rather, the most relevant question may be whether locus equations can represent consonants (/b/, /d/, and /g/, in this case) adequately when they are produced under a wider variety of speaking conditions, e.g., different degrees of coarticulation. Sussman *et al.* (1991) showed that, with degree of coarticulation effectively held constant by using one speaking condition, slope and intercept parameters together can distinguish the stops /b,d,g/. Further, they analyzed slope and intercept parameters separately and found that either parameter does a fair job of distinguishing the stops. This finding is not surprising, given the high correlation between slope and intercept. In the current study, then, it is of interest to test the ability of locus equations to resolve /b/, /d/, and /g/ when coarticulation is a variable. As a graphical aid to see how well slope and intercept distinguish these stops, Figure 4 depicts the locus equation parameters derived from the data in Experiment 1 in "slope-intercept" space.

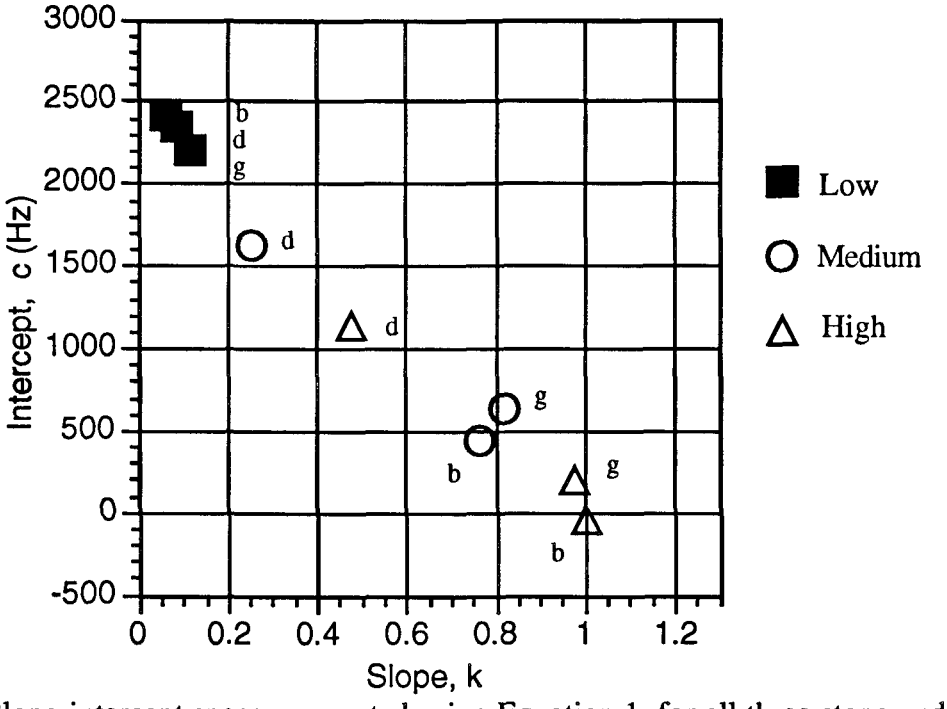


Figure 4. Slope-intercept space, computed using Equation 1, for all three stops under all three degrees of coarticulation. Slope is given on the x-axis, and intercept (in Hz) on the y-axis.

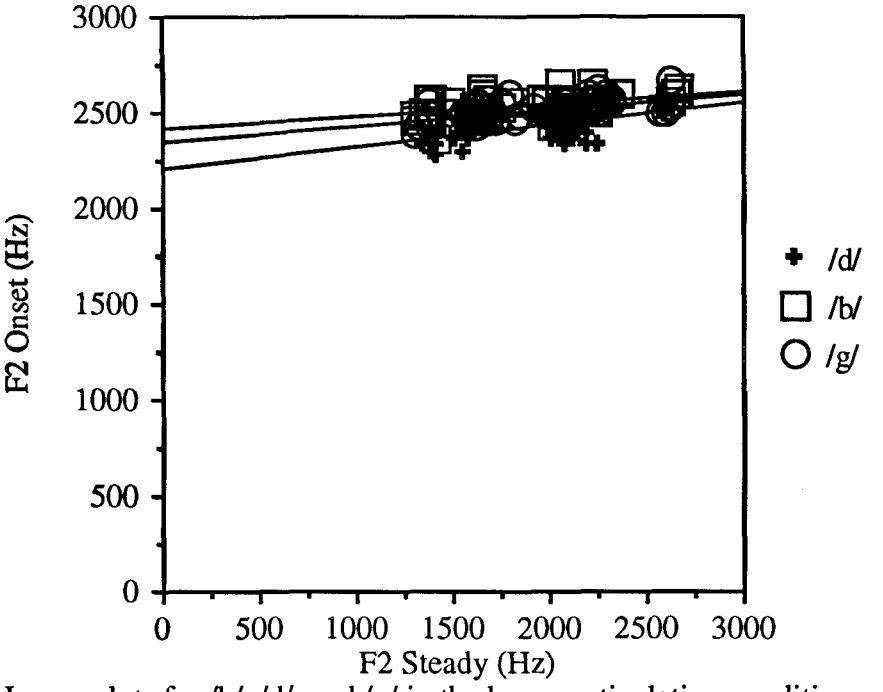


Figure 5. Locus plots for /b/, /d/, and /g/ in the low coarticulation condition.

Points in the space corresponding to **/b/**, **/d/**, and **/g/** in the Low condition are very close, suggesting that locus equations may not distinguish these stops adequately for the speaker in Experiment 1. That some tokens fall close to more than one locus line can be seen clearly in Figure 5, which shows the locus lines together with *F2on*, *F2t* data points, for /b/, /d/, and /g/ in the Low condition on one plot.

From Figure 4, the points in slope-intercept space corresponding to **/g/** and **/b/** are quite close for the Medium and also the High conditions, but those for **/d/** in the same conditions are distant from the corresponding points for **/b/** and **/g/**. Thus /b/ and /b/ are not well distinguished by Equation 1. Were /b/ and /g/ to be identified solely on the basis of locus equations derived for the speaker in Experiment 1, then it is likely that some of her tokens would be misidentified.

Interpreting the intercept, c

There is not an obvious, concrete interpretation of *c*, the intercept, which is in units of Hertz. Of course, to use locus equations as Sussman and colleagues have chosen to use them, an interpretation is not necessary. Intuitively, though, one might suppose that the intercept represents the classic Haskins locus concept as proposed by Delattre, Liberman, and Cooper (1955). In that view, an F2 locus corresponds to a "relatively fixed place of production of the consonant" (p. 283). For each stop, F2 transitions are thought to "point to" a virtual locus, and /g/ may have two loci, one for palatal and one for velar vowel contexts. In simple linear regression, often the intercept is interpreted as the value assumed by the dependent variable (*F2on*, in this case) when not under the influence of the independent variable (*F2t*). This linear regression interpretation of *c* seems to correspond to the context free, Haskins virtual locus concept. This interpretation is unreasonable, though, because the intercept estimates for Equation 1 sometimes are impossibly low (even negative) to be plausible F2 locus values. For example, the intercepts for /b/ and /g/ in the High condition were -59.5 and 191.6 Hz, respectively.

Sussman, *et al.* (1991) attempted a different approach to reconcile their locus concept with the Haskins locus concept. Specifically, they claimed that the point of intersection of Equation 1 with the 45° line, *F2on* = *F2t*, is the Haskins locus. They tested this idea by computing the intersection point of each derived locus lines with the 45° diagonal (slope = 1) for each subject and each stop (data for /g/ were divided into velar and palatal allophones). The locus values computed in this manner were reasonably consistent with the Haskins locus values for /d/ tokens, but less so for /g/ and /b/ tokens. Furthermore, some of the locus values obtained in this way were not realistic values for F2 (Sussman, *et al.*, 1991).

In the next section, Klatt's (1987) locus equation is used to model the *F2on, F2t* data to determine if it will: (a) differentiate /b/, /d/, and /g/ when coarticulation is a variable; (b) generate parameters that dissociate place of articulation and coarticulation; and (c) produce *F2L* estimates that are plausible as F2 locus values. Klatt's equation and Equation 1 are compared in terms of their ability to meet the goals in (a), (b), and (c).

IV. Modeling the *F2on, F2t* Data with Klatt's Locus Equation

In the context of speech synthesis by rule, Klatt (1987) proposes a locus equation that is a slight variation on Equation 1 to fit the *F2on, F2t* data points:

$$F2on = k*(F2t - F2L) + F2L = k*F2t + (1-k)*F2L \qquad (2)$$

where *F2L* is the F2 "locus frequency", and *k* "the degree of coarticulation"(p.754). *F2t* and *F2on* are defined as they were for Equation 1. Klatt suggested that a modified locus concept is manifested by the fact that *F2on, F2t* data points plot on a straight line. This being the case, a look-up table of *F2on, F2t* values for every CV would not be necessary for synthesis -- *F2on* could be predicted from *F2t* using an equation of the form of Equation 2 (or Equation 1).

Equations 1 and 2 describe the same line, and thus have the same slope parameter. The second parameter in Klatt's formula, *F2L*, can be expressed in terms of the formula that Sussman and colleagues use by subtracting Equation 1 from Equation 2:

$$[F2L + k*(F2t - F2L)] - [c + k*(F2t)] = 0.$$

and solving for F2L,

$$F2L = \frac{c}{1-k}. \qquad (3)$$

It turns out that *F2L* is the point that Sussman, *et al.* (1991) asserted to be equivalent to the Haskins F2 locus. That is, *F2L* is the point along the locus line at which *F2on* = *F2t*.

To determine whether Klatt's (1987) equation will (a) differentiate /b/, /d/, and /g/ when coarticulation is a variable; (b) generate parameters that dissociate place of articulation and coarticulation; and (c) produce *F2L* estimates that are plausible as F2 locus values, Equation 2 was fit, using the least squares criterion, to data gathered in Experiment 1. The estimated parameters are given Table 3. Consistent with the Haskins locus values computed in Sussman *et al.* (1991), some of the *F2L* estimates in Table 3 do not match the original Haskins values. The *F2L* estimate for /g/ in the Medium condition is greater than 3500 Hz, and estimates for /g/ (9276 Hz; SE = 12815) and /b/ (4091 Hz; SE = 3195) in the High condition are not realistic F2 values. Standard errors in the later two cases were extremely large, exceeding 75% of the estimates. There is one obvious explanation for some of the F2L estimates having high standard errors and being unrealistic as F2 values. Recall that

Coarticulation		/b/	/d/	/g/
Low	*F2L*	2578.6	2493.8	2571.8
	k	[0.062]	[0.116]	[0.088]
Medium	*F2L*	1883.4	2168.0	3542.7
	k	[0.77]	[0.258]	[0.823]
High	*F2L*	4091.2*	2169.1	9276.0*
	k	[1.00]	[0.48]	[0.98]

Table 3. *F2L* values (in Hertz) and slope (*k*) values computed using Equation 2 for /b/, /d/, and /g/ under Low, Medium, and High degrees of coarticulation. (Slope in brackets). * indicates that the standard error of the estimate was very large.

F2L represents the point where the locus line intersects the 45° line (slope = 1). When *F2on* and *F2t* are equal, the locus line slope is one (i.e., $k = 1$), so the locus line and the 45° line are parallel. In such cases, *F2L* would be (theoretically) undefined, hence the high standard error and unrealistic *F2L* estimate (and see Equation 3 for $k = 1$).

Interpreting locus equations for the low coarticulation condition

The problem of different stops sharing similar locus equations was striking for the Low condition. /b/, /d/, and /g/ were palatalized, so place of articulation was more similar for these stops in this condition than in the other conditions. Given that place was similar and coarticulation was minimal, one should expect the corresponding locus equations to be similar. Because each stop was coarticulated with stressed /i/, a front vowel, it is possible that, were the acoustic data divided into front and back vowel contexts, locus equations fit to back vowel data may be less similar than those fit to front vowel data. For the back vowel context, place of articulation was similar, but the coarticulatory influence of the target vowels may have been different than was the case for the front vowel context. Testing this hypothesis would involve computing separate locus equations for front and back vowel contexts.

Dividing the stops into front and back vowel contexts

It is possible that dividing the acoustic data into front and back vowel contexts would help to (a) better differentiate /b/, /d/, and /g/ when coarticulation is a variable; (b) generate parameters that dissociate place of articulation and coarticulation; and (c) produce *F2L* estimates that are plausible as F2 locus values. For synthesis of CV syllables, Klatt (1987) recommended that data be divided into three categories, according to the frontness and roundness of the vowels. However, Sussman *et al.* (1991) provides two arguments against such a division. First, when divided in this manner, slope values obtained in their study became statistically indistinguishable for velar /g/ vs. /b/, and also for palatal /g/ vs. /d/ (but intercept estimates remained statistically distinct). Second, the standard error of the locus equation for velar /g/ was somewhat larger than that computed for /g/ when velar and palatal /g/ were grouped together. Fowler (1994) argues against a front/back grouping on the grounds that grouping does not improve R^2 values, and, further, that the evidence from perception of velar and palatal allophones does not warrant such a grouping.

On the other hand, the raw acoustic data suggest that a front/back grouping may be in order. Figure 6 is a frequency histogram for *F2on* values for tokens in the Medium condition in Experiment 1. The distributions for /b/ and /g/ are bimodal, one mode

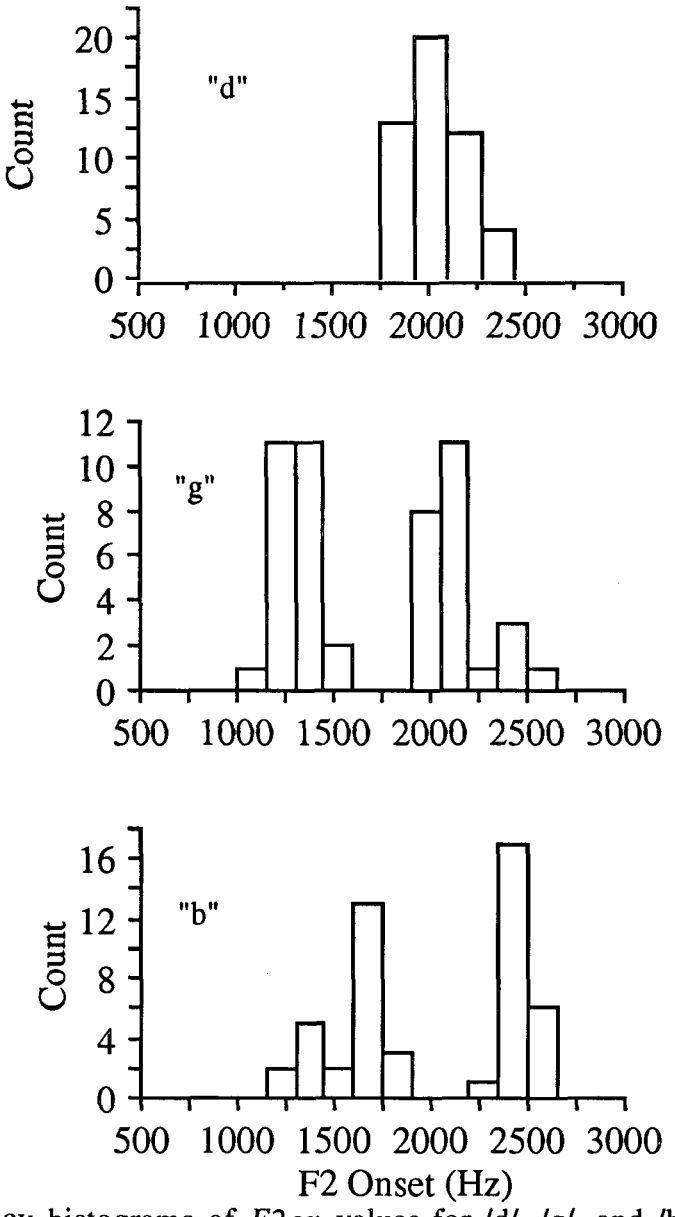


Figure 6. Frequency histograms of *F2on* values for /d/, /g/, and /b/ in the Medium condition.

corresponding to the front vowel context, and the other corresponding to the back vowel context, while the distribution for /d/ is unimodal. Given the bimodal nature of the *F2on* values for /b/ and /g/, it is not clear that we should expect a single F2 locus (or a single locus equation) that covers the front and back vowel contexts. It seems more reasonable to expect two loci, one for each vowel context, suggesting, in turn, that there should be one locus equation for each context for /b/ and /g/, though perhaps only one equation for /d/.

Spectral characteristics of release bursts for /b,d,g/ also support a front/back vowel context division. Klatt (1987) examined burst spectra for /b,d,g/ grouped into [+ FRONT], [+ ROUND], and [- FRONT, -ROUND] vowel contexts, and found that within-group similarities were stronger than between-group similarities. That burst spectra and F2 transitions may be the most salient place cues, and that both cues pattern together when

grouped into [+ FRONT], and [- FRONT, -ROUND] vowel contexts, would seem to call for separate locus equations for /b/, /d/, and /g/ in these contexts.

A perceptual study of synthesized speech could help to determine whether a dichotomous, front/back grouping is justified. One could synthesize CV syllables based on locus equations computed for ungrouped data, and also for grouped data. The intelligibility or quality of the synthesized tokens could be tested to determine which synthesis strategy yielded better results. There is indirect evidence that the front/back grouping is justified. Klatt (1979b, as cited in Klatt, 1987) found that recognition accuracy was 95% when CV's were synthesized according to locus lines derived using the [+FRONT], [+ROUND], and [-FRONT, -ROUND] groupings. This level of accuracy is impressive, but, unfortunately, recognition accuracy for tokens synthesized on the basis of locus equations for ungrouped data are not given for comparison.

To determine whether a dichotomous vowel context grouping would lead to improved locus equation results (in terms of (a), (b), and (c) above), the data from Experiment 1 were divided into Front /i, I, eI, ɛ, æ/ and Back Unrounded /a, ʌ, ɔ/ categories. Equation 2 was not fit to the final category, Back Rounded /o^{W}, u/, because the derived locus equations, based on only two vowels, might be misleading. Table 4 shows the *k* and *F2L* estimates for each stop and each degree of coarticulation. The estimates in Table 4 are plotted in slope-intercept space in Figure 7.

Coarticulation				
(Back Vowels)		**/b/**	**/d/**	**/g/**
Low	*F2L*	2595.0	2906.6	2707.6
	k	[0.074]	[0.38]	[0.18]
Medium	*F2L*	1380.8	1955.3	636.3
	k	[0.51]	[0.146]	[1.26]
High	*F2L*	190.3	1788.9	406.7
	k	[0.94]	[0.168]	[1.18]
(Front Vowels)		**/b/**	**/d/**	**/g/**
Low	*F2L*	2588.6	2499.8	2576.1
	k	[0.121]	[0.206]	[0.107]
Medium	*F2L*	2099.2	2179.3	2533.7
	k	[0.54]	[0.337]	[0.168]
High	*F2L*	2125.8	2168.0	2567.2
	k	[0.814]	[0.589]	[0.399]

Table 4. Slopes (*k*) and *F2L* derived using Klatt's (1987) locus equation (Equation 2) when vowel context is divided into Front and Back Unrounded vowel categories.

Front Vowel Context

For the Front vowel context (open markers), slope increases with increasing coarticulation. For /b/ and also for /d/, *F2L* values are stable over the Medium and the High conditions, but are different for the Low condition. For /g/, *F2L* values are quite stable across conditions, suggesting the existence of an F2 locus for /g/ in the front vowel context that does not depend upon coarticulation. F2L values for all three stops may be more interpretable as F2 locus values (though not necessarily as Haskins locus values) with the velar/palatal grouping. *F2L* estimates range from a minimum of 2099.2 Hz for /b/ in the

Medium condition to a maximum of 2588.6 Hz for /b/ in the Low condition, which may be a plausible range for F2 locus values. However, because *F2L* estimates for /b/ and /d/ in the High condition are similar, (2125.8 and 2168.0 Hz, respectively) it is doubtful that they are sufficiently contrastive to represent /b/ and /d/.

Back Unrounded Vowel Context

For the Back Unrounded vowel context (filled markers) slope increases with increasing coarticulation for /b/. However, for /d/, the slope is largest in the Low condition, and for /g/, the slope was slightly larger in the Medium condition than in the High condition. Also, *F2L* values change for all three stops as a function of coarticulation. With the exception of /b/ in the High condition (190.3 Hz), the *F2L* estimates represent conceivable F2 locus values.

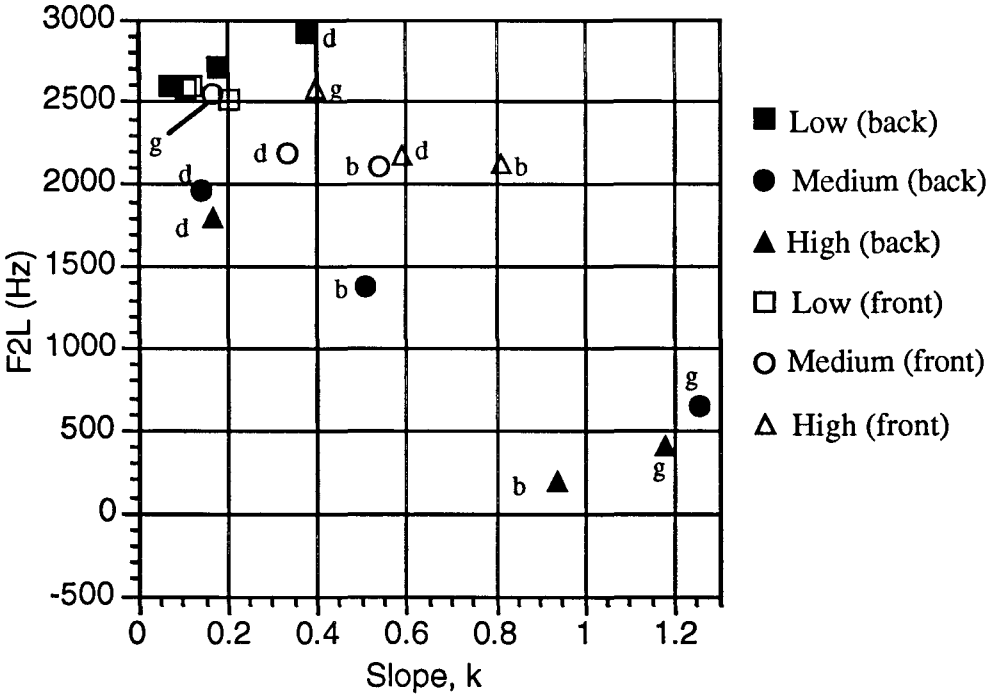


Figure 7. Slope-intercept space computed with Equation 2 for all /b/, /d/, and /g/ in each condition. /back, unrounded/ indicated by filled markers, and /front/ by open markers. Slope is given on the x-axis, and *F2L* (in Hz) on the y-axis.

For comparison, estimates of *k* and *c* (Equation 1) computed using the same Front and Back Unrounded grouping are given in Figure 8. For Equation 2 (Figure 7), with some exceptions to be discussed below, the points seem well separated. However, for front vowels, points corresponding to /b/ in the Medium condition and to /d/ in the High condition are very close. For both equations, the point for /g/ (front vowel context) in the Medium condition falls within the region in slope-intercept space densely populated with points for the Low condition, Front vowel context. The front/back vowel grouping improved separation for Equation 1 (Figure 8). For both equations, however, the points for the Low condition for /b/, /d/, and /g/ in the front vowel context remain tightly clustered the in the upper left corner. As can be seen in Figure 8, the parameters generated by Equation 1 are highly correlated (-.95). The correlation is weaker for parameters generated with Equation 2 (-.82), possibly indicating that more information could be obtained with its use.

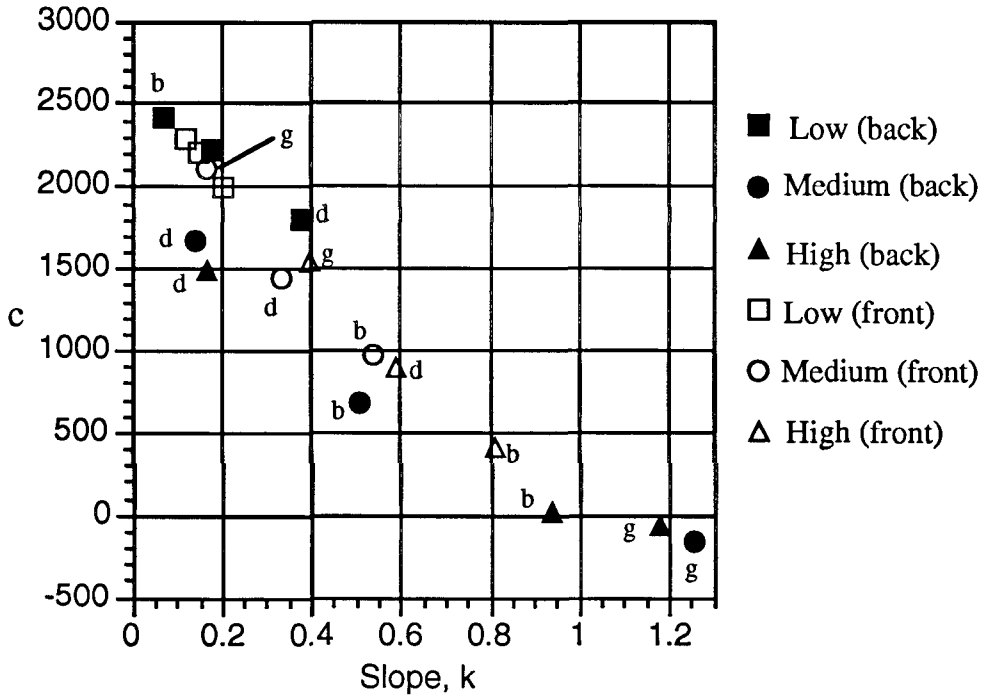


Figure 8. Slope-intercept space computed with Equation 1 for /b/, /d/, and /g/ in each condition. /back, unrounded/ indicated by filled markers, and /front/ by open markers. Slope is given on the x-axis, and intercept (in Hz) on the y-axis.

Earlier we raised the possibility that locus equations would distinguish /b/, /d/, and /g/ better in the Low condition if the data were divided into Front and Back Unrounded vowel contexts. As predicted, the points (computed using either Equation 1 or 2) corresponding to /b/, /d/, and /g/ in the front vowel context (open squares) are more similar than the points for the back vowel context (filled squares), which are well separated.

Experiment 2

It is important to determine the perceptual relevance of locus equations, particularly the relationship between similarity in physical, slope-intercept space, and similarity in perceptual space. One reasonable prediction about the mapping between physical space and perceptual space is that tokens of consonants that have points that are in close proximity in slope-intercept space (i.e., those that have similar locus equations) may be "close" in perceptual space, and therefore would be likely to be confused in perception.

When *F2on*, *F2t* data were not grouped into velar and palatal categories, Equation 1 yielded parameter estimates that did not seem to distinguish between /b/, /d/, and /g/ in the Low condition, because the locus equations derived for that condition were similar. Both equations had difficulty distinguishing the stops in the Low condition for the front vowel context. Therefore, an identification experiment based on the low coarticulation tokens would seem to constitute the best test case of the perceptual relevance of locus equations. To test the prediction that tokens of consonants that are represented in close proximity in slope-intercept space would be perceptually confusable, listeners identified /b/, /d/, and /g/ in tokens produced in the Low condition in Experiment 1.

Method

Stimuli

Two of the five tokens of each íCíV syllable produced in the Low condition of Experiment 1 were recorded from digitized speech files in random order with a 5 s ISI on audio tape using a Tascam cassette recorder. There were ten vowels, three stop consonants and two tokens of each syllable, yielding a total of 60 stimulus words on the audio tape.

Procedure

Stimuli were presented over earphones at a comfortable listening level, and subjects listened to the tape one time. They were asked to identify the consonant in each stimulus word as /b/, /d/, or /g/, and to guess if uncertain. They responded by circling the letter, "b", "d", or "g" on a response sheet.

Subjects

Five members of the UCLA Phonetics Laboratory participated as subjects in the listening experiment.

Results

60 was a perfect score. Two subjects scored 60, one scored 59, and two scored 58. Four of the five errors involved identifying a "g" as "d", or vis versa, and one involved identifying a "d" token as "b". Three errors involved front vowels, and two involved back vowels. The high scores suggest that the task was not difficult, and indicate performance ceiling effects.

GENERAL DISCUSSION

This paper examined locus equation representations of /b/, /d/, and /g/. The main goal was to determine whether locus equations differentiate /b/, /d/, and /g/ when coarticulation is an independent variable. It was also of interest to compare the resolving power of the locus equation that Sussman and colleagues use (Equation 1) with Klatt's (1987) locus equation (Equation 2). Both equations had difficulty distinguishing /b/, /d/, and /g/ in the Low condition. Overall, the best differentiation of /b/, /d/, and /g/ occurred when separate locus equations were fit to acoustic data divided into Front and Back Unrounded vowel contexts, and this was true regardless of whether Equation 1 (see Figure 8) or Equation 2 (see Figure 7) was used. For the Low condition, both equations had difficulty distinguishing the stops in the front vowel context, but did a better job of distinguishing them in the back vowel context.

Equation 2 may have differentiated /b/, /d/, and /g/ better than Equation 1. However, for the Front vowel context, the locus equation for /b/ in the Medium condition was quite similar to that for /d/ in the High condition (see Figure 8). To illustrate the consequences of the similarity of these equations, consider a situation in which *F2on* and *F2t* were the only pieces of information available for identifying a stop consonant. If a given *F2on*,*F2t* pair happened to fall close to the locus line for /b/ in the Medium condition, then it would also fall close to the line for /d/ in the High condition. Therefore, without additional information, the *F2on*,*F2t* pair may be misclassified. Generally, in cases in which two different stops that are produced with different degrees of coarticulation have similar locus equation representations, then, given only *F2on*, *F2t* information, one could identify the stop if the degree of coarticulation were known; or, one could determine the degree of coarticulation if the identity of the stop identity were known. On the other hand, some might question whether locus equations should be expected to differentiate stops across coarticulation conditions. If the scope of the application did not require the locus equation model to determine *both* stop identity *and* coarticulation, then either Equation 1 or

2 would be satisfactory if applied to data divided into Front and Back Unrounded vowel contexts.

Two previous studies reported that the parameters of Equation 1 were highly correlated. Sussman, *et al.* (1991) reported a slope-intercept correlation of -0.94, and Fowler reported a correlation of -0.9, both studies suggesting that the parameters convey redundant information. Earlier we questioned whether locus parameters would be able to dissociate place of articulation from coarticulation, when coarticulation was an independent variable. When Equation 1 was fit to the data from Experiment 1, the slope-intercept correlation was -0.99, and when the stops were divided into Front and Back Unrounded vowel contexts, the correlation was -0.95. For the parameters of Equation 2, *F2L* and *k*, the correlation was -0.82 for the divided data. Although this correlation is still quite high, it is possible that more information could be gleaned by using Equation 2 in place of Equation 1. Fowler (1994) reported that a discriminant analysis, based on the parameters of Equation 1, did a poor job of classifying seven stops (differing in place and/or manner). It is possible that classification accuracy for the seven stops would be improved if Equation 2 were used in place of Equation 1.

Were it the case that only one parameter, say, the slope, changed with coarticulation, and the other changed with stop place, then the parameters could serve as indices for inter-speaker and cross-linguistic comparisons of place of articulation and coarticulation. Unfortunately, this dissociation seemed to have occurred only for /g/ in the front vowel context. There the *F2L* parameter was stable across the three coarticulation conditions, while *k* increased with increasing coarticulation (see Table 4 and Figure 7). In this case, *F2L* could be said to represent an F2 locus (though not necessarily the Haskins locus), while *k* could represent degree of coarticulation.

Another reason for exploring the consequences of using Equation 2 was that the *F2L* parameter may have proven to be interpretable as some kind of F2 "locus". Although the *F2L* estimates in this study may be conceivable as F2 values (with the exception of /b/, back vowel context, High coarticulation) they probably can not represent loci for /b/ and /d/: *F2L* for /b/ and /d/ (front vowel context) in the High condition differed by only about 38 Hz. When both *k* (slope) and *F2L* are considered together, these stops are differentiated well (see Table 4). Parameter interpretability is probably not an issue if one wishes to apply locus equations in the manner of Sussman and colleagues.

Experiment 2 was intended to assess the "psychological reality" of locus equations. Earlier we predicted that tokens of different consonants that have similar locus equations may be perceptually similar, and thus confusable (see Sussman, *et al.*, 1991, for similar predictions). Although locus equations were similar for /b/, /d/, and /g/ in the front vowel context for the Low condition, subjects identified the tokens accurately. There are potential difficulties involved with interpreting these findings, however. Sussman, *et al.* (1991) point out that the F2 transition is not the only source of place information available to the listener. In fact, they outline a "brain-based" model that operates on release burst information as well as on *F2on*, *F2t* information, and Sussman (1991) proposes that inclusion of F3 information may further enhance stop identification. In Experiment 2, all stop closures were released in the stimulus words, so it is likely that, even if F2 information was ambiguous, subjects were able to identify the tokens on the basis of the release bursts. One might argue that, for cases in which the locus information is ambiguous, listeners should be expected to rely more heavily on release burst characteristics to identify stop consonants. This may be the case for CV syllables, but then what about the role of locus equations for stop consonants occurring in syllable final position? Crystal and House (1982, 1988) and Byrd (1993) report that often syllable final stops are not released. Consequently, listeners can not always rely on the burst to identify stops. This being the case, locus equations might be expected to play an even larger role for stop consonant identification in syllable final position. However, Krull (1988) reported that F2 loci in VC syllables are dependent on the target vowel to the extent that locus equations

did not differ substantially for /b/ and /d/. Consequently, locus equation representations may be of little use for stop identification in VC syllables.

Given that listeners probably use information in addition to *F2on* and *F2t* for stop identification, perhaps locus equation predictions should be made in terms of perceptual similarity, rather than absolute identification. For example, although the tokens in Experiment 2 were identified accurately, it is possible that, in a similarity scaling experiment, these tokens would be rated as more similar than tokens of stops that had very different locus equation representations. Alternatively, a categorization paradigm like Experiment 2 could be made more informative if the release burst cues were neutralized (removed), forcing the listener to rely more on F2 transition information for stop identification. This measure would likely eliminate the ceiling effect seen in Experiment 2. We are currently designing several experiments along these lines.

Locus equations when coarticulation is extreme

Locus equations have an impressive history of fitting data. However, we have seen that the equations may have difficulty when coarticulation (this study) or manner (Fowler, 1994) are variables. The behavior of Klatt's (1987) locus equation at extreme levels of coarticulation can be seen by analyzing Equation 2:

$$F2on = k*F2t + (1-k)*F2L \qquad (2)$$

For high coarticulation ($k = 1$), the locus parameter, *F2L*, does not contribute to prediction of *F2on*: *F2on* = *F2t*. Looking at this from a somewhat different perspective, perhaps the fact that the *F2L* parameter is undefined when $k = 1$ (see also Equation 3) is consistent with intuition: when there is very high dependence of *F2on* on *F2t*, one should not expect to find a "locus". Therefore, when there is very high coarticulation, as was the case for the VC syllables that Krull (1988) analyzed, locus equations may be less informative for stop identification.

For very low coarticulation (k = zero), *F2t* (i.e., the vowel) becomes less predictive of *F2on*, and *F2on* = *F2L*. In such cases, the concept of "locus" could be made quite tangible by defining it as the mean of the distribution of *F2on* values computed over the vowel inventory.

The relationship between locus equation slopes and F2 transition slopes

Locus equation slopes are inversely proportional to the physical F2 transition slopes: Smaller locus slopes indicate that the formant trajectories tend to be more radical (steeper), whereas larger slopes indicate flatter trajectories. However, the correspondence between locus line slopes and F2 transition slopes is exact only when k is near 1 and c is near zero. In such cases, the locus equation slope is steep and the F2 trajectory is flat. As k becomes smaller, however, the correspondence between k and the formant trajectory weakens, because locus equations do not take into account formant transition duration. Figure 9 is a schematic diagram showing three F2 transitions (A, B, C) that would be treated as equivalent (i.e., assigned the same *F2on* and *F2t* values in a typical locus analysis), even though the dynamic nature of each transition is quite different. The formant transition slope is shallowest for A, and steepest for C. Clearly, substantial dynamic

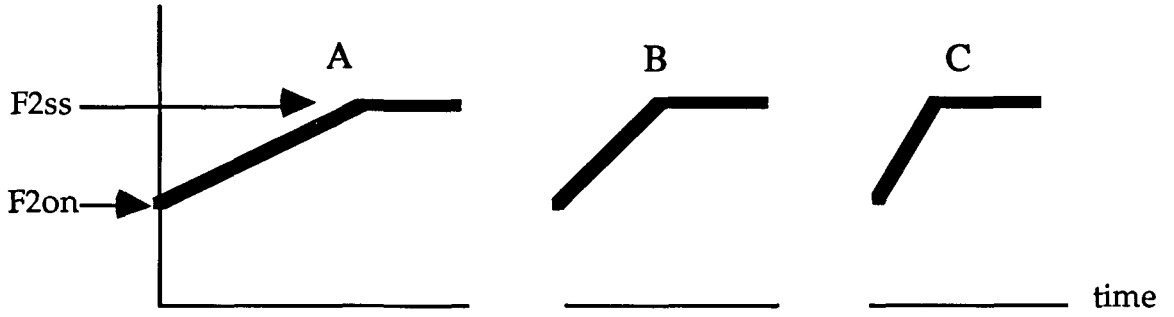


Figure 9. Three F2 transitions (A, B, and C) that would be treated as equivalent using locus techniques (Equations 1 and 2)

information would be lost were F2 transitions to be modeled using locus equations as currently formulated. This consideration suggests that additional coarticulation information (and possibly place information, as well) could be gained if the locus equation model were modified by adding a third variable, *transition duration*, or *transition velocity*. Alternatively, it is possible to include transition velocity information implicitly in Equation 1 or 2 by modeling F2 at onset (i.e., *F2on*) and F2 after a fixed time interval (say, 30 ms). Nearey and Shammass (1987) measured F2 at onset and after 60 ms, and successfully modeled F2 at these points using Equation 1. Further testing would be required to determine whether locus equations could model F2 successfully using a shorter sampling interval.

Acknowledgments

I thank Patricia Keating for invaluable discussions and input regarding this project, Cecile Fougeron and Peter Ladefoged for helpful comments on an earlier version of the manuscript, and Dani Byrd for use of her EPG template analysis software. Work on this project was supported by NIH.

References

Blumstein, S.E., & Stevens, K.N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. J. Acoust. Soc. Am., 66, 1001-1017.

Byrd, D. (1993). 54,000 American stops. UCLA Working Papers in Phonetics, 83, 97-115.

Crowther, C.S. (1994). Inferring degree of coarticulation and relational invariance of stop place in CV's using locus equations. Journal of the Acoustical Society of America, 95, 2922.

Crystal, T.H., & House, A.S. (1982). Segmental durations in connected speech: Preliminary results. Journal of the Acoustical Society of America, 72, 705-716.

Crystal, T.H., & House, A.S. (1988). Segmental durations in connected speech: Current results. Journal of the Acoustical Society of America, 83, 1553-1573.

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

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R.N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. Journal of the Acoustical Society of America., 84, 115-123.

Fowler, C.A. (1994). Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. Perception & Psychophysics, 55, 597-610.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. Journal of the Acoustical Society of America, 73, 322-335.
Klatt, D.H. (1987). Review of text-to-speech conversion for English. Journal of the Acoustical Society of America, 82, 737-793.
Knudsen, E.I., du Lac, S., & Esterly, S.D. (1987). Computational maps in the brain. Annual Review of Neuroscience, 10, 41-65.
Krull, D. (1987). Second formant locus patterns as a measure of consonant-vowel coarticulation. PERILUS, 5, 43-61.
Krull, D. (1988). Acoustic properties as predictors of perceptual responses: A study of Swedish voiced stops. PERILUS, 7, 1-149.
Lahiri, A., Gewirth, L., & Blumstein, S.E. (1984). A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. Journal of the Acoustical Society of America, 76, 391-404.
Liberman, A.M., & Mattingly, I.G. (1985). The motor theory of speech perception revised. Cognition, 21, 1-36.
Lindblom, B. (1963). On vowel reduction. The Royal Institute of Technology, Speech Transmission Laboratory, Stockholm, Sweden, 29.
Massaro, D.W., & Oden, G.C. (1980). Evaluation and integration of acoustic features in speech perception. Journal of the Acoustical Society of America, 67, 996-1013.
Nearey, T.M., and Shammass, S.E. (1987). Formant transitions as partly distinctive invariant properties in the identification of voiced stops. Canadian Acoustics, 15, 17-24.
Oden, G.C., & Massaro, D.W. (1978). Integration of featural information in speech perception. Psychological Review, 85, 172-191.
Stevens, K.N., & Blumstein, S.E. (1978). Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America, 64, 1358-1368.
Sussman, H.M. (1989). Neural coding of relational invariance in speech: Human language analogs to the barn owl. Psychological Review, 96, 631-642.
Sussman, H.M. (1991). The representation of stop consonants in three-dimensional space. Phonetica, 48, 18-31.
Sussman, H.M. (1994). Locus equations during compensatory articulation: A test of the "orderly output constraint." Journal of the Acoustical Society of America, 95, 2922.
Sussman, H.M., Hoemeke, K., & Ahmed, F. (1993). A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation. Journal of the Acoustical Society of America, 94, 1256-1268.
Sussman, H.M., McCaffrey, H.A., & Matthews, S.A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. Journal of the Acoustical Society of America, 90, 1309-1325.
Wagner, H., Takahashi, T., & Konishi, M. (1987). Representation of interaural time difference in the central nucleus of the barn owl's inferior colliculus. The Journal of Neuroscience, 7, 3105-3116.