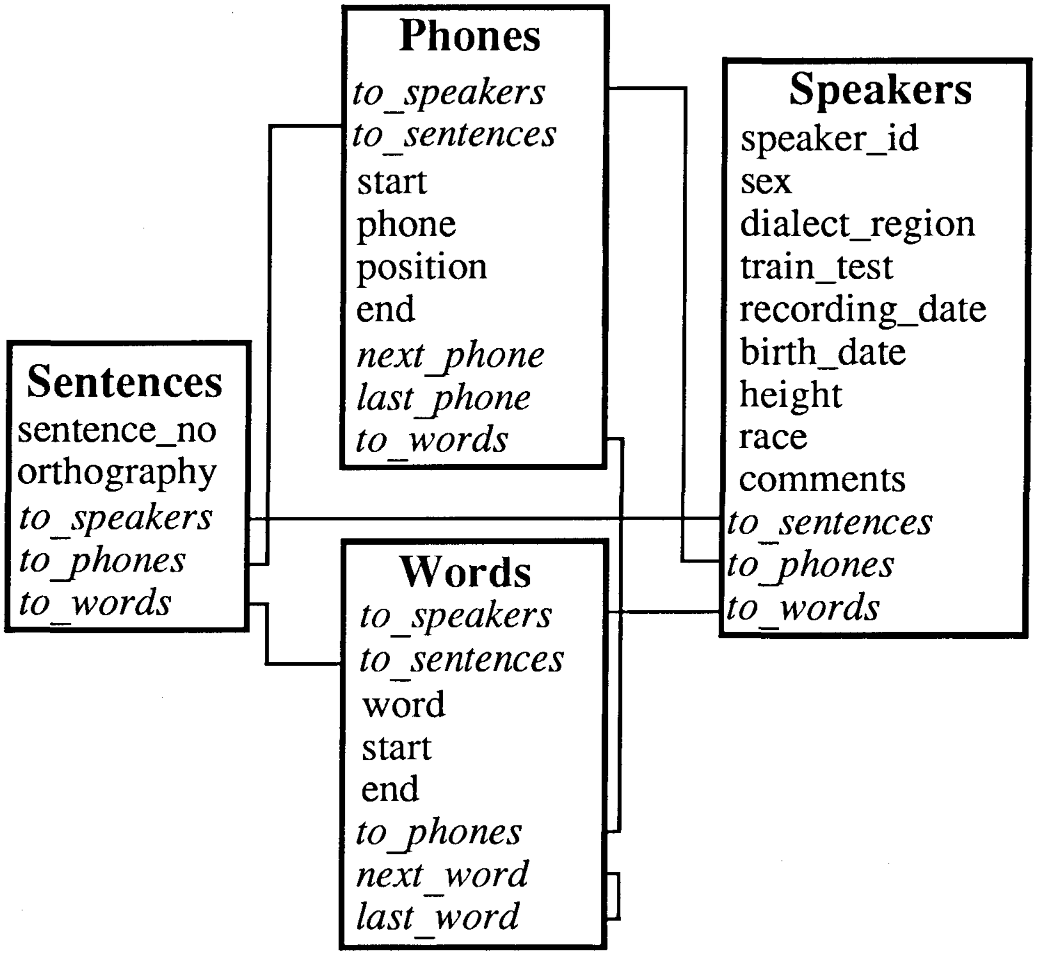
UCLA Working Papers in Phonetics



Number 81

July, 1992

The UCLA Phonetics Laboratory Group

Beatriz Amos
Victoria Anderson
Michael Inouye
Sue Banner Inouye
Barbara Blankenship
Dani Byrd
John D. Choi
Ken de Jong
Sandy Disner
Edward Flemming
Vicki Fromkin
Robert Hagiwara
Bruce Hayes
Susan Hess
Keith Johnson
Pat Keating
Karn King
Paul Kirk
Jenny Ladefoged
Peter Ladefoged
Mona Lindau
Ian Maddieson
Joyce McDonough
Pam Maelzer
Benjamin Munson
Dionne Ramey
Bonny Sands
Stephan Schuetze-Coburn
Michael Shalev
Aaron Shryock
Siniša Spajić
Donca Steriade
Henry Teheranizadeh
Kimberly Thomas
Yuichi Todaka
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 1 R01 DC00642
USHHS grant 5 T32 DC00029
NSF grant BNS 9107004
NSF grant DBS 9213604
and the UCLA Humanities Division

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

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

This issue of UCLA Working Papers in Phonetics was edited by Pat Keating and Dani Byrd.

UCLA Working Papers in Phonetics 81

July 1992

Table of Contents

I. TIMIT

II. OTHER

PHONETIC ANALYSES OF THE TIMIT CORPUS OF AMERICAN ENGLISH AT UCLA

P. Keating, B. Blankenship, D. Byrd, E. Flemming, Y. Todaka

Expanded version of paper to appear in Proceedings of the International Conference on Spoken Language Processing (ICSLP-92)

ABSTRACT

This paper serves as an introduction to others in this issue. It reports a set of studies of some phonetic characteristics of the American English represented in the TIMIT speech database. First we describe generally how we use the non-speech files on the TIMIT CD with a commercial database program. Some studies of pronunciation variation using only the segmental transcriptions and durations of TIMIT, and information about the TIMIT speakers, are then described. Next a study of velar stop variation depending on vowel context, involving acoustic analysis of selected tokens from TIMIT, is presented; it is shown that the effect of a following vowel is stronger than that of a preceding vowel on release bursts, but the two have similar effects on formant transitions. Results of such studies should be useful not only for linguistic phonetics but also for speech recognition lexicons and text-to-speech systems.

I. INTRODUCTION

Although the acquisition of acoustic phonetic knowledge was a design goal of the TIMIT project, very little such work has been reported so far. In the UCLA Phonetics Lab we have been using TIMIT to evaluate a number of claims in the phonetic, phonological, and TESL (Teaching English as a Second Language) literature. Though TIMIT contains only read speech, and has various other limitations, some of which will be discussed below, it is a valuable tool for at least some kinds of phonetic analyses.

The TIMIT speech database, developed at Texas Instruments and MIT and distributed by the National Institute of Standards and Technology on a CD, consists of 2342 sentences read by 630 native speakers of American English. It is described in [8], [10], and [14]. Three types of sentences are included. Two "calibration sentences", designed to allow dialect comparison, were read by all 630 speakers. The speakers are coded as belonging to one of 8 "dialect regions" (New England, New York City, North Midland, South Midland, Southern, Northern, Western, and "Army brat"). 450 "phonetically compact" sentences were designed to provide examples of phonemes in all possible left and right contexts. Each of these sentences was read by seven speakers. The remaining 1890 sentences are "diverse sentences" selected mostly from the Brown corpus. Each of these was read by only one speaker. Each speaker read ten sentences altogether (about 30 sec of speech) and there is a total of 6300 utterances in the database. (There was also a partial, Prototype version released earlier, and one of the studies in this paper used that version.) All sentences in TIMIT are segmented and labeled. The transcriptions are based on a combination of acoustic and auditory criteria [11] and have been rechecked for the final version of TIMIT.

Purchasers of the TIMIT CD can be surprised to find that it is not so much a "database" in the usual sense as it is a collection of related files with mnemonic names. It is not obvious how all the information on the CD can be easily accessed and used, without developing custom software. To work with the information that accompanies the speech recordings, we use a commercial relational database, Borland's Reflex Plus. This outdated program is not perfectly suited to our tasks, and perhaps another product would be preferable, but it was inexpensive and goes far in making TIMIT a useful research tool. All the non-speech files on the TIMIT CD were imported into four database files:

sentences
phones
words
speakers.

The organization of the database -- the fields used in each database file and the links between the files -- is shown in Figure 1. Each record in the "sentences" database contains the orthographic form of a sentence together with its TIMIT ID code and is linked to the speakers who spoke it. Links to the words and phones contained in the sentences are possible but require too much memory for our machine to calculate. Each record in the "phones" database contains one phonetic symbol (a phone), its start and end times (in msec) in the speech signal, and a coding (provided by us) of its position within its word. Each phone record is linked to the sentence containing it and the speaker who read it, and could be linked to the word containing it. "Next_phone" provides a link to the record of the following phone. Each record in the "words" database contains a word in an utterance and is linked to the speaker and to the phones it contains. Each record in the "speakers" database contains information about one speaker, such as their TIMIT ID code, sex, etc., is linked to the sentences uttered by that speaker, and could be linked to the phones and words uttered by that speaker.

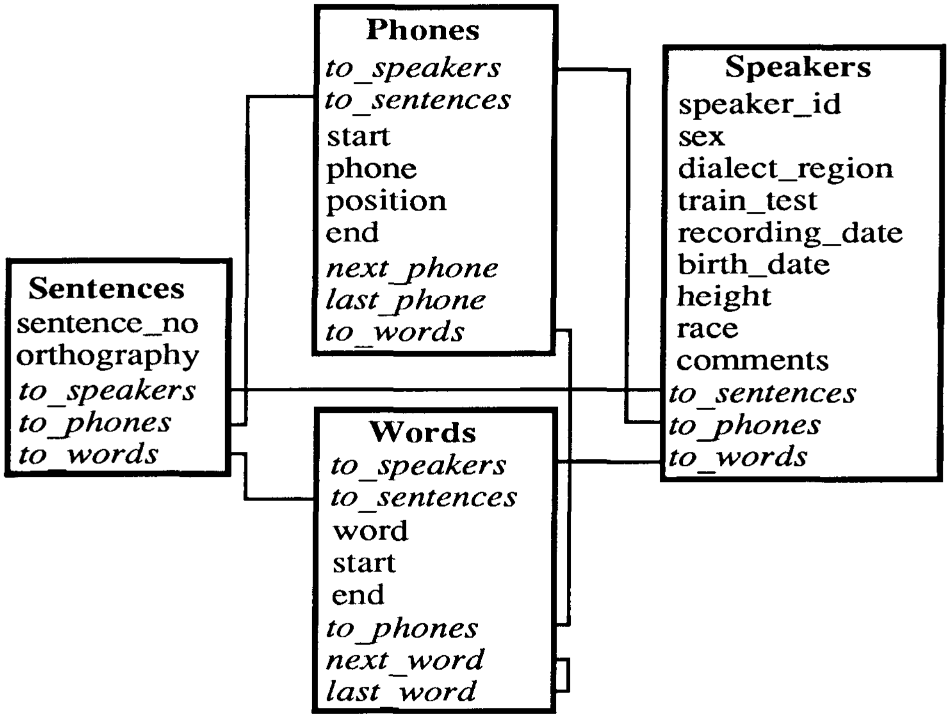


Figure 1 - Organization of TIMIT in a Reflex database.

Such a database allows us to search for tokens described by any combination of these kinds of information, such as all tokens of a particular word spoken by a given subgroup of speakers. It also allows us to search for tokens and then extract information about them. We can search for each instance of a given phone and ask to know its duration, the identity of surrounding phones, the sentence in which it occurs, or personal characteristics of the speaker who produced it. Output from Reflex searches can be exported into commercial spreadsheet, graphing, and statistical programs for analysis.

II. TRANSCRIPTION STUDIES

Because the phonetic transcriptions in TIMIT are fairly narrow and are largely acoustically-defined, many research questions can be answered using the Reflex database and these transcriptions alone, without acoustic analysis of the speech files. For example, we can look at actual pronunciations of words or phonemes over the corpus. As one small example, consider the distribution of vowels in the sequence "ing" at the ends of words, both in monosyllables like "thing" and as a suffix. Standard descriptions lead us to expect [ɪ] in the stressed cases and probably [ɨ] in the stressless cases. However, UCLA undergraduates typically use [i] (in stressless "ing", a pronunciation textbooks don't mention. Figure 2 shows the distribution of vowels in "ing" in monosyllables vs. polysyllables in TIMIT. In agreement with textbooks and UCLA undergraduates, most monosyllables are pronounced with [ɪ], and in agreement with textbooks, polysyllables are most likely to have [ɨ] (41%). At the same time, [i] is close behind at 34%, showing that the UCLA undergraduate pronunciation is not rare. Clearly TIMIT would lend itself to a great variety of studies of how particular sequences are most often pronounced by speakers overall. In some of the studies which follow, the influence of phonetic context and the speaker characteristics sex, dialect region and age are considered when such pronunciation variants are found.

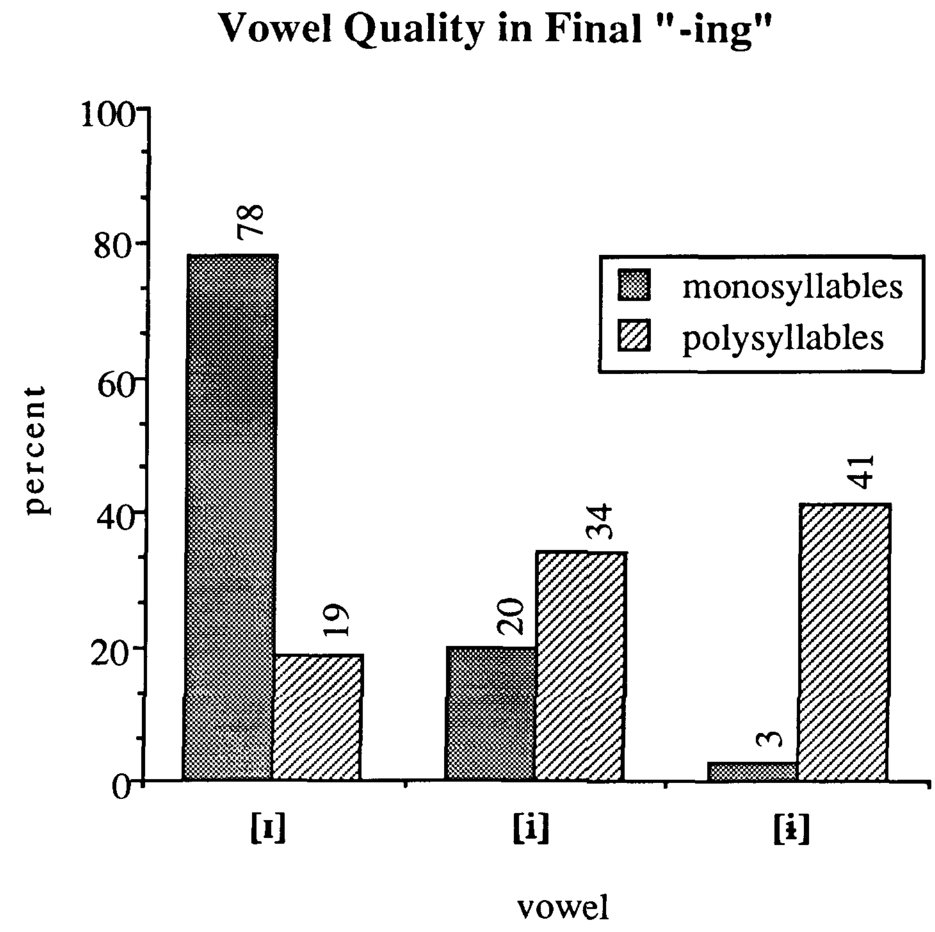


Figure 2 - Vowels in word-final "ing" in monosyllables and polysyllables.

Consider the pronunciation of the word "the", the most common word in TIMIT [3]. It occurs with a wide array of vowel qualities, including several which appear in only a very few tokens (for example, [æ]). Figure 3 shows the five most frequent vowel qualities in "the" in TIMIT; [ə] and [ɨ] are most common overall, followed by [i]. This variation is significantly influenced by the speaker's sex: women use [ɨ] and [ɪ] more than men, and [ə] and [i] less. As Figure 3 shows, the choice of vowel in "the" is also highly dependent on the first segment in the following word. Before consonants, [ə] and [ɨ] dominate, while before vowels [i] is by far the most common; this difference is significant.

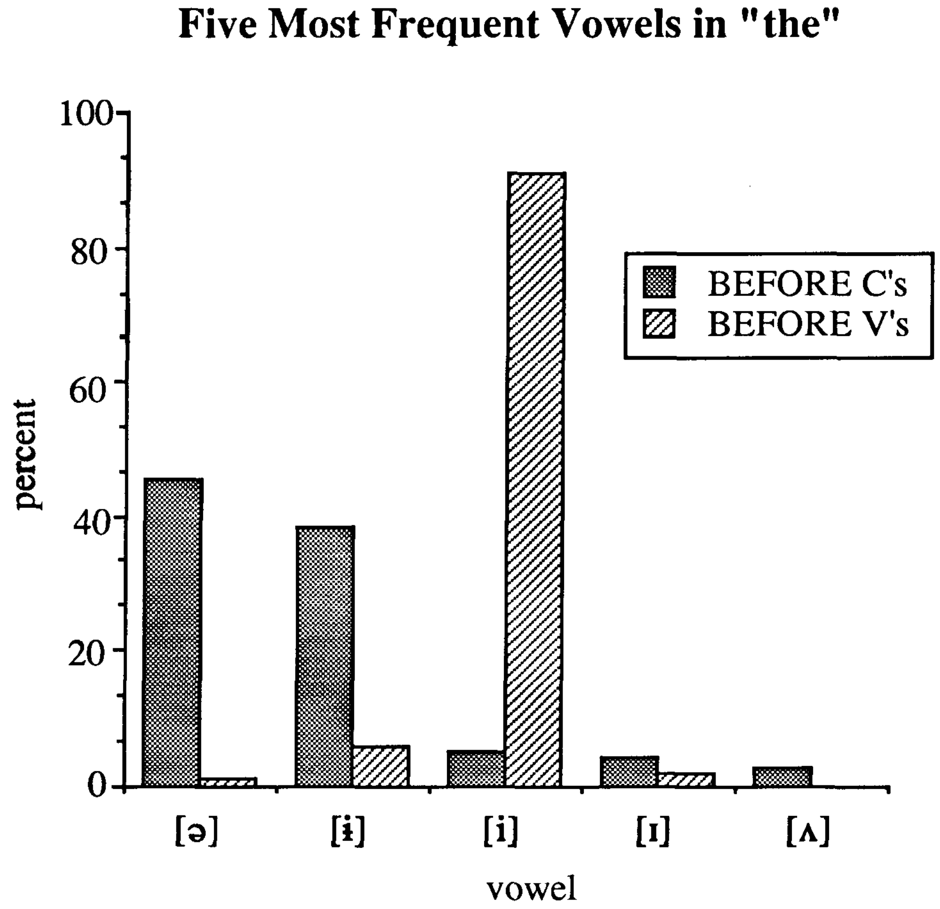


Figure 3 - Most frequent vowels in "the" before words beginning with vowels and with consonants.

Standard and normative descriptions of the pronunciation of "the" follow this pattern: they suggest that it is pronounced [ði] or [ðɪ] before a word beginning with a vowel, possibly with an intervening palatal glide, or with an intervening glottal stop before stressed vowels; and as [ðə] before a word beginning with a consonant. This difference is painstakingly taught to ESL students. However, among UCLA undergraduates the norm seems to be [ðə] before a consonant and [ðəʔ] before a vowel (see also [5]). How can [i] be by far the most common vowel in "the" before vowels in TIMIT, yet be rare among UCLA undergraduates? A striking finding is that in TIMIT the choice of vowel in "the" before a phonemic vowel is age-dependent. Figure 4 (from Todaka, this volume) shows the use of [i] vs. all other vowels, by speaker age. No one over 50 years old uses any vowel but [i] in "the" before a vowel, and while [i] remains the most common vowel even for younger speakers, other vowels occur in more than a third of the tokens for the youngest speakers. This difference is highly significant. Since some TIMIT speakers were recorded several years ago, the UCLA undergraduates are probably the next age bin down and have advanced further along in what appears to be a current change in a pronunciation norm.

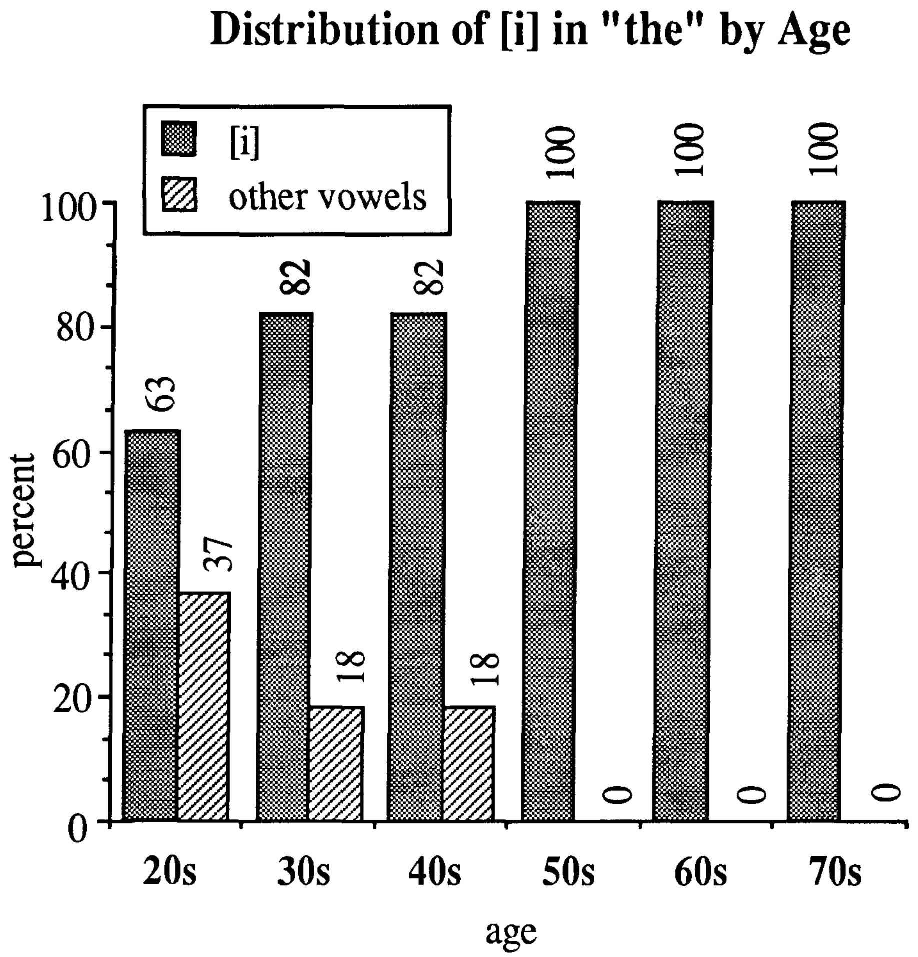


Figure 4 - Use of [i] in "the" before words beginning with vowels according to age of speaker.

Turning to the use of glottal stop after “the,” 65 of the 242 Prototype tokens of "the" before a phonemic vowel have a glottal stop between the two words. Almost all of these occur when the vowel after "the" has primary stress. Figure 5 (also from Todaka, this volume) shows that across the sample tokens of prevocalic “the” are followed by a glottal stop more often when “the” has a reduced vowel and less often when “the” contains [i] and [ɪ].

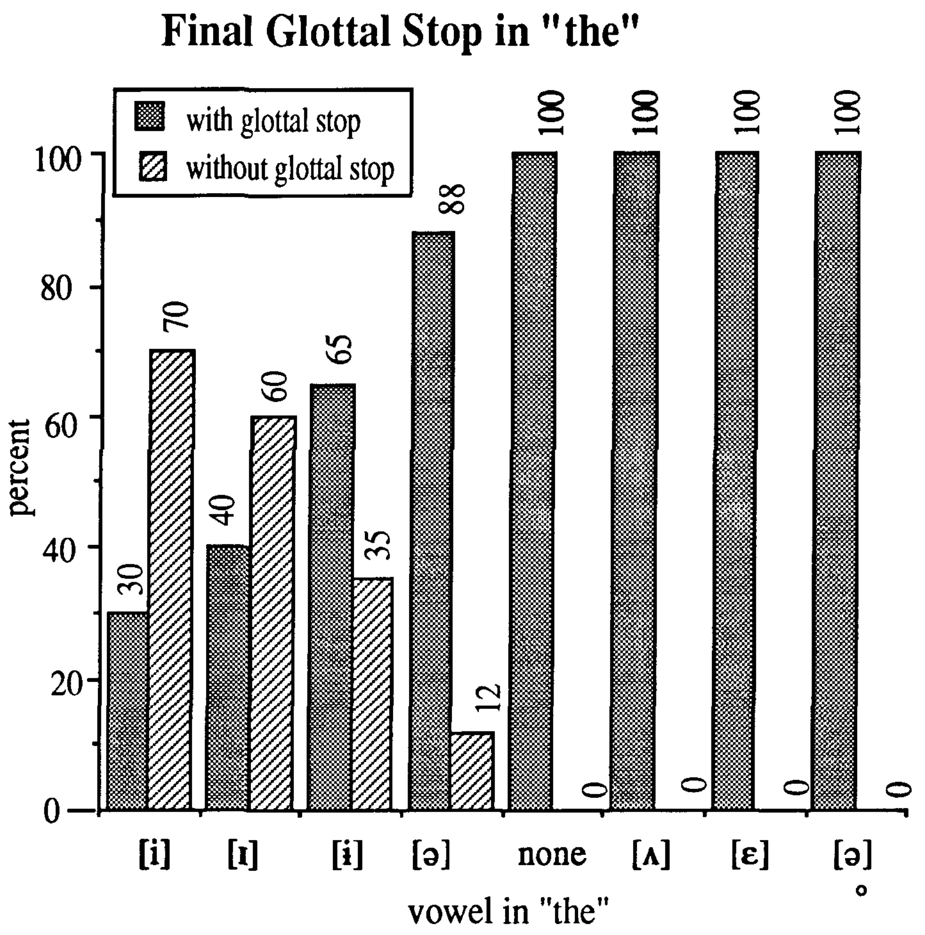


Figure 5 - Glottal stop vs. no glottal stop between "the" and words beginning with vowels, according to the vowel found in "the".

The study of the phonetic form of "the" before words beginning with a phonemic vowel shows that several vowel qualities occur in "the", with [i] the most frequent, but that younger speakers show an emerging trend away from [i]. The implication for TESL and for text-to-speech of these results with "the" is that there is probably no point in forcing the distinction between [ði] vs. [ðə]; the implication for speech recognition, however, is that both variants still need to be listed in a recognition lexicon. The glottal stop study suggests that use of glottal stop variants should be conditioned by the vowel of "the".

Since all phones in TIMIT are listed with their start and end times (to allow alignment with the waveforms), the "phones" database can also be used to study segment durations. For example, the durations of the vowels in all tokens of "the" can be compared across different speaker variables. Speakers' dialect region has a significant effect on this measure, with Southern speakers having longer vowels in "the" than other speakers.

A second study using the TIMIT segment durations concerns the occurrence of so-called "epenthetic [t]" in certain contexts. The most common context in TIMIT is between underlying /n/ and /s/. With [t] epenthesized between these, "sense" can be pronounced like "cents" and "prince" can be pronounced like "prints", etc. Many phonologists and phoneticians have discussed this phenomenon, which is seemingly pervasive among speakers of American English. Blankenship (this volume) explores the contextual effects on the occurrence of t-epenthesis in the TIMIT sentences. Although various patterns can be seen in the data, it appears that TIMIT does not contain enough speech to allow detailed comparisons of specific contexts of the sort usually implicated in phonological generalizations.

Here we will consider instead a further aspect of [t] epenthesis: a particularly interesting claim about the phonetics of epenthesis that has evoked much attention is that epenthetic [t] is shorter in duration than [t] from underlying /t/ [4]. This claim was based on 60 tokens of each type only some of which showed the difference. TIMIT provides many more tokens: 187 (26%) of the 712 tokens with underlying /ns/ were transcribed with an intervening [t] (closure or release or both), and there are 129 tokens of /nts/. These tokens were subdivided according to the stresses of the preceding and following vowels and according to utterance-final vs. nonfinal position. Durations of the two kinds of [t] are compared in Figure 6. This figure shows three different divisions of the data into subgroups: one (the first two pairs of bars) based on stress of preceding context; another (the next two pairs of bars) based on stress of following context; the last (the last two pairs of bars) based on position in utterance. The effect of utterance-final lengthening on both epenthetic and phonemic [t] is the most salient aspect of the figure, but what we are interested in is the local comparison of each pair of bars. In every comparison the epenthetic [t]s were a few msec shorter than the phonemic [t]s, but never significantly so. Thus TIMIT does not provide clear evidence for or against a durational difference, though it does make clear that if such differences exist they are not robust enough for lexical decision in a speech recognition system. Small differences that may be significant in controlled laboratory speech samples or with homogeneous groups of speakers may be washed out in a speech database like TIMIT (see also [2]).

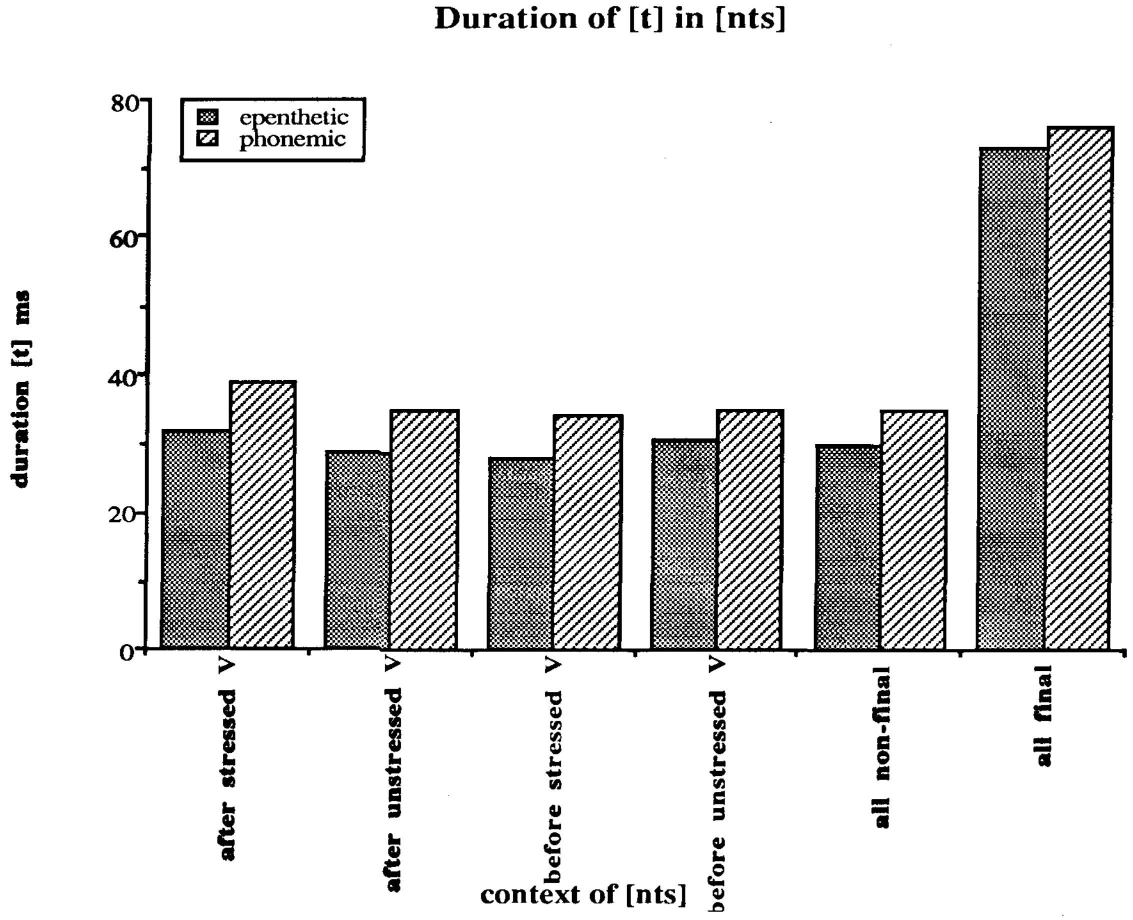


Figure 6 - Epenthetic vs. phonemic [t] durations in several context comparisons

We have found TIMIT to be inappropriate for other kinds of studies as well. We had thought that TIMIT would be ideal for testing an intriguing claim in the literature [3] that prevoicing of English /bdg/ is more common among men than women. Prevoicing refers to the presence of voicing (glottal pulsing) during the closure of a stop after a pause or a voiceless segment. One of the two sentences spoken by all the talkers in TIMIT begins with /d/, and with other postpausal tokens there is a total of 858 /bdg/ eligible for prevoicing in the Prototype version of TIMIT. Given the transcription system of TIMIT, prevoicing can be identified from the "phones" file, without acoustic analysis, as follows. The closures and releases of stop consonants are segmented and labeled as separate phones. The only way the closure of a stop can be found postpausally is by any voicing that occurs during it. Otherwise the cosure cannot be distinguished from the pause; both appear as silence. Therefore any "closure" transcribed after a pause must be a voiced closure, making it a simple matter to tally the proportion of /bdg/ with prevoicing. Unfortunately, almost no tokens in TIMIT show prevoicing: where the prior literature leads us to expect about 40% of tokens to be prevoiced, in TIMIT only 2.4% are. The reason is probably that the speech recordings distributed on the TIMIT CD were made with a close-talking microphone which would be unlikely to pick up signals radiated primarily through the neck walls, as prevoicing is. In any event, the question of a sex difference in prevoicing is moot in TIMIT.

Another feature of TIMIT with limited usefulness is the dialect coding. Recall that TIMIT assigns speakers to one of eight dialect categories, of which seven are regional and the last is "more than one of the above". (Presumably speakers were asked where they grew up.) These regional divisions are an odd combination of very broad and very narrow. (See [6] and [13].) The "New England" region is a fairly narrow subdivision and appears to be Kurath's "Eastern New England". The "Northern" region appears to be Kurath's "North" minus New England and New York City, extended westward through the Great Lakes into Montana. The "North Midland" region appears to be Kurath's North Midland (largely Pennsylvania and DC), extended westward through Kansas, Nebraska, and southern South Dakota. The "South Midland" region appears to be Kurath's South Midland, extended westward and southward through Texas and Oklahoma. The "Southern" region appears to include Kurath's Upper South and Lower South, or Coastal and Gulf South, plus western Tennessee. "New York City" is the only metropolitan region given a dialect coding. The "Western" region includes everything else, some 10 states in the western third of the US, but its speaker sample appears to be weighted towards the Southwest (the part nearest Texas, where recording was done). No further information about speakers' origins is given in TIMIT.

These are largely macro-dialects. The usefulness of such divisions depends on how a given phonetic difference is distributed across the country. We briefly examined this aspect of TIMIT by looking at two well-known speaker differences in American English that the dialect calibration sentences (spoken by all 630 speakers) were designed to study. The occurrence of /s/ vs. /z/ in "greasy" is a dialect marker that should be replicable in TIMIT, since it follows a large geographical division which is part of the basis for traditional dialect distinctions. Indeed, in TIMIT the use of /z/ is concentrated in two dialect regions (South Midlands and Southern), as seen in Figure 7.

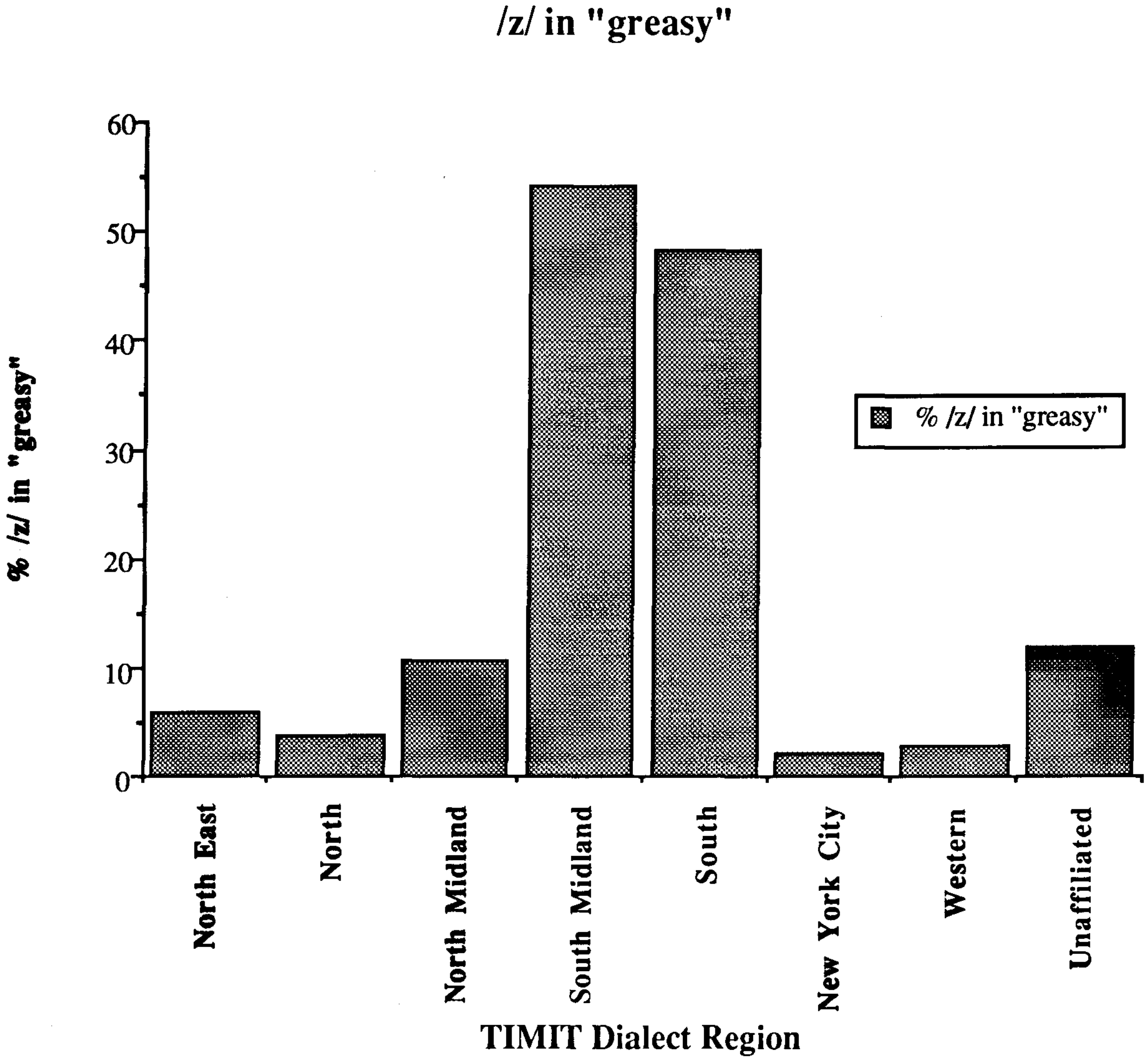


Figure7 - Percent tokens of "greasy" with /z/ by TIMIT dialect regions

On the other hand, the merger of [ɑ] and [ɔ], potentially seen in all or part of "wash water all", is more scattered over the country. The TIMIT dialect regions show no difference in the use of [ɑ] vs. [ɔ] in these words. Thus the usefulness of the dialect regions coded in TIMIT should not be overestimated, even for geographical features.

We imagine that pronunciations in the dialect calibration sentences themselves could be used as the basis for further dialect coding of speakers, in a way that would allow further study of regional dialect differences, but this would require careful planning and possibly a great deal of work.

Social and ethnic dialects are as important as regional ones in the study of variation in pronunciation. TIMIT reports education level and race for speakers. There are few people of color included in TIMIT, and we have not attempted to use these two characteristics in studying variation, except to verify that the men and women speakers in TIMIT do not differ in education levels [1].

III. ACOUSTIC ANALYSIS OF VELAR/VOWEL COARTICULATION

Acoustic analysis can also be performed on speech tokens identified by database searches. We have used Milenkovic's CSpeech acoustic analysis program for IBM PCs, which has a command allowing it to read TIMIT files, with good results. However, we are currently using Kay Elemetrics's CSL, which also reads the TIMIT phonetic transcription, converts it to IPA symbols, and displays it time-aligned with the waveform. Neither the TIMIT CD nor CSL provides any tools for automatic searching of speech files or automatic acoustic analysis based on such searches, nor have we develped any. Instead, relevant tokens are identified from a Reflex database search, file names and paths are printed out, and files are called up and analyzed in CSL by the user, one at a time.

We have conducted one analysis of this sort to date. It is well-known that in English (and many other languages), velar stops are fronted in their place of articulation on the palate when they occur before front vowels. (See Keating and Lahiri this volume for references and discussion.) We looked at whether this fronting was equally strong after front vowels.

The first step was to locate in the Reflex database all velar stops coded as having releases and with a pause either before or after. In such tokens, there is only one adjacent vowel affecting the velar. Additional tokens with a schwa adjacent were also located to boost the sample size. Each token was loaded into CSL (at the TIMIT sampling rate of 16kHz) and a 100 point spectrogram displayed. From the spectrogram, several timepoints were located: the release burst, the onset of formant transitions after the burst (whether voiced or aspirated) for prevocalic stops, and the offset of formant transitions before the closure for postvocalic stops, and the midpoint of the vowel. At each such timepoint, a 20 ms Blackman window was centered and an autocorrelation LPC spectrum (14 coefficients, 512 points) was displayed, with an FFT spectrum if necessary, for measurement of spectral peaks. In the burst, the measured value was the frequency of the highest- amplitude peak below 4 kHz; at transition onset, the frequencies of F2, F3, and F4; and similarly for the vowel mid-point and the transition onset. For any one token, then, seven measurements were made: for CVs, the burst, F2/F3/F4 onset, F2/F3/F4 midpoint; for VCs, F2/F3/F4 midpoint, F2/F3/F4 offset, and the burst. All values were logged in CSL from the screen into a text file, for subsequent statistical analysis.

First, ANOVA was used to show that velar stops do coarticulate with following vowels in the kind of speech represented in TIMIT. All designs for TIMIT data are non-repeated-measures and unbalanced datasets. The spectral peak of the release bursts in CVs were compared across the vowel contexts. It should be born in mind that in this kind of data sample, vowel contexts cannot be matched exactly across conditions. Because the sample includes very few tokens of some of the vowels transcribed, the analysis was not based on the phonemic identity of the vowels. Instead, the vowels were categorized acoustically, into four groups according to their steady-state F2 values (high, mid-high, mid-low, low F2). These correspond roughly to [i], other front vowels, back or central unrounded vowels, and back rounded vowels. Data from both male and female speakers were analyzed, but different cut-off values were used to group the formants for the two sexes. An ANOVA with factors vowel- group, speaker-sex, and consonant-voicing showed that burst frequency does vary significantly with the F2 of the following vowel, indicating very strong coarticulatory effects. Figure 8 shows the main effect only; speaker sex was also a significant determinant of burst frequency, but consonant voicing was not.

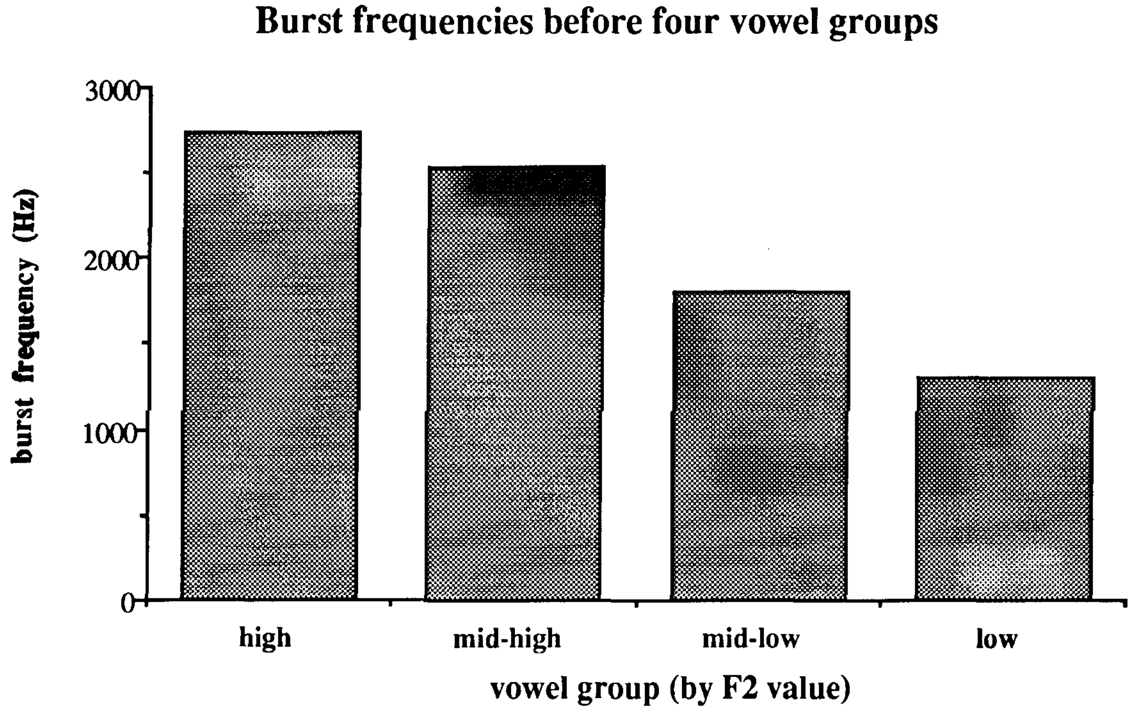


Figure 8 - mean values for burst frequencies of velars BEFORE 4 vowel sets

Turning then to the comparison of velars in CV vs VC position, again tokens from both male and female speakers were used and the vowels were divided into groups based on their F2 values. There are fewer VC than CV tokens in TIMIT, so the vowel groups were limited to the mid-high and mid-low groups where both CV and VC tokens are numerous. An ANOVA again used the factors speaker-sex and vowel-group, and also position of the velar (CV vs VC). Again both sex and vowel group were highly significant. The effect of position was significant overall (p < .032), but its direction depends on the vowel group. Figure 9 shows the effects of vowel group and position. The key comparisons are the vowel groups within each position. A higher burst frequency adjacent to a mid-high vowel means that the consonants are coarticulated with the vowels. Both CVs and VCs show such coarticulation. However, the effect -- viewed as a spread between the values in the two vowel contexts -- is much weaker in VCs. Prevocalic velars show a large effect of the following vowel, over 800 Hz, while postvocalic velars show a much smaller effect of the preceding vowel, under 300 Hz. Thus we can conclude that in terms of the burst, velars are indeed more fronted (acoustically speaking, at least) before than after the fronter vowels.

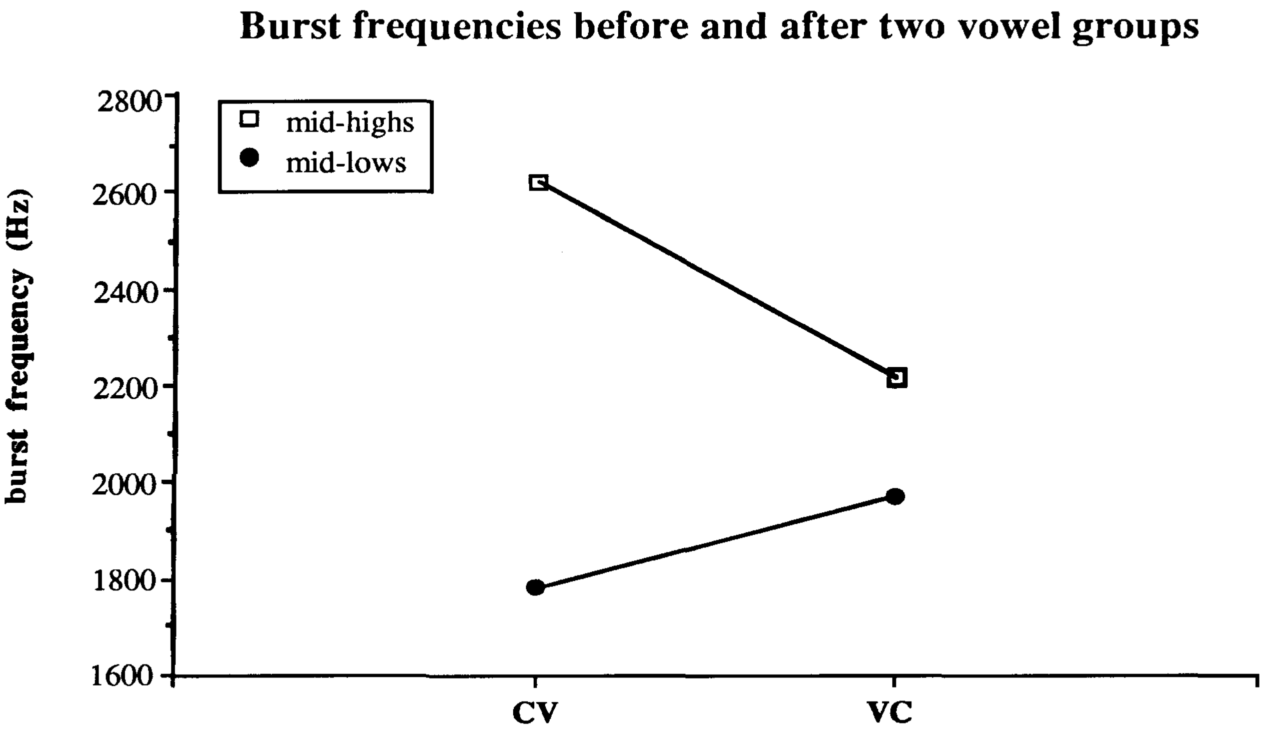
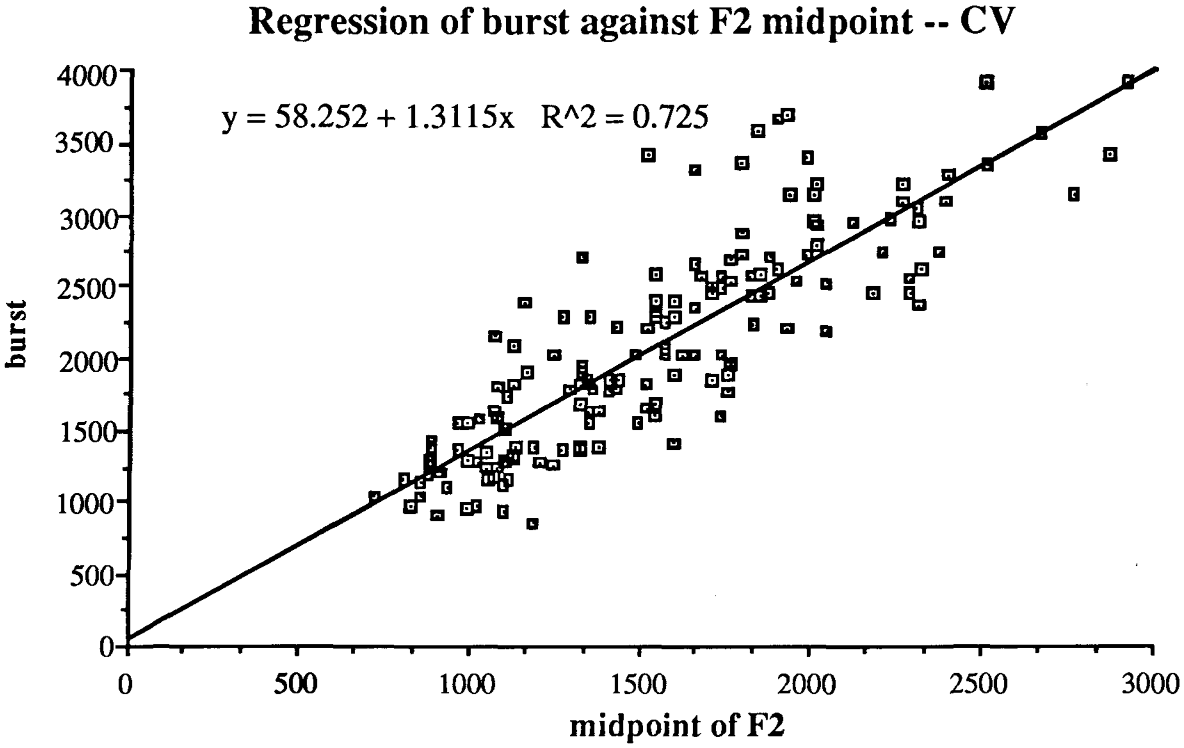


Figure 9 - mean values for burst frequencies of velars before and after 2 vowel sets

Various regression analyses were also performed to understand the relation between bursts, transitions, and vowel mid-points. Burst frequency was regressed against each of the three formant values at vowel midpoint (F2, F3, F4), somewhat in the style of [12]. All speakers and vowels are plotted together. In both CV and VC positions, the burst value was better predicted the lower the formant (F2 better than F3 better than F4), and for each formant, the fit was better for CV than VC samples, though it was significant in every case. Figure 10 shows the relation of burst to F2 midpoint in CV vs. VC samples. This analysis, like the ANOVA, shows that the burst value is more dependent on the vowel value in CVs than in VCs (higher r^2 value), but that in both cases the relation is strong (p=.0001).



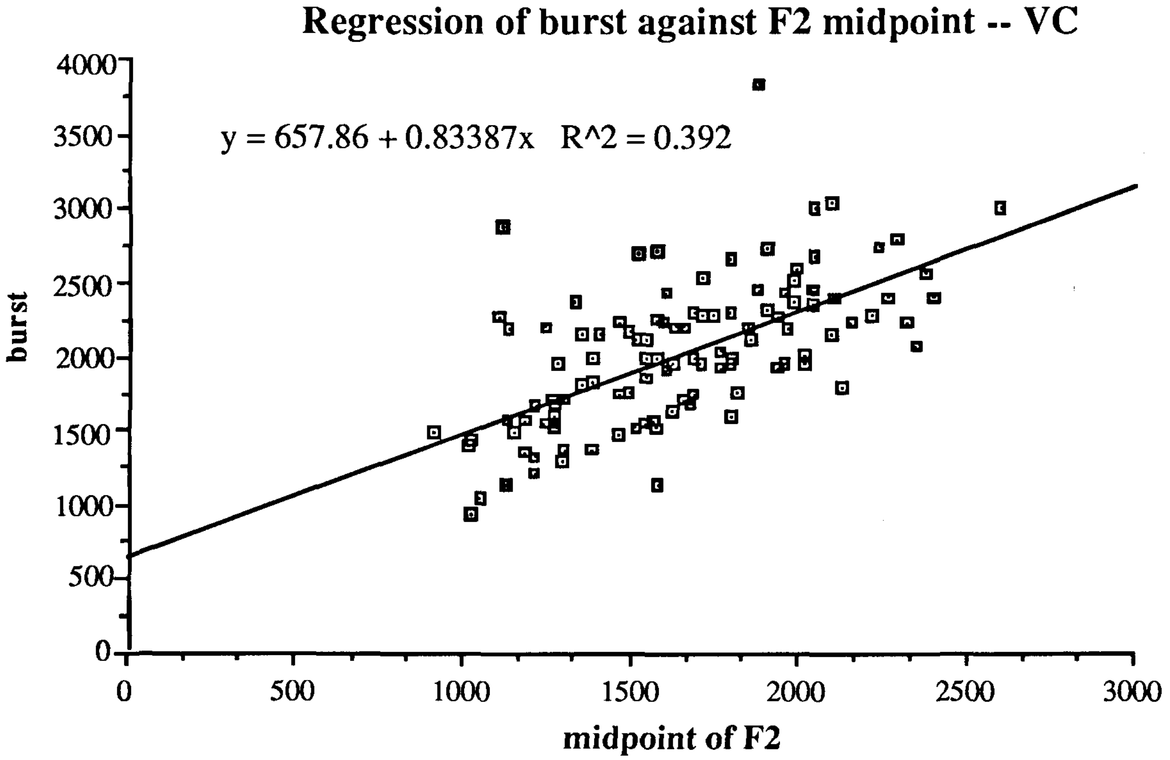
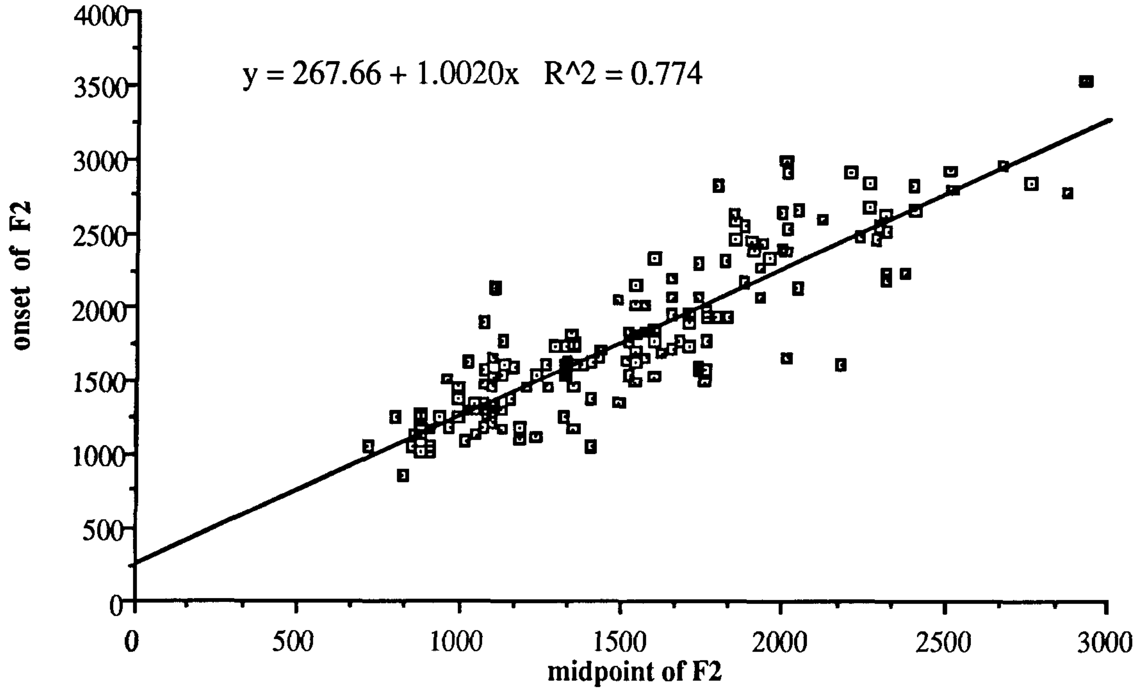


Figure 10 - regressions of burst against F2 midpoint, CV and VC

These analyses, then, show that the release of a velar is more influenced by the adjacent vowel in CV than in VC tokens. If, however, we look at formant transitions rather than bursts, this difference disappears or is even reversed. For CV tokens, the value of the onset of each formant transition (whether voiced or aspirated) was regressed against the value of that formant at vowel midpoint, and for VC tokens, the value of the offset of each formant transition was regressed against the value of that formant at vowel midpoint. (This analysis is more like that in [12], but the difference is that here, formant edges are not determined by voicing onset/offset.) As would be expected, these fits of formant edges to midpoints were better than those above, of bursts to midpoints. Also, in both CV and VC positions, again the fits were better the lower the formant. What is interesting is that the VC fits are as good as or better than the CV fits. Figure 11 shows the relation of F2 edge (onset or offset) to F2 midpoint in CV vs. VC samples. Thus we can conclude that in terms of formant transitions alone, velars about equally affected by vowels on either side.

Regression of F2 onset against F2 midpoint -- CV



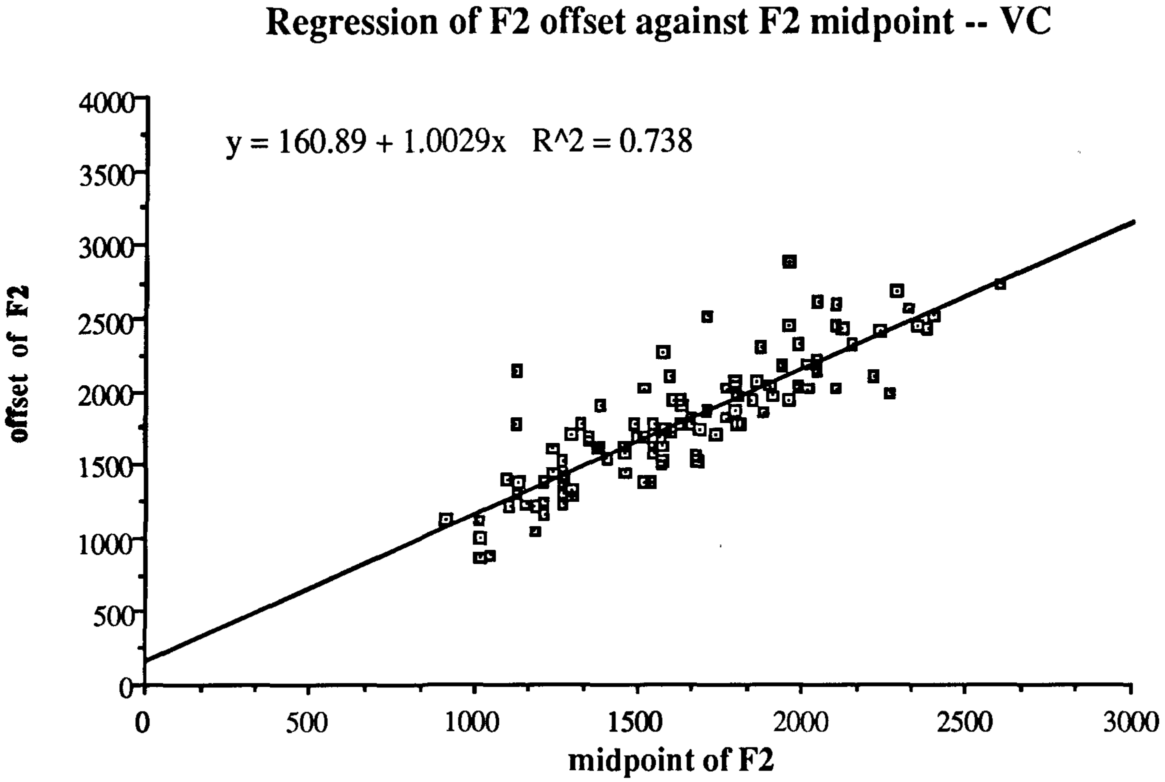


Figure 11 - regressions of F2 edge against midpoint, CV and VC

This makes sense when we consider that transition onset and offset are about equi-distant from the vowel midpoint, but that the bursts in CV and VC differ in this respect. The burst in VC is much further from the midpoint of V than is the burst in CV. Therefore we would expect the VC burst to be freer of the influence of V, just as we would expect the onset of closure for C in CV to be. What this suggests is that the perceived frontedness of a velar after a front vowel will depend on whether the listener attends more to the formant transition into the closure, or the later release. This interpretation accords with that in a recent small laboratory study by Nolan [9], in which velars were paired with [ɪ] and [ʌ]. Despite Nolan's intuition that his velars are more fronted before front vowels than after them, EPG data showed the reverse to be true. Nolan attributed his "intuition" to the burst properties of the utterances.

IV. CONCLUSION

The TIMIT CD-ROM database can be a useful tool for linguistic phoneticians interested in describing the phonetic characteristics of American English. It does have some limitations, however, even beyond the obvious limitation to read speech. First, it requires additional software for any kind of analysis; however, commercial database programs and acoustic analysis systems are viable options for at least some research questions. Second, the regional dialect coding of speakers is only marginally useful, since the dialect regions are broad areas of the country of very different sizes, which tend to wash out all but the most pronounced phonetic differences and underrepresent large population groups such as California. Third, the recording conditions place some limitations on what can be analyzed. Fourth, though the speech sample has been designed to contain all phoneme combinations, the number of tokens of any one combination is typically small. If comparisons of particular sequences are desired, there will not be enough tokens to offset all the variation in other aspects of the tokens, such as prosodic environment. Thus TIMIT does not comprise enough speech for reliable study of many phonetic and phonological hypotheses, which typically refer to highly specific contexts; our study of epenthesis is an example. Nonetheless, we hope to have shown that if such limitations are kept in mind, TIMIT provides a powerful new kind of resource for phonetics.

The results of such phonetic research should be useful for work on text- to-speech and speech recognition systems. In both areas, researchers may want to know about both the range of variation and the most common variant in the pronunciation of a sound sequence or a given lexical item. Data from TIMIT of the sort presented here can help determine variant pronunciations across the population that should be listed in a lexicon for a speaker-independent recognition system. It can also help determine a typical pronunciation for men/women, older/younger speakers, tall/short speakers, or whatever speaker characteristics are being modeled for synthesis or recognition.

REFERENCES

[1] Byrd, D. “Preliminary results on speaker-dependent variation in the TIMIT database” JASA, vol. 92, no. 1, pp. 593-596, 1992.

[2] T. Crystal and A. House. "Segmental durations in connected-speech signals: Current results," JASA, vol. 83, no. 4, pp. 1553-73, 1988.

[3] J. Flege and K. Massey (1980) English prevoicing: Random or Controlled? LSA summer meeting, Alberquerque, NM (unpublished)

[4] M. Fourakis and R. Port. "Stop epenthesis in English," Journal of Phonetics, vol. 14, pp. 197-221, 1986.

[5] C. Henton and A. Bladon. "Developing computerized transcription exercises for American English," Journal of the IPA, vol. 17, pp. 72- 82, 1987.

[6] H. Kurath. A Word Geography of the Eastern United States. Ann Arbor: University of Michigan Press. 1949.

[7] P. Ladefoged. A course in phonetics. 2nd ed. New York: Harcourt Brace Jovanovich. 1982.

[8] L. Lamel, R. Kassel, and S. Seneff. "Speech database development: design and analysis of the acoustic-phonetic corpus," Proceedings DARPA speech recognition workshop, 1986.

[9] F. Nolan. "Phonetic correlates of syllable affiliation," to appear in Papers in Laboratory Phonology 3, ed. P. Keating, Academic Press, 1993.

[10] D. Pallett. "Speech corpora and performance assessment in the DARPA SLS program," Proceedings ICSLP 90 (Kobe), 1990.

[11] S. Seneff and V. Zue. "Transcription and alignment of the TIMIT database," distributed with the NIST TIMIT CD-ROM database, 1988.

[12] H. Sussman, H. McCaffrey, S. Matthews. "An investigation of locus equations as a source of relational invariance for stop place of articulation," JASA vol. 90, pp. 1309-1325, 1991.

[13] J. C. Wells. Accents of English, vol.3. Cambridge: CUP. 1982.

[14] V. Zue, S. Seneff, and J. Glass. "Speech database development at MIT: TIMIT and beyond," Speech Communication vol. 9, pp. 351-356, 1990.

WHAT TIMIT CAN TELL US ABOUT EPENTHESIS

Barbara Blankenship

1. Introduction

In American English, an epenthetic stop can occur between a nasal or lateral and a following fricative. Thus for some speakers there is no difference between the members of these pairs: *sense/cents,false/faults*. The mapping between the phonological representation and the phonetic outcome of words with intrusive stops is an important issue in linguistic theory. Table 1 presents a summary of recent theories and findings, to be discussed below.

Table 1. Theories and findings about epenthesis

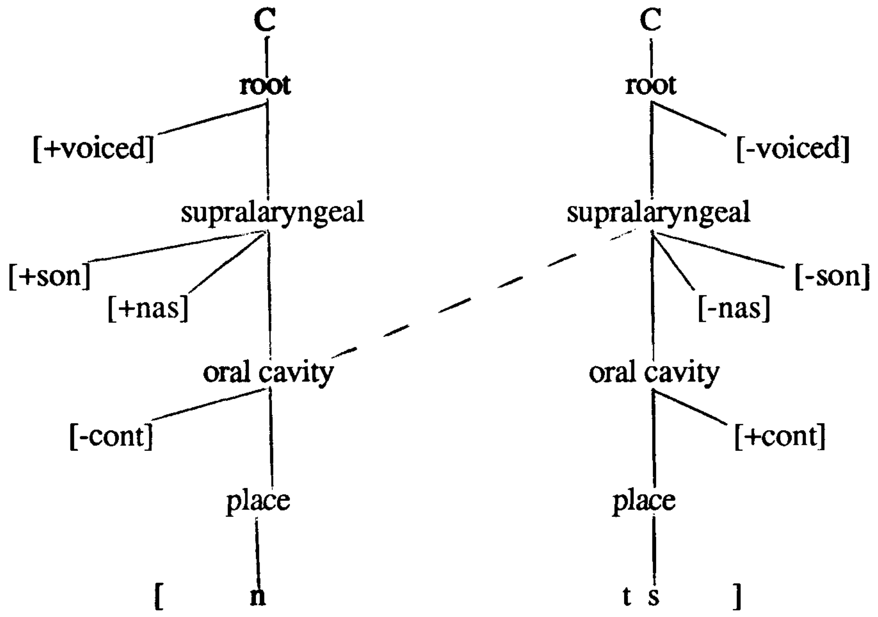
Zwicky (1972)	Dialect-specific phonogical rules
Ohala (1975)	Result of articulatory constraints
Fourakis and Port (1986) experiment:	
	1. Dialect-specific 2. Phonemic [t] longer than epenthetic [t] due to articulatory phase rules.
Clements (1987)	Phonemic [t] has its own C-slot; epenthetic [t] shares slot with the following fricative.
Lee (1990) experiment:	
	No difference in duration between epenthetic and phonemic [t], (intense / in tents).

Some researchers (for example, Ohala, 1975) have claimed that these intrusive stops are the automatic result of articulatory constraints. Others (for example, Zwicky, 1972) claim that the stops are produced by dialect specific phonological rules. Fourakis and Port (1986) found that four subjects who spoke South African English did not produce epenthetic stops, thus invalidating any claims based on articulatory constraints. They found that five subjects who spoke American English always inserted epenthetic stops when the following fricative was voiceless, a result supporting theories that epenthesis is a phonological process.

Fourakis and Port found that for the American speakers, the average duration of the epenthetic stops was shorter than that of the stops with an underlying phonemic representation, leading to the conclusion that the phonetic distinction between *dense* and *dents* was not entirely neutralized by epenthesis. They suggested that epenthesis was controlled by language specific "phase rules" that govern the timing of local articulatory gestures but make reference to phonological properties.

An alternative explanation was proposed by Clements (1987) within the framework of autosegmental phonology. The underlying stop occupies its own C-slot, whereas the epenthetic stop shares a C-slot with the following fricative, as shown in figure 1. The epenthetic [t] results from spreading of the oral cavity node of the preceding /n/ to the following /s/.

Figure 1: The Clements (1987) model of epenthetic stop formation



But the duration differences found by Fourakis and Port may not constitute an adequate basis for the assumption that epenthetic stops are regularly shorter than phonemic stops. As Lee (1990) pointed out, Fourakis and Port found consistent differences only for the pair *dense/dents* but not for other similar pairs such as *tense/tents*. Lee's follow-up study found no significant difference in duration for three speakers of Midwestern American English. Moreover, Lee's perception experiment using tokens from these speakers showed no correlation between the stop duration and listeners' identification of the words *intense* and *in tents* in citation form. Thus any duration differences that were present were not being exploited to make perceptual judgments.

2. The TIMIT corpus

The phonetic section of the TIMIT corpus contains phonetic transcriptions with segment durations for 2342 sentences read by 630 native speakers of American English from a wide range of dialect groups.

Three types of sentence are included, as summarized in table 2. Two "calibration sentences", designed to compare dialects, were read by all 630 speakers. 450 "phonetically compact" sentences were designed to provide examples of American English phonemes in all possible left and right segment contexts. Each of these sentences was read by seven speakers. The remaining 1890 are "diverse sentences" selected from the Brown corpus. Each of these was read by only one speaker. Thus there is a total of 6300 utterances.

The transcriptions are consistent and accurate. They are based on computer-aided segmentation, observation of spectrograms, and listening to the recordings.

Table 2. Samples in the TIMIT corpus.

2,342 sentences:		
	2 "calibration" sentences for dialect identification	(630 speakers each)
	450 "phonetically compact" sentences: all American English phones in all possible segment contexts	(7 speakers each)
	1890 "diverse" sentences: Brown corpus	(1 speaker each)
630 speakers. Each reads 10 of the sentences:		2 calibration 5 phonetically compact 3 diverse

3. Data set for this study

The phonetic transcriptions containing [ns] and [nts] were selected for this study, since /ns/ is by far the most frequent environment for epenthesis in English. The TIMIT transcriptions contain 1009 strings of [ns] and [nts], either within words or across word boundaries. Table 3 summarizes them by source. 712 of the strings resulted from an underlying /n(#)s/ and 152 from an underlying /nt(#)s/. These 864 tokens form the data set for this report. The 145 instances of [ns] and [nts] resulting from other underlying segments (e.g., /nd#s/, /ŋ ##s/, /n#z/) were omitted from consideration. The data set includes 66 of the phonetically compact sentences (with three to seven readings of each) and 295 of the diverse sentences (with one reading each). 467 speakers are represented in the data set.

Table 3. Instances of [ns] and [nts] in TIMIT

Underlying:	n(#)s	712
	nt(#)s	152
	other	145
Source sentences for the 864 utterances:		
	66 phonetically compact (3 to 7 readings each)	
	295 diverse (1 reading each)	
467 speakers		

Some proposed rules of epenthesis refer to stress environment. Since the phonetic transcriptions in TIMIT do not include stress information, a simple stress evaluation was developed. The syllables on either side of the target segments were assigned the stress pattern H-L (high-low), L-H, H-H or L-L (with H-# and L-# for the ends of utterances), based on the presumed least-marked reading of the sentence. In most cases only lexical stress was of concern, since the relevant syllables had the same relationship to each other regardless of how the sentence is read.[1] When there was a possibility of ambiguity, the sentence was omitted from consideration in analyses where stress environment was a factor.

The phonetic transcriptions in TIMIT recognize two elements for each stop, the closure and the release. When [t] occurs between [n] and [s], it can be realized as a t-closure, a t-release, or both. Throughout this paper, the t-closure will be referred to as [cl] and the t-release as [rel]; [t] will be used to refer to either kind of element when the distinction is not important.

4. Results

The results will be discussed in two sections. The first will deal with the presence or absence of [t] in various environments. The second will present segment durations.

4.1 Presence of [t]

Both underlying /ns/ and /nts/ can be realized with or without a [t] element in production, but underlying /ns/ is realized with a [t] less frequently than is underlying /nts/. Table 4 gives the percentage of occurrences for each possible sequence of segments.

Table 4. Percentage of occurrences for each possible realization of /ns/ and /nts/
The symbol [n] here refers to both consonantal and syllabic nasals.

Underlying:		/ns/		/nts/	
Realization:	[n s]	74%		14%	
	[n cl s]	18%	26%	57%	86%
	[n rel s]	1%		4%	
	[n cl rel s]	7%		25%	
Total tokens:		712		152	

Underlying /nts/ generated some kind of [t] element in 86% of the tokens, whereas underlying /ns/ did so in only 26% of the tokens. Thus about one fourth of the /ns/ strings in TIMIT have epenthetic stops. The environments in which they can occur will be described below. (This report will not consider the processes that caused 14% of the /nts/ strings to lose their [t] in production.)

Since phonological rules of epenthesis often refer to stress patterns, it is interesting to see how the stress environment relates to realizations in the TIMIT data set. Table 5 shows the percentage of [t] elements generated by underlying /ns/ and /nts/ according to whether the preceding vowel was assigned high or low stress. (Thirteen /ns/ tokens and 14 /nts/ tokens were omitted from this section of the analysis because the preceding vowel could be either stressed or unstressed in an unmarked reading.) Epenthesis occurred in 25% of the instances where underlying /ns/ follows a stressed vowel and in 44% of the instances where it follows an unstressed vowel. Thus the stress of the preceding vowel appears to have no bearing on whether epenthesis can occur.

[1] Comparison of a sample of about 300 of these assignments against the actual recordings showed the assignments to be 87% correct. Errors were divided approximately equally between H and L assignments.

Table 5. Realizations of /ns/ and /nts/, according to stress of preceding vowel. (Figures for epenthesis are underlined.)

Preceding vowel:		high stress		low stress	
Underlying:		/ns/	/nts/	/ns/	/nts/
Realization:	[n s]	75%	18%	56%	13%
	[n t s]	25%	82%	44%	87%
Total tokens:		233	34	66	104

Some phonological models of epenthesis suggest that the /ns/ cluster must be tautosyllabic. Although the TIMIT transcription does not provide syllable information within the utterances, we know that the consonants at the end of an utterance are tautosyllabic. Table 6 shows the percentages for utterances with final /ns/ or /nts/. These items are a subset of those shown in table 5.

Table 6. Realizations of /ns/ and /nts/ in the last syllable of an utterance, according to stress of preceding vowel. (Figures for epenthesis are underlined.)

Preceding vowel:		high stress		low stress	
Underlying:		/ns/	/nts/	/ns/	/nts/
Realization:	[n s]	37%	0	24%	5%
	[n t s]	63%	100%	76%	95%
Total tokens:		16	3	88	20

A comparison of tables 5 and 6 reveals that in TIMIT, epenthesis is much more frequent for /ns/ strings at the ends of utterances than for /ns/ strings in general. But it is not possible to determine whether this is due to the fact that final /ns/ strings are tautosyllabic, or due to a phrasal effect such as pre-pausal lengthening. We will return to the topic of syllable affiliation in the discussion of table 8.

Table 7 shows the percentages according to whether following vowel was assigned high or low stress. (Thirteen /ns/ tokens and 14 /nts/ tokens were omitted from this section of the analysis because the following vowel could be either stressed or unstressed in an unmarked reading. In addition, 104 /ns/ tokens and 23 /nts/ tokens could not be used because they were at the end of an utterance. Therefore the token totals in table 7 are smaller than in table 5.)

Table 7. Realizations of /ns/ and /nts/, according to stress of following vowel. (Figures for epenthesis are underlined.)

Following vowel:		high stress		ow stress	
Underlying:		/ns/	/nts/	/ns/	/nts/
Realization:	[n s]	93%	9%	73%	19%
	[n t s]	7%	91%	27%	81%
Total tokens:		255	43	340	72

Tables 5 and 7 show that epenthesis can occur in any stress environment but is quite rare (7%) when the following vowel is stressed. This observation does not take into account the syllable affiliation of the /s/, information which is not provided in TIMIT. But Clements (1987) states that the rule of epenthesis does not apply if the following consonant (in this case, /s/) initiates a stressed syllable. To test this aspect of the rule, some knowledge of syllable affiliation is required.

Although syllable boundary analysis is beyond the scope of this report, an indication of word boundaries is feasible and may provide useful evidence. Table 8 gives percentages of [ns] and [nts] realizations of /ns/, organized according to the locations of word boundaries. The table includes only those instances where the following vowel is stressed.

Table 8. Realizations of /ns/ followed by a stressed vowel, organized by underlying segment string and word boundary. (C means one or more consonants.)

Location of /ns/:		Over boundary		Word internal		Word final	
Underlying:		/n#sV/	/n#sCV/	/nsV/	nsCV/	/ns#V/	/ns#CV/
Realization:	[ns]	99%	93%	93%	100%		83%
	[nts]	1%	7%	8%	0%		17%
Total tokens:		72	43	80	33	0	12

The first two columns of this table show that, contrary to Clements' statement, epenthesis can occur when the /s/ is at the beginning of a syllable (in this case at the beginning of a word). The fourth column supports Clements' statement for cases where /s/ is part of a word-internal syllable-initial cluster. But in none of these cases do we know whether the /s/ is ambisyllabic, and thus the evidence is weak[2] A look at the individual

[2] Bruce Hayes (personal communication) claims that a word-final consonant is always ambisyllabic when the following word begins with a vowel. Such examples would yield further evidence to test Clements' claim, but unfortunately the TIMIT corpus provides no instances of /ns#V/.

transcriptions also provides interesting evidence regarding stress environment. They include unusual examples of [t] in such words as "considerable", "coincided", "consuming", "consists", and over the word boundaries of "clean slate" and "one sitting". One subject produced "open street" without epenthetic [t] and "open scaffold" with epenthetic [t] in the same sentence. The text of the complete sentences is given in table 9. It is difficult to explain the presence of [t] in such examples by stating that the /s/ is a member of the preceding syllable. Therefore we must accept the possibility that epenthesis can take place when the /s/ initiates a stressed syllable.

Table 9. Individual examples of epenthesis where the following consonant /s/ introduces a stressed syllable

considerable	Both the conditions and the complicity are documented in considerable detail
coincided	Employee layoffs coincided with the company's reorganization.
consuming	Growing well-kept gardens is very time consuming.
consists	Her wardrobe consists of only skirts and blouses.
clean slate	But we cannot start off with a clean slate.
one sitting	Children can consume many fruit candies at one sitting.
open scaffold	Princes and factions clashed in the open street and died on the open scaffold.

4.2 Segment durations

Fourakis and Port (1986) found that epenthetic [t] is shorter than [t] resulting from an underlying phoneme, (with 60 tokens of each type). Lee (1990) found no significant difference in duration between epenthetic and phonemic [t], (with 54 tokens of each type). Table 10 shows the average durations of epenthetic and phonemic [t] in TIMIT.[3] Because segments at the ends of an utterance are extra long, they have been placed in a separate section of the table.

[3] The three sets of measurements appear to be roughly comparable. In both Lee (1990) and Fourakis and Port (1986) a voiceless interval between [n] and [s] that is longer than 10 msec counts as a [t]; otherwise it is measured as part of the [n]. In TIMIT, a voiceless interval of any duration could count as a [t], but in fact no instances shorter than 12 msec occurred.

Table 10. Average duration in milliseconds of epenthetic and phonemic [t], organized by stress context.

	Epenthetic		Phonemic	
	Number of tokens	Average duration (msec)	Number of tokens	Average duration (msec)
Preceded by:				
stressed vowel	**43**	**32**	23	39
unstressed vowel	**65**	**29**	84	35
Followed by:				
stressed vowel	**19**	**28**	40	34
unstressed vowel	**89**	**31**	57	35
Entire set (except utterance-final)	108	**30**	107	35
Utterance final	79	73	22	76

In the entire set of non-final [nts]'s, with 108 epenthetic and 107 phonemic tokens, epenthetic [t] has a shorter average duration (at 30 msec) than phonemic [t] (at 35 msec), but the difference is not significant at the .01 level. When the data are grouped by stress environment, epenthetic [t] averages shorter than phonemic [t] in each group, but again not significantly. Likewise in the utterance-final strings: although the [t]s are much longer, the two kinds are not significantly different from each other.

If the indications of the TIMIT corpus are true, then there is no need for a phonological model such as those of Clements (1987) or Fourakis and Port (1986) to explain different durations for phonemic and epenthetic [t], since the differences are not significant.[4] The differences found by Fourakis and Port would have been due to the small sample size.

Although the TIMIT corpus provides a much larger sample, it has some drawbacks. Since it is not statistically balanced and it does not provide the target segments within a controlled context, the data are more difficult to enterpret. But Lee's carefully designed experiment corroborates the TIMIT evidence. TIMIT's size and diversity make it a valuable tool for exploring phonetic processes, particularly when it is used in combination with laboratory experiments on a more controlled data set.

5. Summary

The TIMIT corpus provides the following information with regard to epenthesis in /ns/ strings.

1. Epenthesis occurs in about one fourth of the /ns/ strings of American English.

[4] There may be other unexplored differences in the surrounding segments, e.g., duration of the preceding vowel, presence or absence of nasalization on the preceding vowel, whether the /t/ becomes glottalized.

2. Epenthesis in American English is either optional or dialect-specific. Speakers whose dialect includes epenthesis do not use it consistently, even within the same utterance.

3. Epenthesis can occur in any stress context but is more frequent when the /ns/ is adjacent to an unstressed syllable, or at the end of an utterance. It is rare when the following vowel is stressed.

4. There is no significant difference in duration between epenthetic and phonemic [t].

References

Clements, G.N. (1987). Phonological Feature Representation and the Description of Intrusive Stops. In A. Bosch et al, (eds.) *Parasession on Autosegmental and Metrical Phonology,* Chicago Linguistic Society.

Fourakis, Marios, and Robert Port (1986). Stop Epenthesis in English. *Journal of Phonetics* 14, 197-221.

Lee, Sook-Hyang (1990). The Duration and Perception of English Epenthetic and Underlying Stops. Paper presented at the fall 1990 conference of the Acoustical Society of America.

Ohala, John (1975). Phonetic Explanations for Nasal Sound Patterns. In C.A. Ferguson, L.M. Hyman, J.J. Ohala, (eds.), *Nasalfest: Papers from a Symposium on Nasals and Nasalization*, 289-316. Language Universals Project, Department of Linguistics, Stanford University. (Did not read; cited in Lee.)

Seneff, Stephanie, and Victor W. Zue, n.d. Transcription and Alignment of the TIMIT Database. In *Getting Started with the DARPA TIMIT CD-ROM*. Cambridge, Massachusetts Institute of Technology, Research Laboratory of Electronics. (Distributed with the TIMIT database by the National Institute of Standards and Technology.)

Zwicky, A. M. (1972). Note on a phonological hierarchy in English. In R. P. Stockwell and R.K.S. Macauley, (eds.), *Linguistic Change and Generative Theory,* 275-301. Bloomington: Indiana University Press. (Did not read; cited in Lee.)

Sex, Dialects and Reduction

Dani Byrd

(to appear in *Proceedings of the International Conference on Spoken Language Processing*, 1992)

ABSTRACT

A set of phonetic studies based on analysis of the TIMIT speech database is presented which addresses topics relevant to the linguistic and speech recognition communities. Using a database methodological approach, these studies detail new results on the effect of speakers' sex and dialect region on pronunciation.[1] This report concerns speaker-dependent effects on certain phonetic characteristics of speech often involved in reduction such as speech rate, stop releases, flapping, central vowels, non-canonical phonation type, syllabic consonants, and palatalization processes.

INTRODUCTION

The majority of acoustic-phonetic studies to date have shared, in the most general terms, a common methodology. Each speaker reads an entire set of carefully controlled experimental speech materials designed to answer the specific questions motivating a given experiment. Much valuable linguistic knowledge has been gathered from experimentation of this sort; however, this general method has limitations. It considers a small number of homogeneous speakers which may be unrepresentative of the diversity found in a larger population of the language's speakers. The limitation to carefully controlled test items may focus the speaker's attention on contrasts, thereby exaggerating them. Finally, a new experiment must be designed and executed for each new question which arises.

A recent trend in new methodologies for speech investigation is the development of general-purpose speech databases for acoustic phonetic analysis. Such databases are well-suited to answering questions about pronunciation where small variations of specific sentential context are irrelevant due to the size and diversity of the data set. General factors in pronunciation variability such as speaker-specific characteristics can also be investigated. In practice, a large speech database will include either long samples from relatively few speakers, or short samples from many speakers. Examples of the first type include sociolinguistic studies, and the emerging New York University database [4]. An example of the second type is the TIMIT database for American English.

The TIMIT database, which is described below, was designed jointly by the Massachusetts Institute of Technology, Texas Instruments, and SRI International under sponsorship from the Defense Advanced Research Projects Agency-Information Science and Technology Office (DARPA-ISTO) for the development and evaluation of automatic speech recognition systems [3]. It was desired that the database incorporate sufficient variability to examine details of the acoustic realization of phonetic segments as affected by canonical characteristics of the phoneme, contextual dependencies, syntactic effects, and such speaker-specific factors as dialect, sex, age, and education [3]. As a large corpus of speech, TIMIT provides an interesting testing ground for the linguist to assess the accuracy of generalizations regarding allophony and regularity in English that have previously been based on more "artificial" laboratory experiments or naturalistic observation. Additionally, the linguist's perspective may highlight fertile areas in which to gather acoustic-phonetic data relevant to the phonetic classification and speech recognition goals which TIMIT serves.

Just as allophonic variation is dependent on phonological context, phonological or phonetic variation can be influenced by speaker-specific factors In what follows, I will describe a series of small studies which exploit TIMIT for the purpose of investigating the influence of speaker sex and dialect on certain gross indicators of reduction in speech. Reduction includes different types of simplification which speakers regularly exhibit in pronunciation. Reduction is often but not necessarily correlated with speech rate and/or casualness of speech, but may also occur as a result of optional (post-lexical) phonological rules. Generally speaking, reduced forms of a word are simplified with respect to the canonical (underlying) form and often involve assimilations, vowel centralization, and the deletion or simplification of segments. Conversely hyper-articulated forms of a word are generally pronounced slowly, have all underlying segments fully articulated, and use more acoustically peripheral vowels.

METHOD

The TIMIT database includes 630 talkers and 2342 different sentences. The sentences are of three types. Two calibration sentences are spoken by every talker. These sentences were designed to "incorporate phonemes in contexts where significant dialectical differences are anticipated" [7]. Additionally, 450 phonetically-compact sentences were designed to incorporate as complete a coverage of phonetic pairs as practical [3]. Each one of these sentences is spoken by seven talkers. Finally, 1890 randomly selected sentences were chosen to provide alternate contexts and multiple occurrences of the same phonetic sequence in different word sequences [[1] in [7]]. Each talker read two calibration sentences, five phonetically compact sentences, and three randomly selected sentences. Eight dialect regions were established for classifying the speakers, and 70% of the speakers were male and 30% female. Information regarding the speaker's age, race, and education are also provided for the user. A map showing the geographical divisions into the seven geographic dialect regions can be found in Fisher, et al., 1986 [1]. Broadly speaking, these include the seven geographical regions of New York City, the Western United States, New England, the northern Midwest, the southern Midwest, the Atlantic seaboard, and the southeastern United States. An additional dialect division called "Army Brat" denotes speakers unaffiliated with a particular geographic region. There appears to be no statement on the part of the database designers as to the motivation for establishing these particular dialect regions; nor is there any explanation of the marked asymmetry between the number of male and female speakers. The distribution of speakers is shown in Table 1.

dialect region	female	male	total	percent
1:North East	18	31	49	7.78%
2:North	31	71	102	16.19%
3:N Midland	23	79	102	16.19%
4:S Midland	31	69	100	15.87%
5:South	36	62	98	15.56%
6:NY City	16	30	46	7.30%
7:West	26	74	100	15.87%
8:"Army Brat"	11	22	33	5.20%
total	192	438	**630**	
percent	30.5%	69.5%		

Table 1 - Sex and Dialect Distribution in the TIMIT database

The digitized recordings are accompanied by a time aligned phonetic transcription. Transcribed phones include stop closures and releases, syllabic and non-syllabic nasals and laterals, flaps (nasal and non-nasal), pauses, epenthetic and glottal stops, and a wide variety of vowels including breathy [ə], [y], [ɚ], [ʌ], and [ɨ]. An orthographic transcription and the waveform are also provided. A description of the inventory of transcribed elements and criteria for segmentation can be found in Zue and Seneff, 1988 [6]. The validity of the results reported in this work depend entirely on the correctness and consistency of the phonetic transcriptions.

RESULTS AND DISCUSSION

Speech rate

The first and perhaps most important quality examined for the TIMIT speakers was their speaking rate. The duration of both calibration sentences was used to examine speech rate as all 630 speakers read these same two sentences. A three-factor analysis of variance (ANOVA) with the factors of sex, dialect region, and calibration sentence number (1 or 2) with all interactions was used to test the effect of sex and dialect region on speaking rate. There is a significant effect of sex on rate (F(1,1228)=37.301, p=.0001) with men speaking 6.2% faster than women. There is also a significant effect of dialect region (F(7,1228)=5.424), p=.0001). The dialects range from slowest to fastest in the following order: South, South Midland, NY City, North, West, North Midland, North East, and "Army Brat." (Although the smaller representation of regions 1, 6, and 8 may impair the overall accuracy of these rankings.) A post-hoc Scheffe's S test yields a significant difference between the "Army Brat" group and the South Midland and between the "Army Brat" group and the South (p<.05). A marginal difference is seen between the North Midland and the South (p=.0558) and between the West and the South (p=.0726). There is also an interaction of sex and dialect region. While the South Midland region is ranked as the slowest speaking region for men, it is only the fourth slowest for women. The North East and West regions are ranked as fastest and third fastest respectively for men but are second slowest and most slow respectively for women. There is no interaction of sex or dialect with the two-level calibration sentence factor. Differences in speaking rate are important to bear in mind in the overall examination of reduction, as rate has a substantial influence on the production of reduced word forms.

In order to determine whether the frequency or duration of pauses could have contributed to the above effects, these were examined in the calibration sentences. A chi-square test determines that pauses are randomly distributed between men and women but are not randomly distributed between dialects (χ^2=19.325, p=.0072). Speakers from the South Midland and the South paused more often than expected while speakers from the North Midland, West, and the "Army Brats" paused less often than expected given a random distribution. This result explains, at least in part, the effect of dialect region on rate described above, as pauses contributed to sentence duration. A three-factor ANOVA shows no effects (or interactions) of sex, dialect region, and sentence number on the duration of pauses. In summary, both sex and dialect have significant influences on speaking rate; influences which we may by extension expect to find on reduction processes which are affected by the rate of speech.

Sentence-final stop releases

All sentence-final oral stops (n=1130) were evaluated as to whether their closures have a release or not. A contingency table analysis was conducted where the expected number of releases is assumed to be randomly distributed with respect to sex and dialect within the group of speakers who produced stops in this position. The contingency table analysis determines that sex has a significant effect on the distribution of final released and unreleased stops (χ^2=11.651, p=.0006).

Women released their sentence-final stops more often than men: 67% versus 56% of the time. There is no significant effect of dialect region on the frequency of releases in sentence-final stops. As place of articulation has a significant effect on the frequency of release of sentence-final stops (χ^2=40.829, p=.0001) with the probability of a release increasing from the bilabial to the alveolar to the velar place, the effect of sex was tested separately at each of these places. This contingency table analysis shows significant effects of sex on the frequency of release at the alveolar and velar place of articulation but not at the bilabial place of articulation.

The final word '*that*' of one of the calibration sentences read by all 630 speakers also ends in an oral stop. This stop is realized as released 23% of the time, unreleased 67% of the time and as a glottal stop 9% of the time. (Five cases were excluded due to collection error in searching the database.) For the speakers who produced a stop in this position, a contingency table analysis determines there to be a significant effect of sex (χ^2=49.146, p=.0001) but no effect of dialect on whether a release was produced. Women released this stop 32.5% of the time and men 23.1% of the time. In summary, the sex of the speaker exerts a significant effect on the frequency of a sentence-final stop release. Such releases are characteristic of hyperarticulated speech and are less often found in reduced pronunciations.

Flaps

Another process found in continuous speech is alveolar flapping. This rule as stated by Oshika, et al., 1975 [4] describes a process whereby an intervocalic stop, optionally preceded by [r] or [n], is realized as a flap when it occurs in a falling stress pattern (as in '*winter*') or between reduced vowels (as in '*ability*') [4]. Across word boundaries, there are no stress conditions (as in '*what#is*' or '*not#equal*') [4]. Two analyses of flaps in TIMIT were conducted. In the first, the frequency distribution of all oral and nasal flaps in the database was considered where the expected distribution given the null hypothesis is assumed to be random over the database as a whole, i.e. men will produce 69.5% of the flaps and women 30.5%. A chi-square test indicates a significant effect of sex on the frequency of both nasal (n=1331) and oral (n=3649) flaps (χ^2=55.341, p=.0001 and χ^2=12.585, p=.0829 respectively). The women produce significantly fewer flaps than the men. No effect of dialect region is found on the frequency of oral or nasal flaps.

A second analysis was conducted on the sequence '*suit in*' and the word '*water*' which are included in one of the calibration sentences which all 630 speakers read. Word final flaps (n=121) were produced by 19% of the speakers and word medial flaps (n=624) by 99%. While there is no effect of sex or dialect on the frequency of a wordmedial flap in '*water*', a chi-square test indicates a significant effect of sex on the frequency of flaps in the '*suit in*' sequence (χ^2=13.934, p=.0002). Only 9% of the women flapped the alveolar consonant in this sequence while 19% of the men did. Dialect region also had an effect on the frequency distribution of the word-final flap (χ^2=20.03, p=.0055). The North and North East speakers flapped less often than expected, and the North Midland speakers more often than expected. (Note: The effect of sex on the distribution of oral flaps across the whole database remains significant χ^2=8.829, p=.003) when the flaps in the calibration sentence are excluded (from analysis.)

Central Vowels

An analysis of the distribution of the three central vowel transcribed in TIMIT (ɨ, ʌ, ə) was conducted. A total of 17858 central vowels were transcribed of which 55% were [ɨ], 18% were [ʌ], and 27% were [ə]. (Note: the calibration sentences were excluded from this analysis so as not to overrepresent any particular central vowel tending to occur there.) A chi-square test determines there to be no effect of sex or dialect region on the frequency distribution of the total number of central vowels. However, when each vowel is considered separately, differences in their use across sex and dialect emerge. Both sex and dialect have a significant effect on the distribution of [ɨ] (χ^2=7.161, p=.0074 and χ^2=14.203, p=.0477 respectively). Men use [ɨ] less frequently than would be predicted by a random distribution. Speakers from the North East, NY City, and the West use [ɨ] more frequently than expected and speakers from the North and N. Midland less frequently than expected. A chi-square test for the effect of sex and dialect region on the distribution of [ʌ] show a significant effect of sex (χ^2=5.79, p=.0161) but no effect of dialect region. Women used this central vowel more frequently than expected given a random distribution. Finally, a chi-square test for the effect of sex and dialect region on the distribution of [ə] show significant effects of sex (χ^2=21.591, p=.0001) and dialect region (χ^2=30.15, p=.0001). Women used this central vowel less frequently than the men. Speakers from the North, NY City, and the West use this vowel less frequently while speakers from the N. Midland, South, and, especially, the South Midland use this vowel more frequently. In summary, we have found that women use the two more peripheral of the central vowels, [ɨ] and [ʌ], more frequently than men and the more central of the vowels, [ə], less frequently than the men. This difference again suggests that the women's speech may be less reduced in certain respects than the men's.

An ANOVA was conducted on the durations of these vowels as duration often corresponds to the degree of acoustic reduction as in cases of articulatory undershoot. A four-factor ANOVA with all two-level interactions was conducted to test effects on vowel duration with the factors of vowel, sex, dialect region, and position in the word (medial, initial, final, or unaffiliated). While, not surprisingly, vowel and position had significant effects, so did dialect region (F(7,17796)=6.668, p=.0001). The South and South Midland have the longest central vowels and differ significantly from most other dialect regions as determined by a post-hoc Scheffe's S test. Sex did not have a significant effect. There were also significant interactions of vowel and dialect region; sex and position; and vowel and position. No other interactions were significant.

Glottal stop, breathy vowel, [h], and [ɦ]

The distribution of the glottal stop, the breathy vowel, the [h] and the voiced [h] were evaluated. Use of the glottal stop can occur in English between vowels, before a vowel, in place of an alveolar stop, and in many other positions. Not much is known about patterns of distribution of glottal stop. This corpus provides a data set for determining general distributional patterns of glottal stop across a variety of prosodic and phonological contexts. The frequency distribution of glottal stops (n=4834) is significantly affected by both sex and dialect region (χ^2=30.906, p=.0001 and χ^2=148.154, p=.0001 respectively) as shown by a chi-square test. Women have significantly more glottal stops than the men. Speakers from the North and South use more glottal stops than expected while speakers from the North Midland and the "Army Brats" use less. When the position of the glottal stop in the word is considered (initial (49%/total), medial (6%/total), final (16%/total), or unaffiliated (not part of a word) (29%/total), the effect of sex on the frequency of glottal stops is significant at all positions, with the effect always in the direction indicated above. The effect of dialect region is significant in initial position and final position only. When we consider the small sample of 57 glottal stops produced in place of the sentence final [t] of the word "that" in one calibration sentence, we find that the production of a glottal stop in this position is not significantly influenced by sex (χ^2=1.763, p=.1843), although the distribution favors the direction demonstrated above.

It is somewhat unexpected to find that the speaker-dependent characteristic of sex was related to the use of glottal stop in this way. In fact, women's voices are often characterized as more breathy, and glottal closure is often related to creakiness in the voice quality of the signal. It may be that the glottal stop is used as a devoicing mechanism more often by women or that it participates in allophonic patterns which are less productive for the men.

All (n=478) instances of the voiceless vowel transcribed in TIMIT were evaluated with respect to sex and dialect region of the speaker. The frequency distribution of the voiceless vowel shows a significant effect of sex (χ^2=36.471, p=.0001) but no effect of dialect region as determined by a chi-square test. Women have significantly fewer voiceless vowels than the men. Again this result is surprising given the commonly accepted conception of women's voice quality as being more breathy than men's. It is, however, not unexpected in light of the findings reported here which suggest that women produce less reduction in their speech than do men. Many voiceless vowels would presumably be created by overlapping a neighboring laryngeal opening movement with the syllable nucleus. The generally less reduced forms apparently produced by women would be less inclined to such overlap.

All (n=1313) [h]'s in TIMIT were evaluated. The frequency distribution of [h] shows a significant effect of sex (χ^2=3.815, p=.0508) but no effect of dialect region as shown by a chi-square test. Women have significantly fewer [h]'s than the men. Here, we again see what appears to be an odd result meriting further investigation.

The frequency distribution of all [ɦ]'s in TIMIT (n=1523) showed no effect of sex or dialect region as determined by a chi-square test. While we argued that overlap of a laryngeal opening movement might be more likely by men, it seems that this argument can not be extended to include the case of the voiced [h] where a breathy voice segment is produced.

Syllabic consonants

All the syllabic consonants transcribed in the database were evaluated for speaker-specific effects on their distribution. Syllabic consonants are the result of complete reduction of the vocalic syllable nucleus. These consonants include [l̩] (n=1291, 52% of total syllabic consonants), [m̩] (n=171, 7% of total), [n̩] (n=974, 39% of total), and [ŋ̩] (n=43, 2% of total). No effect of sex and dialect region is found in a chi-square test on the distribution of [l̩], [m̩], and [ŋ̩] (The chi-square test is not valid for effect of dialect region on [ŋ̩] distribution.) Additionally, no effect of dialect region was found for [n̩]. However, the sex of the speaker did have a significant effect on the frequency of [n̩] (χ^2=12.632, p=.0004), such as might occur in the word *'hidden'* or *'button.'* Women use significantly fewer syllabic [n]'s than the men. As reduction in the environment of alveolar consonants is a particularly common process, it is important to note that men and women appear to produce this, and only this, syllabic consonant with different frequency.

Palatalization

All sequences of *'s_sh'*, *'z_sh'*, *'sh_s'*, and *'sh_z'* occurring across a word boundary in the canonical forms of the TIMIT sentences were evaluated to determine whether both consonants were produced by the speaker or whether assimilation occurred. The presence of a pause (as determined by the TIMIT transcription) between words was also noted. The null hypothesis is that the number of assimilations and pauses are randomly distributed with respect to dialect and sex within the group of speakers saying these sequences. A contingency table analysis determines there to be no significant effect of sex on whether both consonants were produced or whether there was a pause between them. The lack of any significant effect of sex on whether assimilation occurred was seen both when C1 is the post-alveolar consonants, and when C1 is an alveolar consonant.

In a second analysis, one of the calibration sentences was investigated where the sentence included the phrase *'had_your'*. Three types of productions were included in the contingency table analysis. In 44% of the cases the intervocalic sequence [dǰ] was produced; in 20% of the cases [d˺y] was produced, and 36% of the time [dy] was produced where the last two productions differ in the presence or absence of an alveolar release before the glide. Thirty speakers who produced unusual sequences which occurred in less than 1.5% of the cases were not included in the analysis.

The null hypothesis is that each of the three sequences described above are randomly distributed with respect to dialect and sex within the remaining group of 600 speakers. A contingency table analysis determines there to be no significant effect of sex or dialect region on which sequence was produced. Nor is there any effect on whether the stop was released in the cases where it occurred before a glide. Lastly, there is no effect of sex or dialect on whether an affricate or a glide was produced.

CONCLUSION

In conclusion, it has been shown that in this corpus, speaker-specific characteristics of sex and dialect region influence speaking rate, the choice of central vowels, and the frequency distribution of stop releases, flaps, glottal stops, [h]'s, breathy vowels, and the syllabic alveolar nasal. This and similar database analysis offer a promising new methodology for approaching speech analysis. We have seen here a number of indications that sex, and, to a lesser extent, dialect may influence reduction processes even in this relatively formal scripted speech. These results have nothing to suggest with respect to the causes behind this correlation; these effects may be task-specific or not. However, the results do suggest that speech analysis for both synthesis and recognition goals will provide a more comprehensive picture of variation if similar numbers of men and women are included in speech databases. Furthermore, because there are many aspects of pronunciation which seem to differ between men and women, it may be profitable to incorporate within a recognition lexicon different probabilities leading to particular pronunciations for a male as compared to a female speaker. Or, if a single most likely pronunciation is sought, certain differences in the lexicon for male and female speakers might improve accuracy. It is interesting to consider whether many of the effects described are highly enough correlated with rate to make rate rather than sex a driving force for a recognition system.

It is distressing how little is known about the effects of speaker dialect and sex on pronunciation variability or general speech patterns. In particular, this study has pointed out several areas in which our knowledge of speech differences between the sexes is deficient. An improved understanding of the influences of speaker-dependent variables should be valuable in improving the performance of recognition systems which will presumably be employed by a wide variety of users. Speaker-specific variation is also of interest to the linguist attempting to describe the universal, language-specific, and speaker-dependent characteristics of speech. For example, Labov states that "sexual differentiation of speech often plays a major role in the mechanism of linguistic evolution" ([2], p. 303). Exploring how anatomical and social factors interact in the speech of both men and women is vital to understanding the linguistic principles governing language variability. TIMIT has proven to be fertile ground for gathering acoustic-phonetic knowledge which is of interest to all speech scientists attempting to describe regularity and variability in English speech.

Endnotes

[1] A preliminary report by the author on a subset of this research appears in Byrd, *JASA*, July, 1992.

Acknowledgment

This research was supported by the National Science Foundation and the Department of Linguistics at UCLA. The author wishes to thank Patricia Keating and Edward Flemming for their valuable assistance.

References

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

[2] W. Labov. *Sociolinguistic Patterns*. (University of Pennsylvania Press, Philadelphia), 1972.

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

[4] B. Oshika, V.W. Zue, R.V. Weeks, H. Neu, and J. Aurbach,. The role of phonological rules in speech understanding research. *IEEE Transaction on Acoustics, Speech, and Signal Processing*. Vol. ASSP-**23**, No. 1, 104-112, 1975.

[5] N. Umeda. Multimode database and its preliminary results. *JASA* 89,4 pt.2 p. 2010, 1991.

[6] V.W. Zue, and S. Seneff,. Transcription and alignment of the TIMIT database. *Proceedings of the Second Meeting on Advanced Man-Machine Interface through Spoken Language*, pp. 11.1-11.10, 1988.

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

A Note on the English Palatalization Process

Dani Byrd

The TIMIT database of American English speech was used to examine across word sequences likely to undergo palatalization in American English. For an introduction to the use of TIMIT to examine linguistic characteristics of English see Byrd (this volume) and Keating (this volume). One allophonic rule which has long been noted in English is the rule states that alveolar obstruents become palatalized before palatals (see, eg. Oshika, et al., 1975) This rule often occurs across word and morpheme boundaries, but may be bounded by a clitic group boundary (Hayes, 1989). However, the details and productivity of this rule are not clear. In this note, all sequences of <z#sh>, <sh#s>, <s#sh>, and <sh#z> (where # denotes a word boundary) from the TIMIT orthographic corpus were examined for their phonetic realization as indicated by the TIMIT transcription.

Table 1 shows the number of tokens spoken in the TIMIT corpus for each underlying sequence type.

Table 1

underlying sequence	number	percent of total
z#sh	76	49%
sh#s	22	14%
s#sh	50	32%
sh#z	7	5%

A total of 155 tokens.

The results obtained from the transcriptions can be considered in several ways. Both consonants were produced in 30.3% of the utterances, while assimilation (or deletion) yielding a single consonant occurred in 69.7% of the utterances. The post-alveolar fricative was produced in all but two tokens; an alveolar fricative was produced in 49 tokens (31.6%). The following table shows how often assimilation yielding a single consonant occurred in each of the underlying sequences.

Table 2

underlying sequence	% realized as [ʃ]
z#sh	78.9% (note: also two tokens realized as [z] only)
sh#s	31.8%
s#sh	78%
sh#z	0%

The effect of the underlying sequence on whether both consonants were produced was significant as determined by a contingency table analysis (χ=37.744, p=.0001)

When [ʃ] was the first consonant of the sequence it was always produced, but when an alveolar fricative is C1, it was produced only 22% of the time. When C2 was the post-alveolar fricative, there were only two tokens in which it was not produced, but when [s] was C2 it was not produced 32% of the time. However, [z] was always produced when it occurred in the C2 position. The underlying sequence did have a significant effect on whether an alveolar consonant was produced in the sequence as determined by a contingency table analysis (χ=34.809, p=.001). The frequency with which an alveolar consonant was produced is shown in Table 3.

Table 3

underlying sequence	alveolar fricative present
z#sh	21.1%
sh#s	68.2%
s#sh	22%
sh#z	100%

A pause occurred in 12 (7.7%) of the utterances. A contingency table analysis showed no significant effect of the underlying sequence on whether a pause occurred, although seven of the 12 pauses occurred in /z#sh/ sequences. There were three cases where a [z] was devoiced, all occurring in the phrase “redwoods shimmered.”

The syntactic environment was also considered. 60% of the tokens occurred in modifier-noun sequences, 37.4% in verb-noun sequences, and 2.6% in other types of syntactic environments. These syntactic groupings had no effect on whether assimilation occurred or whether a pause occurred as determined by a contingency table analysis. Likewise, the speaker's sex and dialect region had no effect on the frequency of assimilation or pausing.

These results generally support the formulation of the palatalization rule specifying a potential palatalization site as occurring when a post-alveolar fricative follows an alveolar fricative, but the 31.8% occurrence of palatalization when the context is reverse, ie. the alveolar after the post-alveolar, suggests that such a rule needs refinement so as to include this context as a possible, if less likely, palatalization site. (See the discussion of variable rules in Labov, 1972) Oshika et al. (1975) note that “it is possible that preceding palatals may also influence [s] and [z]” (p.108). The asymmetrical nature of the result can be considered further evidence for the preference for anticipatory coarticulation as against carryover coarticulation (Ohala, 1990, 1991; Javkin, 1979; Lindblom, 1983 and others). Note particularly that the post-alveolar fricative is realized as an alveolar in only two out of 155 tokens.

Acknowledgements

This research was supported by the National Science Foundation and the Department of Linguistics at UCLA. Many thanks are due to Edward Flemming for designing and implementing the structure of a relational Macintosh database with which TIMIT is used at UCLA.

References

Byrd, D. (in press and in this volume) Sex, dialects, and reduction. to appear in the proceeding of the International Conference of Speech Processing.

Hayes, B. (1989) The prosodic hierarchy of meter. *Phonetics and Phonology*, 1. Kiparsky, P. and Youmans, G. eds. San Diego:Academic Press, pp. 201-260.

Javkin, H.R. (1979) Phonetic universals and phonological change. *Report of the Phonology Laboratory (Berkeley)*, No. 4.

Keating, P., Blankenship, B., Byrd, D., Flemming, E., Todaka, Y. (in press and in this volume) Phonetics analyses of the TIMIT corpus of American English. to appear in the proceeding of the International Conference of Speech Processing.

Labov, W. (1972) *Sociolinguistic Patterns.* Philadelphia:University of Pennsylvania Press.

Lindblom, B. (1983) Economy of speech gestures. in MacNeilage (ed.) *The Production of Speech.* New York:Springer-Verlag, 217-246.

Ohala, J. (1990) The phonetics and phonology of aspects of assimilation. in J. Kingston and M.E. Beckman (eds) *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech.* Cambridge, Cambridge University Press, 258-275..

Ohala, J. (1992) The segment: primitive or derived. to appear in Docherty G. and Ladd, D.R. (eds.) *Papers in Laboratory Phonology II: Segment, Gesture, and Tone.* Cambridge, Cambridge University Press.

Oshika, B., Zue, V.W., Weeks, R.V., Neu, H., Aurbach, J. (1975). The role of phonological rules in speech understanding research. *IEEE Transaction on Acoustics, Speech, and Signal Processing.* Vol. ASSP-**23**, No. 1, 104-112.

A Note on English Sentence-Final Stops

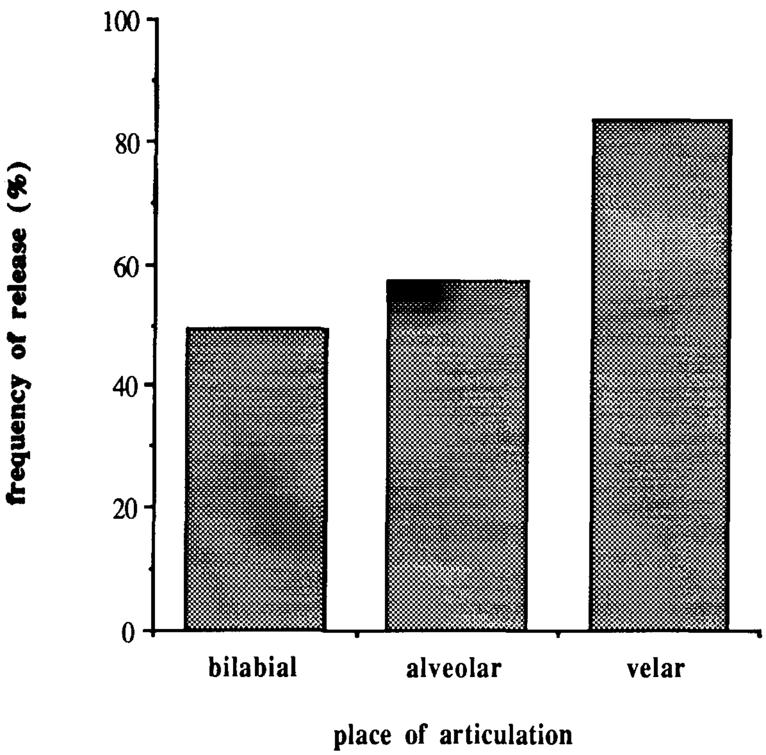
Dani Byrd

The TIMIT database of American English speech was used to examine the character of sentence-final oral stops. For an introduction to the use of TIMIT to examine linguistic characteristics of English see Byrd (this volume) and Keating (this volume). For this note, the transcription given for the sentence-final position of all (non-calibration) sentences was examined for the occurrence of released and unreleased oral stops.

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

The place of articulation had a significant effect on whether a release occurred or not as determined by a contingency table analysis (χ=40.829, p=.0001). Bilabial stops were released 49.5% of the time, alveolar stops 57% of the time and velar stops 83.11% of the time. This result can be seen graphically in Figure 1.

Percent of sentence-final stops released



Voicing, however, did not have an overall effect on whether a release occurred. Voiced stops were released 59.5% of the time and voiceless stops 59.9% of the time. When each of the places of articulation is tested independently in a contingency table analysis, the bilabial and alveolar place show no effect of voicing on whether a released or unreleased stop occurs. In the velar stops, however, voicing did have a significant effect on whether a released or unreleased stop occurred (χ=3.902, p=.0482). The voiced velars were released in 64.3% of the cases while the voiceless velars were released 85.1% of the time.

Crystal and House (1988a) report a tendency in their data set for the velar consonants to be released more often than bilabial or alveolar consonants and state that this tendency is attributable to the behavior of the voiceless velar consonant. The results reported here are in accordance with that argument. A contingency table analysis of the effect of place on the occurrence of a release in which the voiceless velars are excluded yield no significant effect (χ=2.27, p=.3214) although the direction of the trend remains unchanged from that shown in Figure 1. As only 18 velar tokens remain in this analysis, a larger sample might raise the trend to significance even without voiceless velars. Another Crystal and House result reported in 1988(b), that "labials, particularly in unstressed syllables, [tend] to be completed [ie. released] more frequently than alveolars and velars " (p. 1580), was not supported by the TIMIT data considered here. They also report a tendency for voiceless stops to include a release more often than voiced stops (Crystal and House, 1988a). As an overall effect, this was also not evident in the data considered here.

The sex of the talker had a significant effect of whether a release occurred; dialect classification had no such effect. For a description of these results see Byrd (this volume).

Acknowledgements

This research was supported by the National Science Foundation and the Department of Linguistics at UCLA. Many thanks are due to Edward Flemming for designing and implementing the structure of a relational Macintosh database with which TIMIT is used at UCLA.

References

Byrd, D. (in press and in this volume) Sex, dialects, and reduction. to appear in the proceeding of the International Conference of Speech Processing.

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

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

Keating, P., Blankenship, B., Byrd, D., Flemming, E., Todaka, Y. (in press and in this volume) Phonetics analyses of the TIMIT corpus of American English. to appear in the proceeding of the International Conference of Speech Processing.

Phonetic Variants of the Determiner “the”

Yuichi Todaka

Introduction

Recent progress in speech recognition systems takes linguists and system developers into an infinite search for acoustic, phonetic, and intra-and inter-speaker variants of a spoken language. The TIMIT database of spoken American English, designed to have at least limited coverage of different dialects, is a possible tool to gain a better understanding of acoustic-phonetic rules such as velar fronting and flapping above the word level.

In the present study, the special behavior of the determiner "the' before vowels is examined. The word "the" is said to have two or three phonetic variants in natural (unemphatic) speech: (1)[ðə] before consonants; and (2) [ði] (Allen et al., 1987:87-88; Ripman, 1924:106) or (3) [ðɪ] (Kenyon, 1940:107; McLean, 1930:241) before vowels. Ladefoged (1982:98) describes the variants as "the" [ðə] before consonants and often [ði] or [ðɪ] before vowels. It is therefore interesting to examine how the word "the" is actually realized before vowels in the TIMIT corpus, since there are some disagreements among the sources. Furthermore, nothing is said in the sources about the possible inter-speaker differences which might underlie the phonetic variation.

Another interesting characteristic regarding the word "the" is the English glottal stop insertion rule. There is also some dissension among researchers about the environment in which glottal stop insertion takes place. One claim is that the occurrence of glottal stop in ordinary speech is "permitted in General American only after the word "the" before a stressed as in e.g., “the ocean” [ðə ʔo^{ʊ}ʃən]. Liason of the two vowels is also allowed, i.e., [ði o^{ʊ}ʃən]" (cf. Trager & Smith, 1951 in Henton & Bladen, 1987:76). Another view is that glottal stop occurs before word-initial stressed vowel after a syllabic segment (*not* a determiner) or a voiced nonplosive and a phrase boundary, e.g., "Liz eats" (Allen et al., 1987:87-88). Third, it is claimed that glottal stop is found "before initially stressed vowels, sometimes between vowels when the second vowel begins a stressed syllable" (Bronstein, 1960:79). Even though the present study only examines the "the + V" sequence, it is still interesting to find the distribution of glottal stop in that environment since the major disagreement seems to concern it.

Such findings can also serve an important function pedagogically. The word "the" is one of the most frequent words in writing and in speech (West, 1953). It is in fact the most common word in TIMIT, where the word "the" accounts for roughly 7 % of all word tokens (Lamel et al., 1986:101). Thus, the precise understanding of the phonetic variants of the word will help L2 learners' productive use of it, and thus perhaps the perceived fluency of their speech.

Research Questions

The present study considers the following issues:

1. What is the distribution of the phonetic variants of the word "the" before vowels?
2. Is the above distribution rule-driven?
3. Does the glottal stop insertion rule apply?

If so, in what context?

4. Do the above findings differ by dialect or by age of the speaker?

Procedures

Orthographic transcriptions of the MIT ("SX") and TI ("SI") sentences in the Prototype version of TIMIT were searched for all instances of "the" followed by a vowel letter. A total of 272 tokens were located (in 40 SX and 117 SI sentences). For each token, the time-aligned phonetic transcription of the sequence was taken from the corresponding transcription file. All analysis here is based on these transcriptions. These were tabulated according to several variables.

Results/Discussion

First, the vowel variants found in the determiner ([ð_]) are displayed in Table 1 and Fig. 1.

Vowel Variants found in TIMIT (Table 1)

	vowel	tokens	percentage
	iʸ	184	76.0
	ə	20	8.3
	ɨ	18	7.4
	-	9	3.7
	ɪ	5	2.0
	ʌ	3	1.3
	ɛ	2	0.8
	ə̥	1	0.4
total	8	242	99.9

note: - means that no vowel was transcribed in the data.

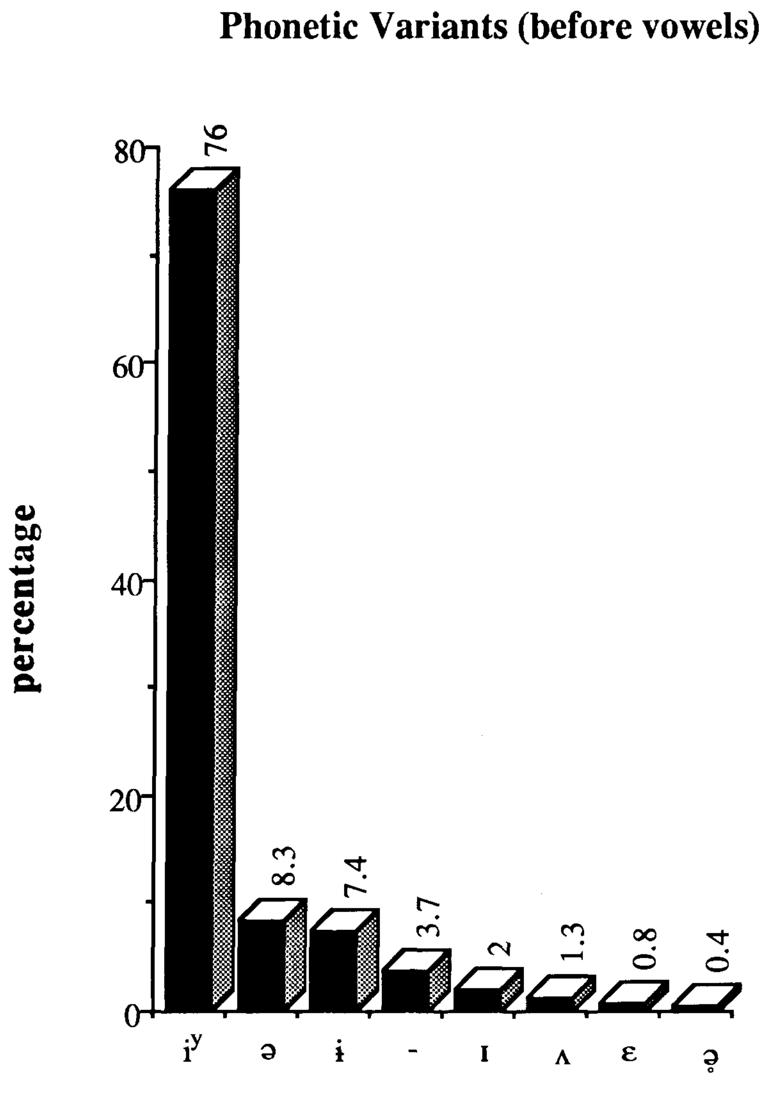


Fig. 1

Regarding the vowel variants of the determiner (Fig. 1), the findings in the TIMIT database coincide with the general description found in the literature. In other words, when the following segment after the determiner is a vowel, 76 % of the total tokens had [i^y]. Only 7.4 % and 8.3 % of the total tokens had [ɨ] or [ə]. The other vowels ([ɪ], [ɛ], [ʌ], and [ə̥]) found in the corpus counted only 4.5 % of all the tokens. This means that most speakers in the database pronounced the determiner as [i^y] when the following segment was a vowel. I also checked approximately 40 tokens of the "the + C" sequence to see if there were any cases where the vowel was pronounced as [i^y] in unemphatic speech. None were found: when the following segment was a consonant, the subjects in this sample pronounced the determiner as a reduced vowel, i.e., [ə] (77 %), [ʌ] (17 %) and [ɨ] (6 %). This result corresponds to the general description that native English speakers differentiate the pronunciation of "the" according to the following segment (vowel or consonant).

It is, however, interesting to point out the regional differences (Table 2 and Fig. 2).

REGIONAL DISTRIBUTION (TABLE 2)

regions/vowels	i^{y}	ɪ	ɛ	ə	ɨ	-	ə̥	ʌ	
New England	16	0	1	0	1	1	0	1	20
Northern	27	1	0	5	3	0	1	0	37
North Midland	32	0	0	3	0	0	0	1	36
South Midland	32	1	0	4	5	3	0	0	45
Southern	28	2	0	7	5	0	0	1	43
New York City	5	1	1	0	1	2	0	0	10
Western	30	0	0	1	3	2	0	0	36
total	170	5	2	20	18	8	1	3	227

note: 15 tokens in "Region 8" are excluded from this Table.

Regional Distribution of /i/

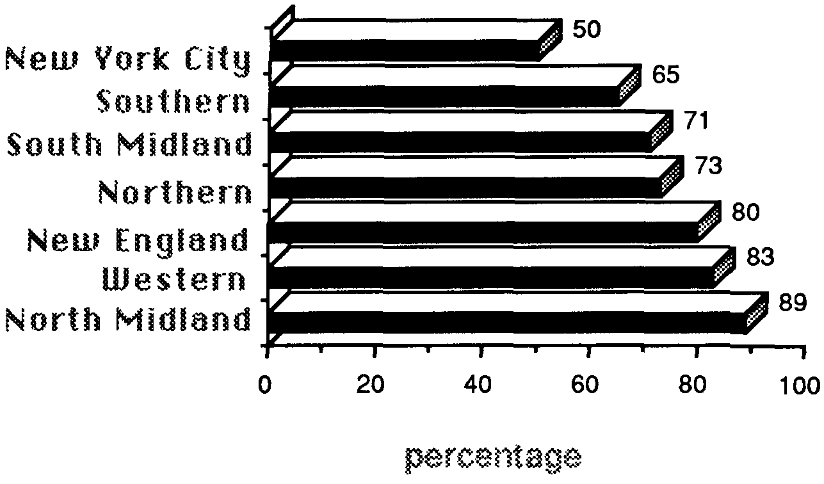


Fig. 2

Fig. 2 indicates the percentage occurrence of [i^{y}] in each region (The above figure shows the 1st divided by the last column in Table 2).

It is found that speakers in region 3 and 7 (i.e., North Midland and Western) had the highest percentage (89 % and 83 % respectively) of the [i^{y}] occurrence, whereas speakers in region 6 and 5 (i.e., New York City and Southern) had the least (50 % and 65 % respectively: refer to Fig. 2). The differences were statistically significant at the 2 % significance level ($X^2 = 11.3$). This finding appears to somewhat contradict Henton & Bladon, and my intuition, that speakers in the Western region (especially, Californians) would pronounce the determiner as a reduced vowel such as [ðə]. However, our intuition may be explained by examining the vowel distribution by age (Fig. 3).

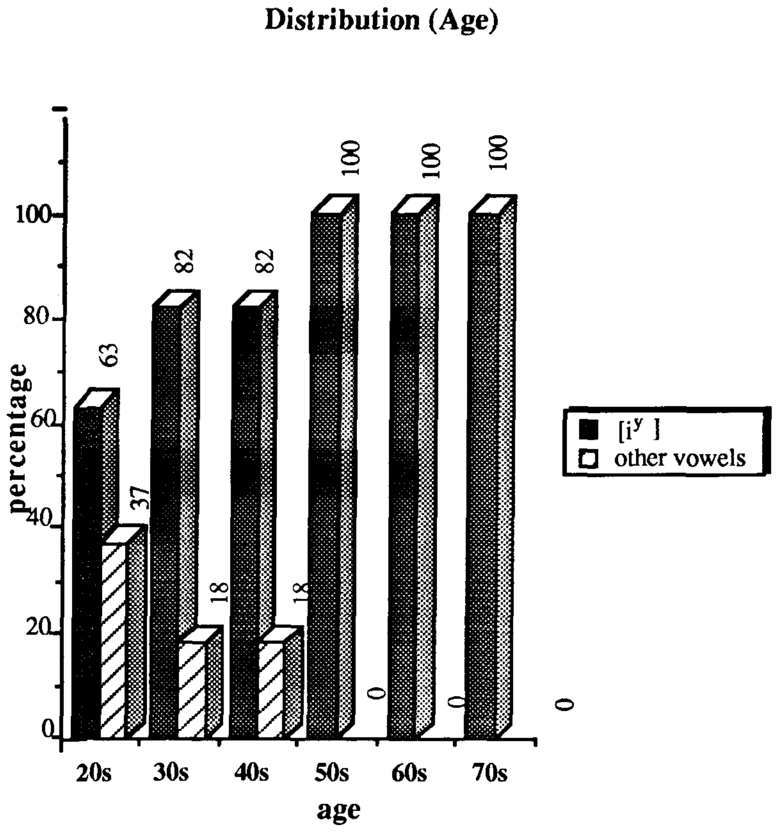


Fig. 3 shows how the use of [i^{y}] in "the" varies Fig. 3

As we can see, no one over 50 years old pronounced the determiner (preceding a vowel) as a reduced one. In other words, all of those speakers (19 speakers) pronounced it as [ði^{y}]. There is also a tendency among the young speakers to pronounce it as a reduced vowel (Fig. 3). A chi-square test indicates that the choice of vowels differ among the different age groups at the 1 % significance level ($X^2 = 13.365$). This finding corresponds to the observation that young Californians (e.g., UCLA and UC Davis students) tend to pronounce the determiner as [ðə] before vowels in unemphatic speech. Even though the description of the phonetic variations in the literature seems to hold true, there is a definite trend among the young that the general distinction, i.e., [ðə] before consonants and [ði^{y}] before vowels, is becoming less obvious. This finding is useful pedagogically. In order to have L2 learners get acquainted with current pronunciation of the English language, new information such as above should be incorporated in the textbooks.

Another interesting observation made in the present study is in respect to the glottal stop insertion rule. As I have mentioned earlier, the description of the rule differed among researchers. Generally speaking, there are two different descriptions of the rule described in the literature. (I have excluded the other possible environments addressed in the literature *in the following generalization*).

1. glottal stop occurs only after the word "the" before a stressed vowel, and that the vowel preceding glottal stop is realized as [ə]. (Trager & Smith, 1951)

2. Glottal stop occurs before word-initial stressed vowel after a syllabic segment (but not after "the"). (Allen et al., 1987)

I found that 65 out of the 242 tokens had an inserted glottal stop after "the". All the tokens (except for two, both "the idiotic") were ones in which the initial vowel of the word following the determiner received primary lexical stress. Fig. 4 shows the regional distribution of the glottal stop, while vowel quality of the determiner preceding the glottal stop is summarized in Fig. 5.

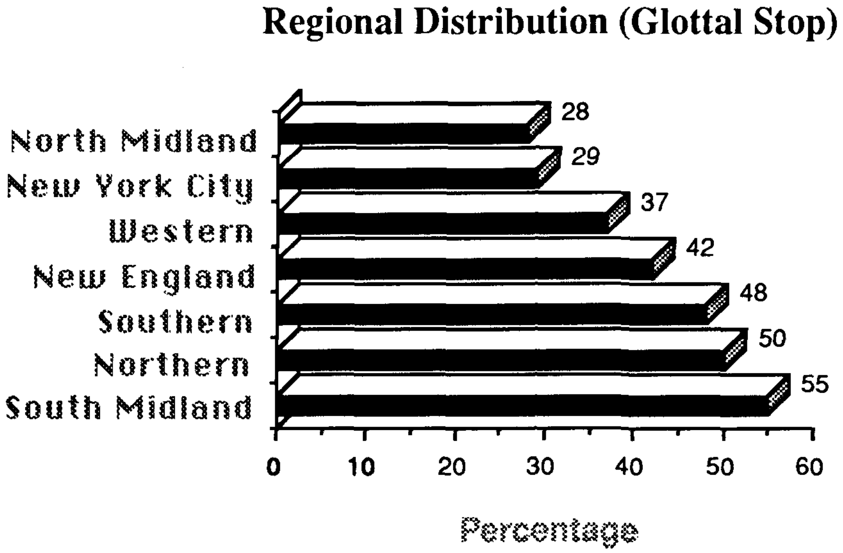


Fig. 4

Distribution of Vowels Preceding Glottal Stop

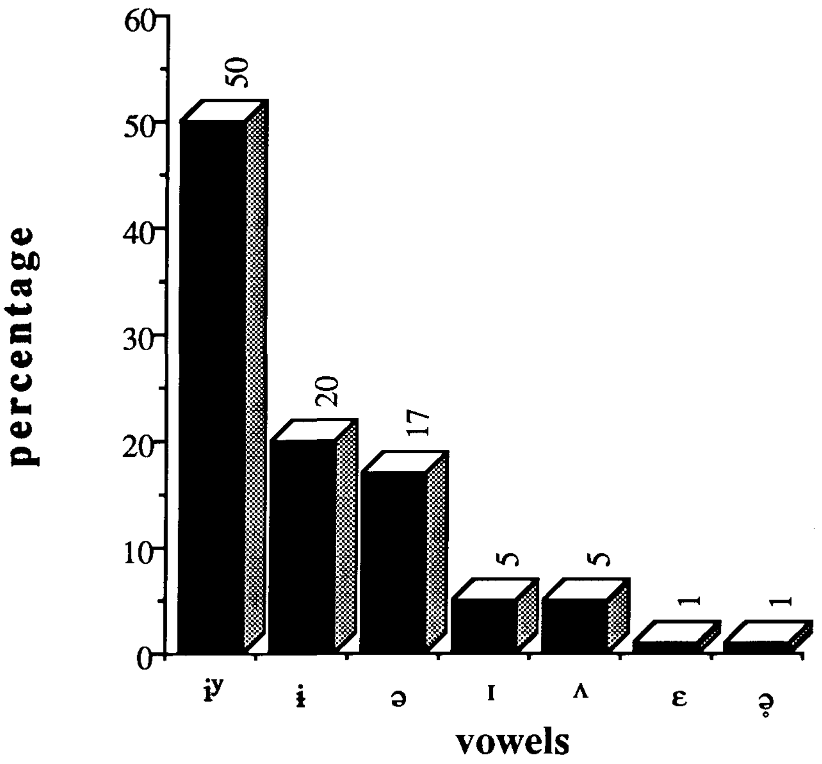


Fig. 5

Thus, my findings in the TIMIT database indicate the following:

1. All the tokens (except for two, both *the idiotic*) of glottal stop occurred before a word-initial stressed vowel (Fig. 4). When glottal stop didn't occur in that environment, 86 % of the vowel of the determiner was [i^y].

2. The vowel preceding the glottal stop is realized as i^y] (50%), [ɨ] (20%) and [ə] (17%).

3. Some regional difference was observed. Speakers in region 4 (i.e., Southern Midland) had the highest occurrence (55 %), whereas speakers in region 3 (i.e., North Midland) had the lowest (28 %).(Fig. 4)

It seems that Trager & Smith's description is in a way correct in that 97 % of the total tokens of glottal stop are in fact observed before a word-initial stressed vowel. However, the vowel preceding glottal stop was not realized as [ə] as was described in Trager & Smith. It was rather realized most often as [i^y] (50 %). (Fig. 5) Then, I compared the percentage occurrence of vowels with an inserted glottal stop to that of the vowels without a glottal stop in the same environments, and found the following.

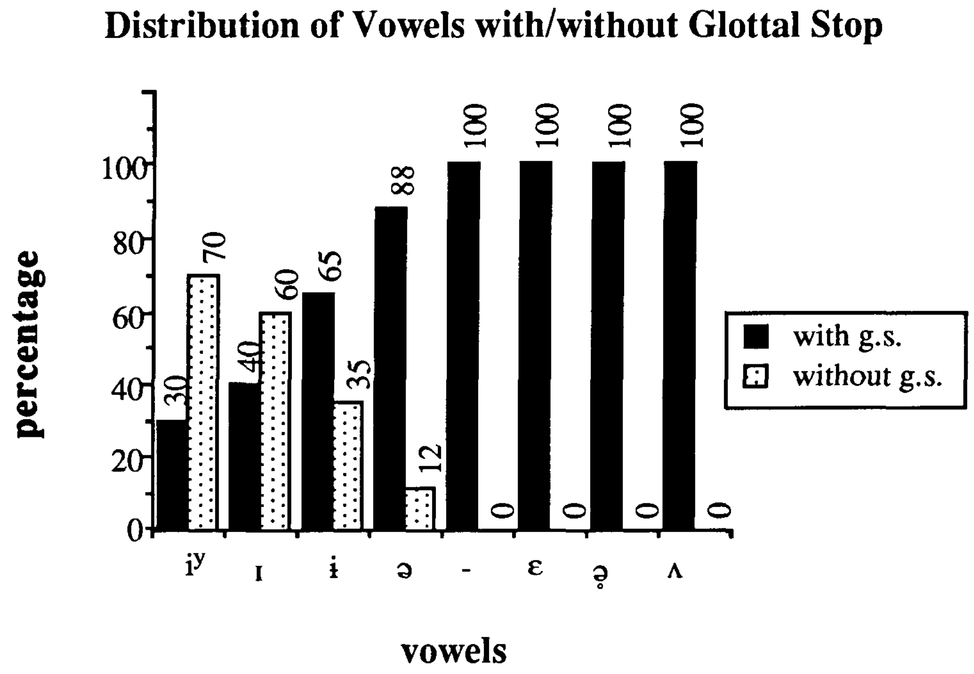


Fig. 6

The above Figure indicates that the occurrence of an inserted glottal stop is much higher when the preceding vowels are reduced ones. I ran a chi-square test ($X^2 = 20.6$, $p<.0005$) to find the relationship between vowels and glottal stop, and found that the observed differences are statistically significant (I only selected [ɨ] and [ə] for the above statistical test, since the tokens of other phonetic variants were less than 5). It means that Trager & Smith's description is correct as a whole. However, we have to keep in mind the fact that the vowel [iʸ] can also precede a glottal stop.

Another point is that speakers in region 4 had approximately twice the occurrence of glottal stop compared with that of speakers in region 3 (Fig. 4). The differences observed were in fact statistically significant at the 5 % significance level.

I also checked the vowel quality of a lexical item following glottal stop. It is generally believed that if the following vowel is high, it is more likely to have the glottal stop insertion. No conclusive evidence was found regarding this notion, however.

Finally, the larger environments in which different vowel variants of the determiner occurred were examined: I tried to classify the different vowels according to syntactic or prosodic boundaries. However no conclusive evidence for a relationship was found in the data. It seems that lexical stress is the main criterion for the phonetic variants.

Final comments

The present study examined the phonetic variants of the determiner "the" when preceding vowel in terms of vowel quality and glottal stop insertion. The following is the summary of the findings:

1. Seven different phonetic variants of the vowel in the determiner were found in the TIMIT database. Of those 7, [iʸ] had the highest frequency (76 %), followed by [ə] (8.3 %) and [ɨ] (7.4 %).

2. Speakers in dialect regions 3 and 7 had the highest occurrence (89 % and 83 % respectively) of [i^y], whereas speakers in region 5 and 6 had the least (65 % and 50 % respectively). The differences were statistically significant at the 2 % significance level ($X^2 = 11.3$).

3. The younger the speakers were, the more frequently they pronounced the vowel of the determiner as a reduced vowel, i.e.,
[ə] or [ɨ]). The above trend was statistically supported at the 1 % significance level.

4. The glottal stop insertion took place before a word-initial stressed vowel (97 %).

5. The occurrence of an inserted glottal stop was much higher when the preceding vowels were reduced ones compared with that preceding the vowel [i^y]. The observed differences were statistically significant ($X^2 = 20.6$, $p<.0005$), even though it is also possible to have [i^y] preceding it.

Within the limited scope of the present study, it was not possible to capture all the phonetic variants of the word "the", i.e., before consonants and under stress; nor was it possible to check all the possible environments for the glottal stop insertion rule. However, the study does provide a foundation on which to base further research in the analysis of phonetic variants of the word "the" and the glottal stop insertion rule.

References:

Allen, J., M. S. Hunnicott, & D. Klatt. 1987. *From the text to speech: the MITalk System.* Cambridge University Press, London.

Bronstein, A. J. 1960. *The pronunciation of American English.* Appleton Century Crafts, New York.

Henton, C & A. Bladon. 1987. Developing computerized transcription exercises for American English. *JIPA* 17: 72-82.

Kenyon, J. T. 1940. *American Pronunciation: A Textbook of Phonetics for Students of English.* George Wahr, Michigan.

Ladefoged, P. 1982. *A Course in Phonetics.* Harcourt Brace Jovanovich, New York.

Lamel, L. F., R. H. Kassel, S. Seneff. 1986. Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus. *Proceedings DAPRA Speech Recognition Workshop*, 100-104.

McLean, M. P. 1930. *Good American Speech.* E.P. Dutton & Company, New York.

Ripman, W. 1924. *The Sounds of Spoken English: with Specimen Passages in Phonetic Transcription, Annotated, and with a Glossary and Index.* Dotton, New York.

West, M. 1953. *A General Service List of English Words.* Longman, Green and Co., London.

Pitch and Duration of Yes-No Questions in Nchufie.

Dani Byrd

This paper will present a preliminary phonetic description of yes-no questions in Nchufie (also known as Bafanji), a Grassfields Bantoid language of the Nun group in the Mbam-Nkam family spoken in Northwestern Cameroon by approximately 8,500 people (Grimes, 1988). As there is no published description of this language, a very brief review of the Nchufie segment inventory and phonology will be in order. Following this a phonetic description of the yes-no questions in the language will be presented, focusing on the prosodic cues of duration and pitch. Of special interest will be the interaction of intonation with lexical tone and the representation and cross-linguistic significance of Nchufie yes-no question formation.

1. Introduction

Summarized below is a preliminary description of the phonetic segment inventory of Nchufie.

Consonants. The language has voiceless aspirated and unaspirated oral stop consonants at the bilabial and dental places of articulation and a velar voiceless aspirated stop. There are prenasalized (voiced) stops at each of these places, and nasals at these places in addition to a palatal nasal. There is also a voiceless alveolar affricate and its prenasalized counterpart. Labialized velar stops also occur. The language has a lateral dental liquid, voiced and voiceless alveolar and labio-dental fricatives, a voiceless velar fricative, and a palatal and a labio-velar glide. (A palato-alveolar fricative alternates with the alveolar fricative before high vowels.) The velar nasal is the only coda consonant permitted.

Vowels. The language has three vowel heights and front, central, and back places of articulation. The back vowels are rounded and the front and central unrounded. These vowels will be transcribed: i, e̞, a; ɯ, ə, ɑ[1]; u, o, ɔ. ([i] alternates with [ɪ] in closed syllables). Vowels may be contrastively nasal and oral, long and short, and, if long, creaky and modal. Only short vowels may occur in a closed syllable, and the diphthongs [uɔ], [ɑi], [ɯə], and [ie] occur.

Tones. Four tones occur underlyingly on lexical items: H, L, HL, and LH. Mid[2] and Superhigh arise in context, and downstepped Highs also occur.

A language's yes-no questions may be marked morphologically, syntactically, intonationally, or by some combination of these mechanisms. When marked intonationally, this kind of question generally has either a terminal rise or a higher overall fundamental frequency. Ultan (1978) found in a sample of yes-no questions in 53 languages that 71.7% had a rising intonation and 34% an overall higher F0 (Shen, 1991). Only 5.7% had a falling contour (Ultan, 1978). Nchufie provides a fascinating type of yes-no question formation. In Nchufie, yes-no questions are not marked segmentally but rather marked by intonation and localized temporal lengthening. I will suggest that Nchufie presents a case of question marking which is very unusual in the world's languages. First, final lengthening is the most salient cue marking a yes-no question. We will see that the final rhyme in the question domain is lengthened almost 70% as compared to the declarative statement. Secondly, Nchufie shows two intonational mechanisms at work. In the question, a higher pitch range occurs which differs from a final rise or high raising (upstep), or overall higher F0. Final lexical low tones are not effected however by this raising. Most interestingly, in the question, we see an optional final lowering of phrase final high tones. I will discuss the interaction of this intonational low tone with underlying lexical tones. While this is a preliminary study on a language which has never been studied previously, the data to be presented suggest that Nchufie is an important language in the investigation of universal regularity and variability in the formation of yes-no questions. Finally a brief discussion of apparently similar data in Hausa and topics requiuring future research will be presented

2. Method

Data Collection. Monosyllabic, modal voice nouns were selected as target words and inserted in the carrier phrase /ɑ̄ ɣé ______ /; 'He has a ______ /Does he have ______ ?' spoken with statement intonation and question intonation. The nouns included all four underlying lexical tones: H, L, HL, and LH, and all possible rhyme types: CV, CV_1V_2, CVN, and CV_1V_1. The experimental target words are shown in Table 1.

TABLE 1. Target words

	CV	CV_1V_2	CVN	CVV
H	tʰɯ 'tree' n͡ɟʷi 'cloth mɑ 'mother'	muɔ 'fire'	kɯŋ 'ant' 'kɔŋ 'bed' sɔŋ 'friend' ŋʷɪŋ 'machete	lɑɑ 'lamp'
L	ncɔ 'mouth' ŋ͡gɔ̃ 'stranger' ŋkɔ 'box'	fuɔ 'chief' fuɔ̃ 'key	pɯŋ 'stomach' 'nɪŋ 'animal	kɔɔ 'foot' kɯɯ 'barn'' pẽe 'bag'
LH	ŋki 'water' kɯ 'pot' ŋkũ 'back'	fuɔ 'leaf' fuɔ 'medicine'	ŋ͡gɯŋ 'corn' n͡ɟɪŋ 'brother' zɪŋ 'brothers'	ŋ͡gʷɔ̃ɔ 'plaintain'
HL	n͡ɟə 'egg' li 'name' pʰu 'ashes' ŋkʷe 'firewood'	pʰuɔ 'arm' ŋkɯə 'bush' n͡die 'neck'	pɯŋ 'breast' m͡bɯŋ 'money' lɯŋ 'tongue'	mee 'child' lii 'eye' ɲɔ̃ɔ 'thing' kɯɯ 'dish' too 'ear'

A fully balanced set could not be obtained with the lexical data collected thus far in the study of the language within the constraints of monosyllabicity and modal voice. These constraints were applied in order to minimize variation in the interaction of lexical and intonational tone and to simplify automatic pitch tracking.

The 43 sentences were randomized and recorded in blocks first as declarative statements and then as yes-no questions with the speaker being prompted by a written English translation of each sentence. The sentences were recorded in blocks of statements and questions so that no overt contrast effects were produced by the speaker. A single female speaker raised speaking Nchufie in Cameroon was recorded in a sound insulated booth. Each target word was recorded only once in each condition as the intended independent variables in evaluating F0 and duration were lexical tone and rhyme type.

Acoustic Analysis. Pitch tracking was done on the digitized sentences using the Kay Computer Speech Lab (CSL) package. Pitch synchronous pitch tracking was used by first employing the automatic peak picking capabilities of CSL which marks the division between each voicing impulse in the waveform immediately following positive-going zero crossings that precede the first positive-going amplitude peak of the voicing impulses. The process separates the voiced signal into its periodic components, the inverse of each period being the fundamental frequency (F0) of the signal. Peak picking was done with a 25 ms window and a 20 ms frame advance with a specified range of 50 - 300 Hz. The minimum peak threshold was lowered for target rhymes ending in a coda nasal or a nasalized vowel due to the decreased amplitude. However, the same threshold was always used for a word for both statement and question. Impulse marks were checked by hand and missing divisions added. The pitch tracking function of CSL extracts the fundamental frequency values by computing the inverse of the time between each marked peak. The pitch track of the entire rhyme of the target word was recorded and smoothed using three point smoothing. A single measure for F0 in each word of the carrier phrase at the point of highest amplitudes was also obtained using a simultaneously displayed energy plot. . This procedure provides measurements relatively independent of the effects of adjacent consonants

The duration of the rhyme of the target word was measured in both sentences from the onset to the end of vocalic voicing. The duration of the carrier phrase [ɑ̄ ɣé] "he has" was also recorded as a control across the statement and question conditions. This measure was made from the onset of voicing up to but not including the onset consonant of the target word. Segmentation was done by waveform examination.

Statistical analyses included paired t-tests, correlations, and analysis of variance. Post-hoc Scheffé F-tests were used for pairwise comparisons.

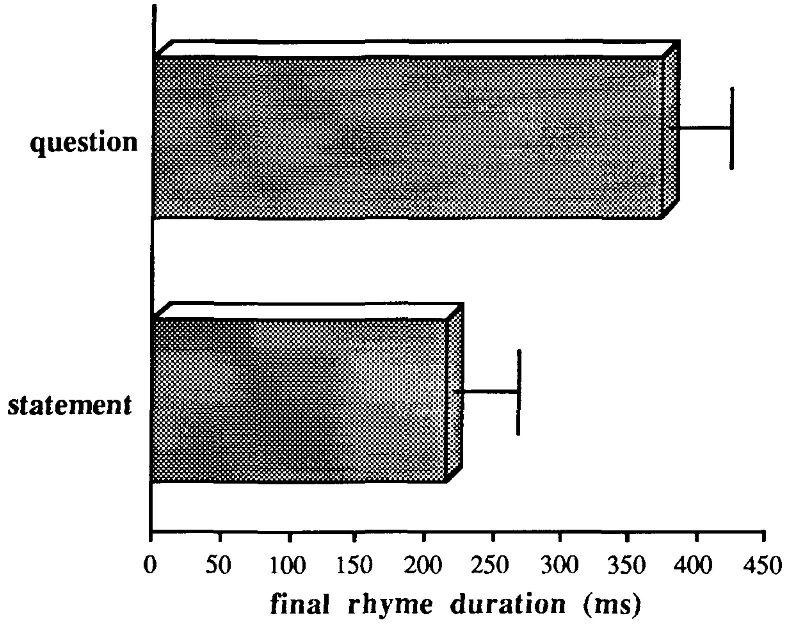
3.0 Results

Lengthening in questions. In order to test the anecdotally observed final lengthening in questions, it first needed to be determined that the speaker maintained the same carrier phrase rate in both conditions. A paired two-tail t-test was conducted on the duration of the carrier phrase in statements and questions. A paired test was used to factor out any effects specific to the onset consonant of the target word on the preceding vowel of the carrier phrase. The mean carrier phrase durations were 311.1ms for statements and 312.4ms for questions. A (two-tail) paired t-test showed there to be no effect of statement vs. question on the rate at which the sentence was spoken: t=-.219, DF=42, p=.8281.

There being no effect of statement vs. question on rate, the rhyme durations could be compared directly for each condition. A paired one-tail t-test showed there to be a significant main effect of statement vs. question on the duration of the final rhyme in the sentence: t=-156.767, DF=42, p=.0001. The difference between the mean final rhyme duration in statements and in questions is 156.8ms; questions being 72.7% longer than statements. This main effect can be seen graphically in Figure 1.

Figure 1

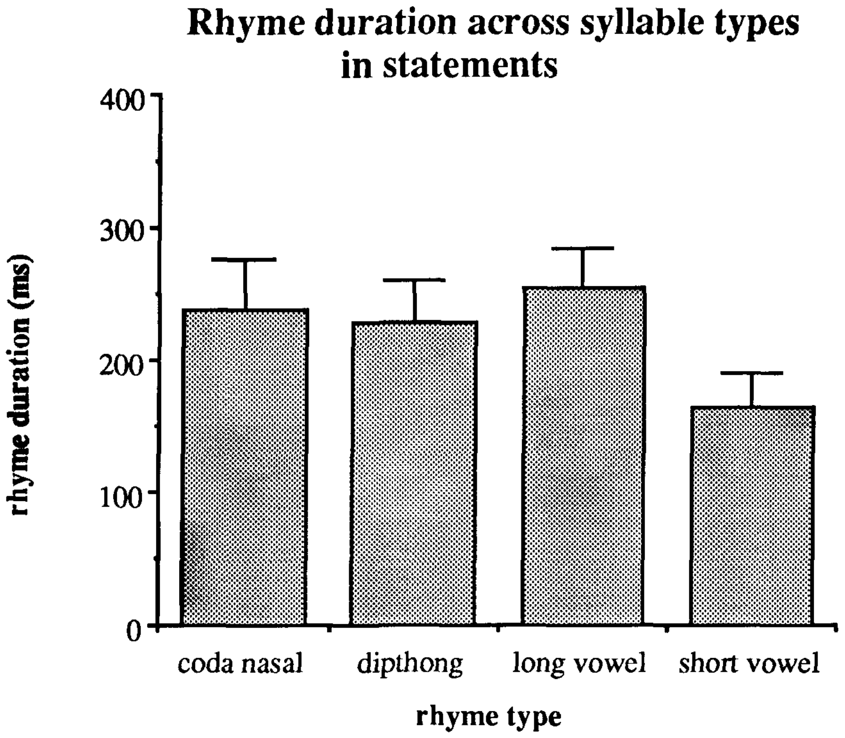
Main Effect-Lengthening of Final Rhyme in Questions



The duration of final rhymes in statement and question sentences. Standard deviations are shown by error bars. The difference between the statement and question values is 156.8ms or 72.8%; p = .0001.

Next, we need to ask what effect, if any, the four experimental syllable types have on lengthening. In order to do this, it needed to be determined what syllable types were significantly different from one another in rhyme duration in the *statement* condition as the initial divisions into CV, CV_1V_2, CVN, and CVV were arbitrary with respect to duration. A one factor ANOVA showed there to be a significant effect of syllable type on rhyme duration in statements: F(3,42)=17.838, p=.0001. Post-hoc pairwise comparisons using the Scheffe F-test showed there to be no significant difference at the 95% level in rhyme duration between the diphthong, the closed syllable, and the long vowel syllable types. The CV syllable was significantly different at the 95% level from the other three syllable types. The mean rhyme duration in statements for each syllable type can be seen graphically in Figure 2.

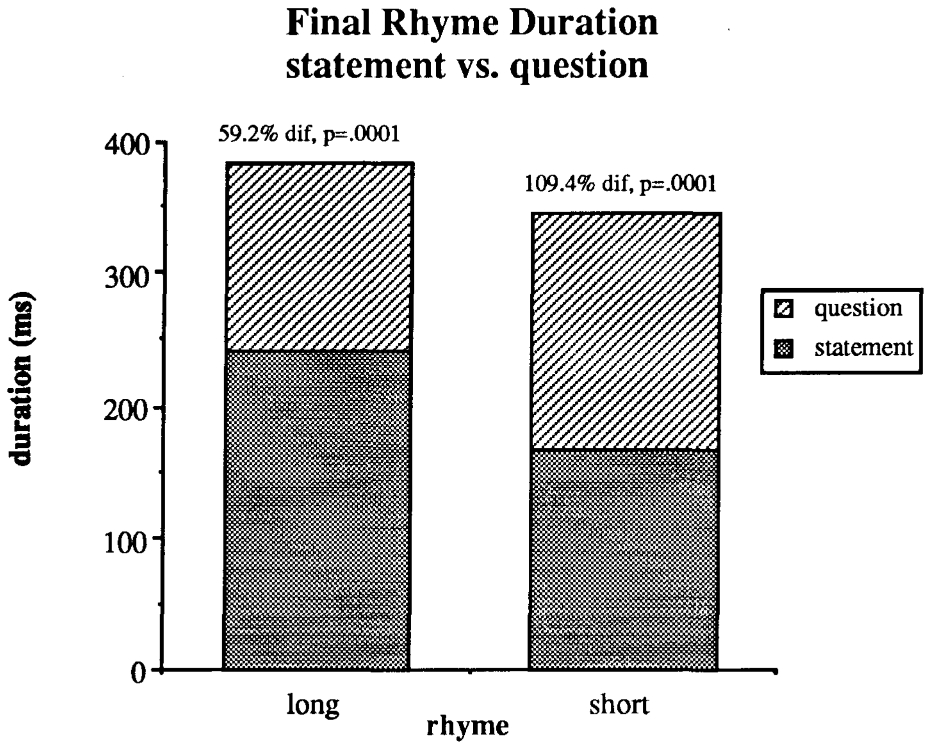
Figure 2



Final rhyme duration as a function of syllable type in statements. Standard deviation shown by error bars.

As expected, the categories of long versus short rhyme have a significant effect on rhyme duration in statements: F(1,42)=48.81, p=.0001. In light of this analysis, all subsequent analyses involving the categorical variable of rhyme type will be made using two types: long and short, where CV syllables are referred to as short and CV_1V_2, CVN, and CVV as long. The long rhymes have a mean duration of 240.7ms in the statements and the short rhymes 164.8ms in the statements. Mean rhyme durations in statements for these types is shown by the shaded areas in Figure 3.

Figure 3

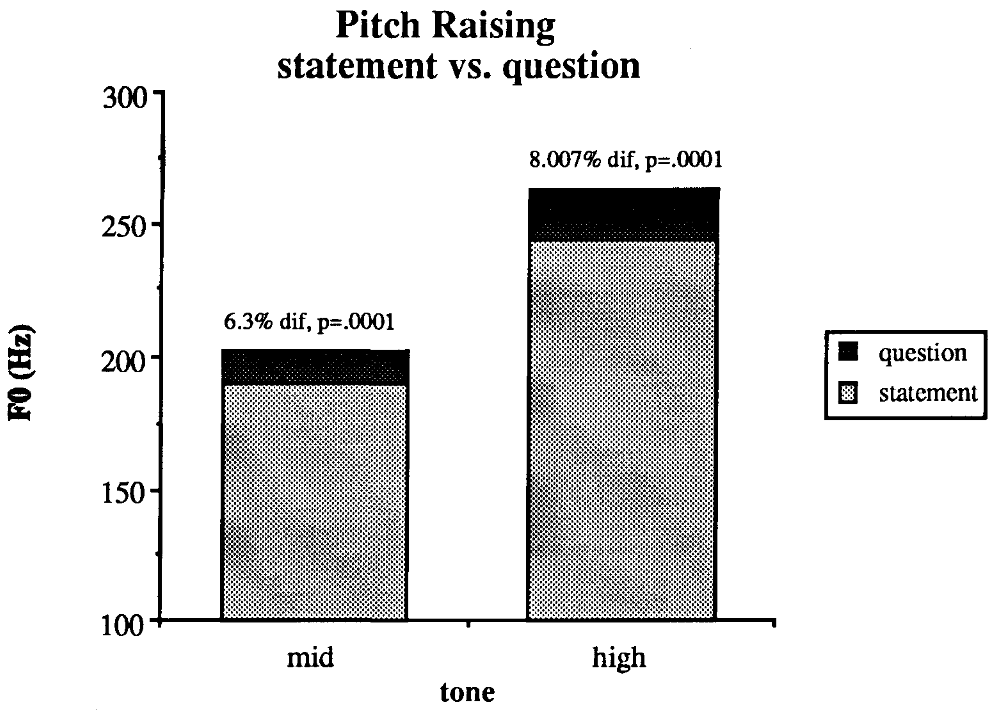


Final rhyme duration in statements and questions as a function of rhyme type.

While all final rhymes are lengthened in a question as compared to a statement, there is a significant interaction of duration with rhyme type in questions as shown by a two-factor ANOVA: F(1,82)=4.6113, p=.0347. Short rhymes lengthen by 109.4% in questions, and long rhymes by 59.2%. This lengthening is shown in Figure 3. Although short rhymes lengthen by a considerably larger percentage than long rhymes, the underlyingly long and short rhymes are still significantly different in length even in questions. A one-factor ANOVA shows there to be a significant effect of rhyme type on rhyme duration in questions: F(1,42)=7.544, p=.0089. Long rhymes exhibit a mean duration of 383.2ms (sd=43.5ms) in questions, and short rhymes a mean duration of 345.1ms (sd=37.2ms). Although long rhymes are lengthened in questions less than short rhymes , the underlying lexical length distinction in (final) rhymes is not neutralized. It appears rather that the long rhymes are displaying a ceiling effect which prevents them from lengthening by the same amount as the short rhymes.

Pitch Range. Two F0 measures were made of the carrier phrase in both the statement and question condition: one of the mid tone on the first word and one of the high tone on the second. A one factor ANOVA shows there to be a significant effect of statement vs. question on the mid tone: F(1,85)=16.685, p=.0001. The questions have a significantly higher F0 for the mid tone in the carrier phrase than statements: 201Hz (sd=18Hz) vs. 189Hz (sd=6Hz). This is a difference of 6.3%. Likewise, the carrier phrase high tone is significantly higher in the questions than in the statements: F(1,85)=67.361, p=.0001. The question high has a mean F0 of 263Hz (sd=13Hz) and the statement a mean F0 of 243Hz (sd=9Hz); a difference of 8%. This question raising is shown graphically in Figure 4.

Figure 4



Pitch raising in statements and questions for carrier mid and high tones.

The H tone on target nouns occuring at the end of the phrase were raised by a like amount compared to the highs in the carrier phrase. The mean maximum F0 value for these words was 276Hz (sd=9) in questions vs. 258Hz in statements (sd=13Hz). This is a difference of 7%. The overall mean F0 value for H target words (which also averages in tokens where optional final lowering took place as will be discussed below) was 261Hz (sd=15Hz) in questions vs. 246Hz (sd=10Hz) in statements; a difference of 6%. While there was no low tone word in the carrier phrase, the low tone target nouns were examined to determine if low tones were realized differently in statements than in questions. Four measures were considered for all low tone target nouns: overall mean F0, minimum F0, endpoint F0, and maximum F0. One-factor ANOVAs showed that the statement vs. question condition had no effect on mean F0, minimum F0, and endpoint F0: F(1,19) p=.3342, p=.94, p=.8357 respectively. Maximum F0 was significantly higher in questions F(1,19), p=.0107, but this can probably be explained by the fact that the target noun was falling from a preceding high tone which was raised significantly in the questions. All of these pitch tracks show a downward fall from this preceding high. In summary, this experiment suggests that raising in the questions occurs over the entire clausal domain and affects highs more than mids, and lows not at all.

The raising of mids and highs in the question phrase, highs more than mids, suggests that pitch range expansion rather than upstep or overall register shift is at work here. Lows are not raised, and highs are raised by a greater percent than mids. Upstep generally refers to the raising of highs only, and an overall register shift or key raising would produce a raising of all lexical tones.

Before we can firmly assert that pitch range expansion is at work, we need to determine if for some reason the lexical low tones on the final target word are not undergoing raising whereas a low tone earlier in the sentence would be raised. There is evidence to suggest that the final pitch of a declarative phrase is less variable than other peaks in the phrase (Boyce and Menn, 1979). This suggests that last syllables may be under-informative when examining F0 scaling. For this reason, a second experiment was conducted in which a low tone occured at the beginning of the carrier phrase.

In experiment two, the carrier phrase */ŋ͡gɔ̀ ɣɛ́* ________/; 'The stranger has ________/Does the stranger have ________?' was used. The Nchufie sentence was . The high and low tone nouns from table 1 were used as the target words, yielding a total of nineteen statements and nineteen yes-no questions, eight with high final nouns and nine with low final nouns. The pitch of each word in the carrier phrase was tracked according to the methods outlined in section two. The pitch of the final noun was recorded at its endpoint.

Two-tailed paired t-tests were conducted to test a difference in pitch between statement and question for the low and high tones in the carrier phrase and for the lexical tone of the target word. The high tone of the carrier phrase was shown to be significantly lower in statements than in questions, t=-13.078, DF=18, p=.0001. The average high tone in the statements was 205Hz (sd=7Hz) and in the questions was 228Hz (sd=7Hz). Importantly, the low tone in the carrier phrase was also shown to differ significantly in the two conditions, t=-13.459, DF=18, p=.0001. The average low tone in the declarative carrier phrase was 168Hz (sd=3Hz) and in the question phrase was 182Hz (sd=4Hz).

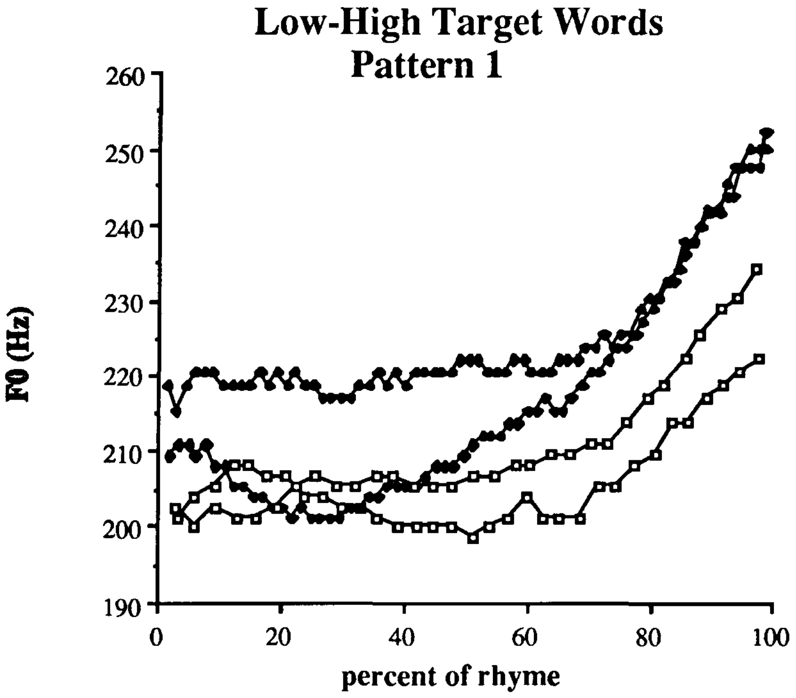
In the target words, the results of experiment two show the lexical highs to be raised significantly in the questions, t=-6.872, DF=8, p=.0001, and the lows *not* to be significantly different, t=1.118, DF=9, p=.2923. This result is in accordance with experiment one in which low toned target words were not raised in questions. However, the results showing a change in the low tone of the carrier phrase suggest that an overall higher pitch range is used in the yes-no questions than is used in the segmentally identical statements. Lexical low tones occurring phrase finally appear not to undergo this raising however.

Optional Final Lowering. If individual tokens of the LH and H, i.e. lexically high-final, target nouns are examined, one observes that in the questions certain of these tokens have a final lowering or a lowering throughout the rhyme which causes the F0 contour in the question to fall even with or below that of the statement. No statements ending in LH or H nouns showed this final lowering.

We have seen in experiment one that the endpoint, minimum, and overall mean for low target words does not differ between statement and question. However, the HL nouns do show a significant difference in both endpoint (F(1,29)=12.344, p=.0015) and in minimum (F(1,29)=12.824, p=.0013) as a function of statement vs. question with questions being significantly lower on both measures. It is unclear however whether this difference is due to any final lowering operating in questions or merely a by-product of the question rhyme being longer in duration than the statement thereby allowing it more time to reach a lower target value in the fall from the initial H of the contour. Because these possibilities can't be evaluated in this dataset, the discussion of final lowering in questions will focus on the LH and H nouns where final lowering in questions can be clearly distinguished due to the presence of an underlying final lexical high.

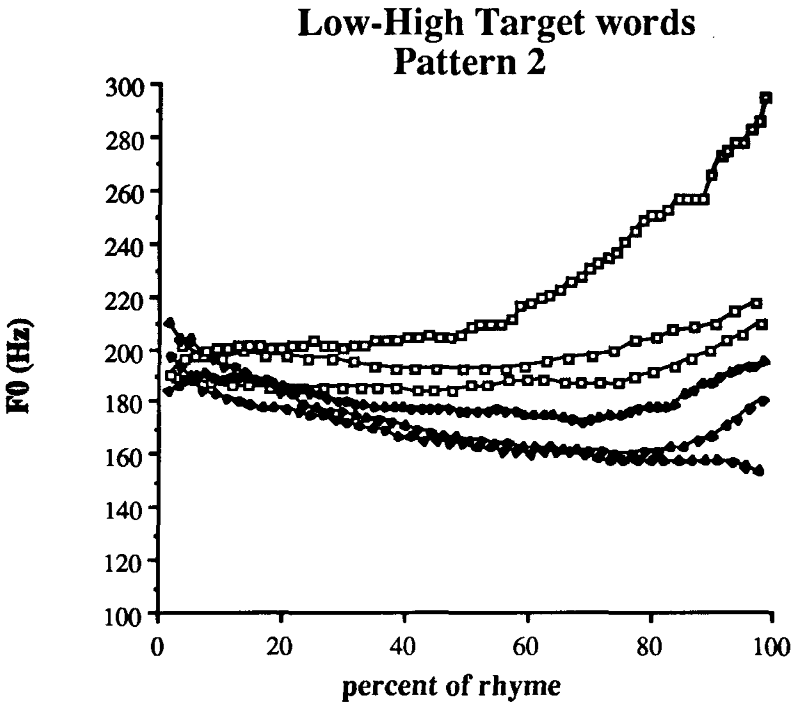
The two patterns found in the LH case can be seen by comparing Figures 5 and 6. In pattern one, the final H in the question condition ends higher than in the statements. In pattern two the entire contour in the question condition is lower than that of the parallel statement. This is contra the general pattern of raising in questions. In pattern two of the LH cases, all the questions end lower than all the statements. Other tokens show fairly identical LH contours for the statements and questions.

Figure 5



Pattern one pitch contours for LH final rhymes. Questions are shown by filled symbols, statements by open symbols.

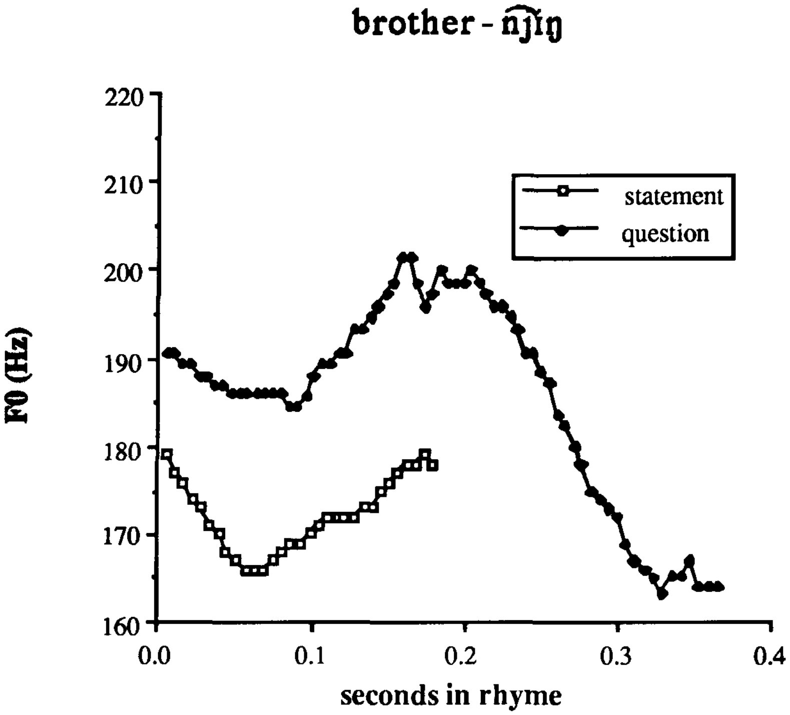
Figure 6



Pattern two pitch contours for LH final rhymes. Questions are shown by filled symbols, statements by open symbols.

In one case, ŋ͡ɟɪŋ brother, the statement contour shows a LH pattern while the question contour shows a LHL pattern. This can be seen in Figure 7.

Figure 7



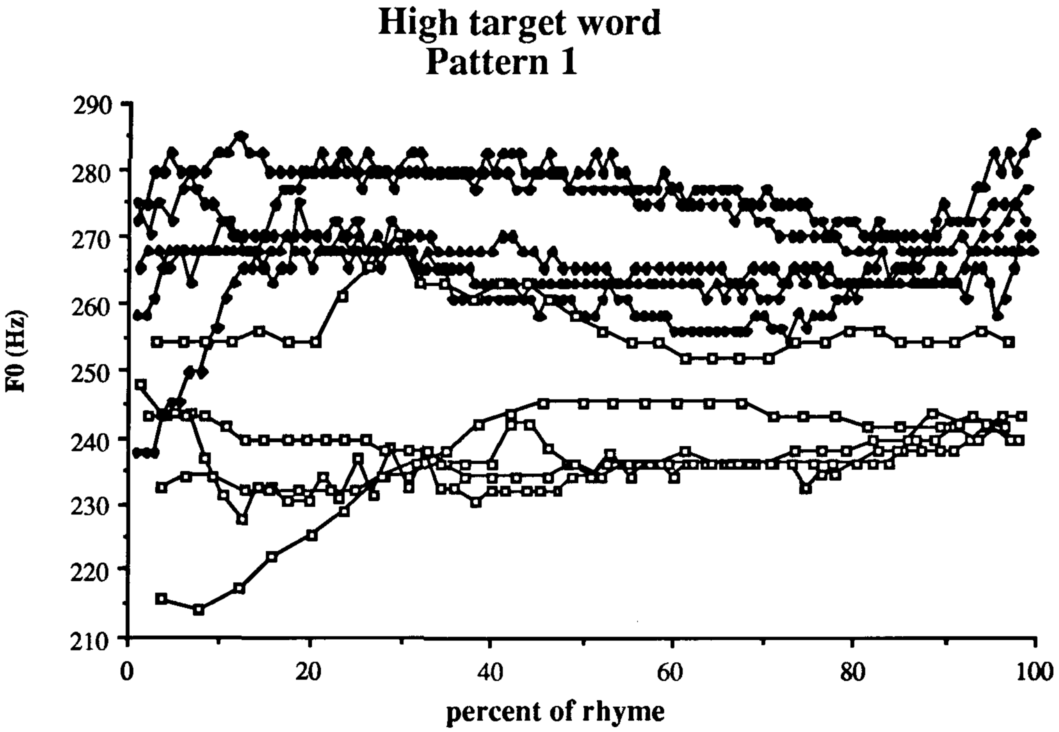
Pitch contours for the word n͡jɪŋ. Questions are shown by filled symbols, statements by open symbols.

The final fall in the question is aligned precisely to occur where the final high ends in the statement condition. This suggests that a final L is optionally added along with the extra length in the question condition. If not added, the endpoint of a phrase final lexical LH in a question is raised above that of the statement level as is expected due to the higher pitch range used in questions. If the final L is added, it may act to lower the entire preceding contour, or it may simply cause a final lowering after the predictably raised contour.

It can be shown that the addition of the final low is optional by considering that the token for *medicine* showed no final lowering while the token for *leaf* showed a pattern two output. These two English words are translated by the same word in Nchufie; identical segmentally, tonally, and in noun class. However, this word was elicited twice, once using each English meaning, yet the word conformed to different patterns for each elicitation. Likewise, the word for *brother* and the word for *brothers* which differ in only the onset consonant show final lowering in the case of *brother* and no final lowering in the case of *brothers*. This suggests that the addition of the final low in questions is optional and not determined by segmental, tonal, or class qualities of the noun.[3]

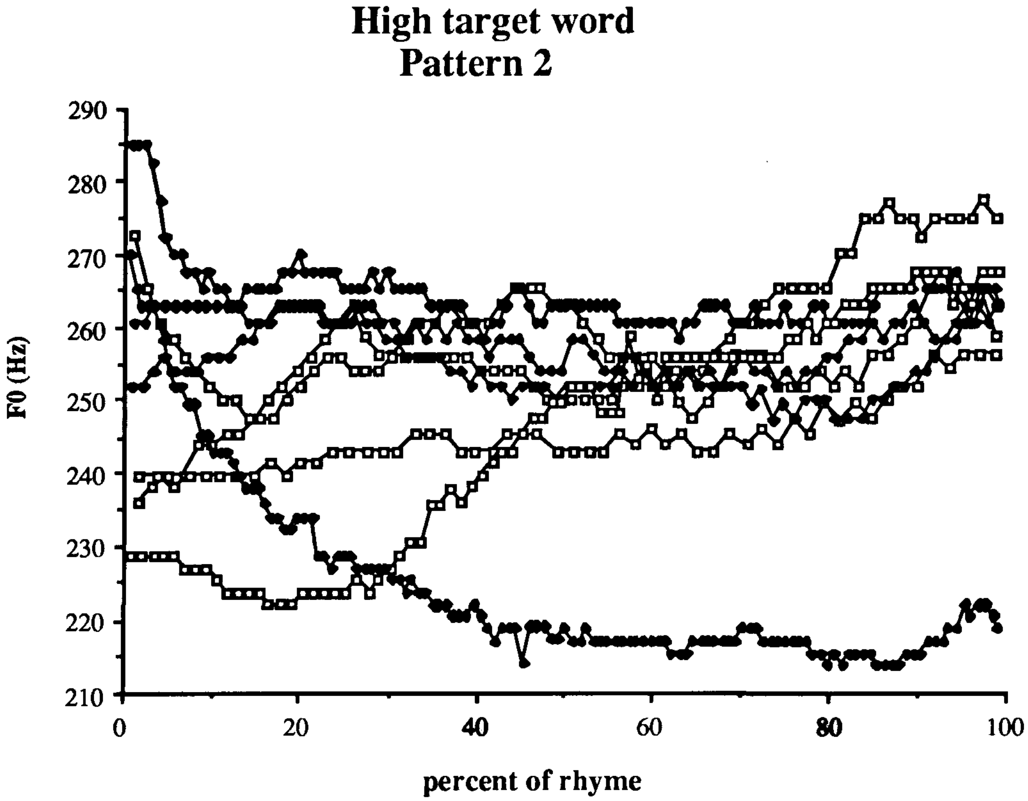
High final nouns show a similar duality in patterning. In pattern one as seen in Figure 8, each statement/question pair shows the question condition to have a higher pitch contour throughout the final rhyme than that found in the statement. Figure 9 shows the second pattern where the normally raised high of the question falls to the same level as the statement or where a final lowering causes the end of the rhyme in the question condition to fall below its level in the statement.[4]

Figure 8



Pattern one pitch contours for H final rhymes. Questions are shown by filled symbols, statements by open symbols.

Figure 9



Pattern two pitch contours for H final rhymes. Questions are shown by filled symbols, statements by open symbols.

To determine if speaking rate (i.e. carrier phrase duration) or rhyme duration are correlated with the appearance of the optional final low, the pattern one and two groups for the H and LH nouns were recoded as integer values. R^2 values were calculated for carrier phrase duration and rhyme duration. Neither rate nor rhyme duration were correlated with the appearance of a final low; R^2=.0001 and R^2=.022 respectively.

3.1 Summary of Results

In summary, it has been shown that yes-no questions in Nchufie are marked by several cues. First, question prosody is characterized by a specific duration cue. The final rhyme of the phrase is lengthened substantially, more than doubled in the case of CV syllables. The amount of lengthening is dependent on whether the rhyme is underlyingly long or short with long rhymes showing a ceiling effect causing them to lengthen by 59% as compared to the short rhymes' 109%. Secondly, it has been shown that a higher pitch range is used in questions, causing high, mid, and low tones to raise in the phrasal domain but not affecting phrase final lexical low tones. Finally, we have seen that the lengthening of the final rhyme in yes-no questions may optionally be accompanied by a final lowering. If present, this low tone may have the effect of lowering the entire contour of the final rhyme or adding a final fall in F0 at the end of the rhyme. Whether this low tone appears or not is not predictable in this dataset from the melodic, tonal, or morphological nature of the word nor from the rate of speaking or length of the word's rhyme.

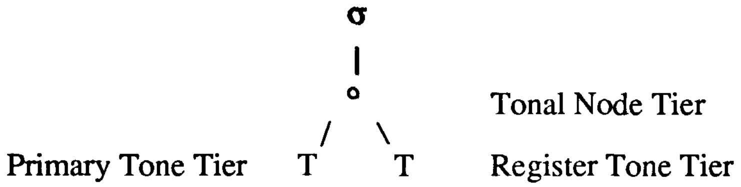
4.0 Discussion

The intonational systems of tonal languages have not been very thoroughly investigated. However, Hausa is an exception to this generalization with respect to data regarding the formation of yes-no questions. As a cross-linguistic comparison to the Nchufie data, I would like to review what has been claimed to occur in the formation of yes-no questions in Hausa. While some of the Hausa facts appear similar to the Nchufie facts described above, I will point out major differences between the two systems.

Hausa, a language with only lexical H and L tones, forms yes-no questions by solely prosodic means like Nchufie. Hausa has been shown to have a locally raised final high tone in yes-no-questions (Cowan and Schuh, 1976, Miller and Tench, 1980, and Lindau 1986). Newman and Newman (1981) have used the term 'key raising' to refer to the systematic upward shift of final H and L tones which also occurs in questions. It has been suggested by Newman and Newman (1981) that the final raised high may optionally be followed by a low tone question morpheme which includes length. The addition of this final low tone neutralizes the distinction between H and HL lexical tones in final position (Inkelas and Leben, 1990). Newman and Newman (1981) have claimed that lengthening takes place on final short syllables in questions. Inkelas and Leben (1990), Schuh (1978), Newman and Newman (1981), and Lindau (1986) have found that downdrift is suspended in questions. Inkelas and Leben (1990) have suggested that the reason for the suspension of downdrift is that downdrift is incompatible with key raising. They claim that key raising is the attachment of a H tone to the register tone tier which raises the tone attached to the primary tone tier. It is claimed that this is incompatible with downdrift which is the insertion of a resister low tone. Lindau (1986) has shown however that Hausa yes-no questions are not marked by a raised register but significantly by the global suspension of the statement downward slope to zero slope, and a local feature of considerably raised pitch of the last high tone. The slopes of the F0 grids for the statement and question pivot around the same starting F0. However, the question grids show a zero slope which contrasts with the statement downward slope. The width (range) of the grid is no different in statements and questions, and the grid is in no sense raised for questions (Lindau, 1986). Lindau (1986) also found that the final fall observed by Newman and Newman (1981) in questions is actually an optional rule of ending low which is not specific to questions but also occurs in statements. In her data, the same speakers who fell finally in questions also fell in statements. Thus the optional low is not part of the question morpheme. Finally, Lindau (1986) found that questions were 10% *shorter* than statements. This shortening was an overall, non-local shortening. She notes that consideration of durational differences in questions and statements is not part of the general literature on the topic of question formation. If this shortening is specific to Hausa, it will have to be part of the rule of question formation in Hausa (Lindau, 1986). Lindau notes that Bannert's (1983) study of German found no significant differences in the duration of statements and questions.

The Nchufie facts appear to differ from the Hausa data in several ways. First, there is no non-local durational differences between Nchufie questions and statements. Second, local lengthening in questions appears to affect both final short and long vowels, not just final short vowels as in Hausa. Third, an overall higher register does appear to be used in Nchufie questions. While an overall higher register is not the most common type of yes-no question intonation, 34% of Ultan's (1978) sampled languages used some form of raised pitch other than a final rise. Some of the languages using this type of yes-no question intonation include Swedish (Hadding-Koch, 1961; Gårding, 1979; Bredvad-Jensen, 1983), Mandarin (Shen, 1990), and Sango (Samarin, 1967). Fourth, a zero slope F0 contour can't entirely explain the higher F0 in questions in Nchufie as the starting fundamental frequency in questions and statements was shown to be significantly different. Fifth, although a single speaker is considered, the optional low tone seen in questions was never seen in statements suggesting that this is a process which is part of yes-no question formation in Nchufie, not a general characteristic of the language or this speaker.

Speculation as to how the yes-no question intonation of Nchufie can be represented phonologically is presented below as a means of introducing some of the relevant representational questions raised by this data. The representation of key raising and downstep outlined by Inkelas and Leben (1990) and Hyman (1985, 1986) provides for a Primary Tone Tier and a Register Tone Tier which are linked to a syllable by an intermediate Tonal Node Tier. This structure is shown below.



(from Inkelas and Leben, 1990)

A high Register Tone will raise a Primary Tone and a low Register Tone will lower or downstep a Primary tone. I will assume for the sake of this discussion that a Tonal Node may be attached to only one Primary Tone and that falling and rising contours are created by adjacent Tonal Nodes which differ in their Primary Tone specification. This implies by extension that a syllable may be attached to more than one Tonal Node. In Nchufie yes-no questions, a H register tone will be added to all tonal nodes within the relevant intonational phrase. This will be realized as a globally higher pitch during phonetic implementation. The final lengthening can be represented as the addition of a mora to the final rhyme of the phrase in the Discourse Phonology , i.e. the phonological component in which the formation of intonation contours and other discourse related phenomena take place (Ken de Jong, personal communication). The phonetic interpretation of this additional mora will result in final lengthening. Note that no restrictions on the number of moras in a syllable hold at this level of the phonology. Finally, recall that the optional final low sometimes produces an abrupt final fall and sometimes has the effect of lowering the final H Primary Tone, neutralizing in that rhyme the pitch range expansion occuring in the question. The final abrupt fall could be represented as the case in which the floating low tone attaches as a *Primary* Tone dependent on the rightmost Tonal Node of the phrase. (This leaves open the question of whether it attaches to a tonal node created by lengthening or whether it projects its own tonal node.) The lowering of the final Primary high tone could be captured in a representation in which the floating low attaches as a *Register* Tone dependent on the rightmost Tonal Node of the phrase, displacing in the process the high register tone which would otherwise be present in question intonation. Crucially, this approach does not explain while final lexical low tones do not undergo raising. The description outlined above is only one of several ways in which the prosodic elements of Nchufie yes-no questions could be represented and is included not as a definitive representation but rather as a means of suggesting some of the relevant representational questions engendered by this and other similar data.

5.0 Future Research

One would like to find some explanation of the appearance of the optional final low tone which occurred in some questions. It has been suggested (Hayes, personal communication) that a difference in meaning might be signaled by the final low. This would have to be a connotation that doesn't occur in statements, as we never saw the final lowering in statements. Other interesting topics not completely explored here include the syntactic and/or phonological domain of question raising, the interaction of question intonation with downdrift and downstep, and the comparison of raised question highs with the syntactic superhigh which marks the past tense in Nchufie. Work on Hausa by Inkelas and Leben (1990) and Hyman (1985, 1986) suggest that the raised highs in Nchufie questions might be represented by a phonological structure linking a high register tone and a high primary tone. As this system of representation offers no obvious way of distinguishing the syntactic superhigh from the raised intonational high, it is significant whether these tones are empirically different. The representation would additionally predict that superhighs in a yes-no question would not undergo any raising and that downdrift will not occur within the the intonational phrase of the question.

6.0 Conclusion

The results described above extend our typology of prosodic marking in questions. This research is an instrumental study of the use of duration and intonation in question formation. Being a tonal language, Nchufie provided the opportunity to explore the interaction of lexical and intonational F0 specification. Both duration and pitch behave unusually in Nchufie questions with respect to the world's languages. Substantial lengthening was found to occur in questions, and the intonational patterns employed in questions, i.e. overall raising and optional final lowering, are types rarely attested in previous literature.

Acknowledgements

Recent work on Nchufie has been undertaken at UCLA in the Department of Linguistics, and the author wishes to thank Sophie Ajeakwa for extensive work as a language consultant, Hilda Koopman for leading the research on Nchufie at UCLA, and Ian Maddieson, Keith Johnson, and Mary Beckman for insightful comments on this study.

Notes

1. This vowel is phonetically back but is considered phonologically central here for the sake of symmetry.

2. Mid may also be a lexical tone although it doesn't occur on nouns in isolation

3 The appearance of the final lowering was also not predictable from the token's location in the word list.

4. In experiment two, designed to test the pitch range questions, of the eight high target words one word, *tɯ́* 'tree' showed no raising in the questions, ending in the same pitch as in the statement condition. This word did undergo raising in experiment one, again suggesting a random process.

References

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

BOYCE, S. and MENN, L. (1979) Peaks vary, endpoints don't: Implications for intonation theory. *Berkeley Linguistic Society*, Chiarello, et al., eds., 5, pp. 373-384.

BREDVAD-JENSEN, A.C. (1983) Perception studies of Swedish interrogative intonation. paper presented at the 10th International Congress of Phonetic Sciences (Utrecht, The Netherlands).

COWAN, R. and SCHUH, R. (1976) *Spoken Hausa* (Spoken Languages Services, Inc., Ithaca, NY.

GÅRDING, E. (1979) Sentence intonation in Swedish. *Phonetica* 36, 207-215.

GRIMES, B. ed. (1988) Pittman, R. & Grimes, J.E. cons. eds. *Ethnologue: Languages of the World.* 11th edition, Dallas: Summer Institute of Linguistics.

HADDING-KOCH, K. (1961) *Acoustico-Phonetic Studies in the Intonation of Southern Swedish.* Lund; Gleerups.

HYMAN, L. (1985) Word domains and downstep in Bamileke-Dschang. *Phonology Yearbook*, 2.

HYMAN, L. (1986) The representation of multiple tone heights. In K. Bogers, H. van der Hulst and M. Mous (eds.) *The Phonological Representation of Suprasegmentals.* Dordrecht: Foris, 109-152.

INKELAS, S. and LEBEN, W. (1990) Where phonology and phonetics intersect: the case of Hausa intonation. in Beckman M. and Kingston J. (eds.) *Papers in Laboratory Phonology: Between the Grammar and Physics of Speech.* Cambridge: Cambridge University Press, pp. 17-34.

LINDAU, MONA. (1986) Testing a model of intonation in a tone language. *JASA* 80(3), pp. 757-764,

MILLER, J. and TENCH, P. (1980) Aspects of Hausa intonation: Utterances in isolation. *JIPA* 10:1-2, 45-63.

NEWMAN, P. and NEWMAN, R. (1981) The q morpheme in Hausa. *Afrika und Ubersee* 64:35-46.

SAMARIN, W. (1966) *A Grammar of Sango.* Janua Linguarum, series practica 38. The Hague.

SCHUH, R. (1978) Tone rules. in Fromkin, V. ed. *Tone: A Linguistic Survey*. New York:Academic Press, 221-256.

SHEN, X.S. (1990) *The Prosody of Mandarin Chinese.* Berkeley: University of California Press.

SHEN, X.S. (1991) Question intonation in natural speech: a study of Changsha Chinese. *JIPA* (1991) 21:1.

ULTAN, R. (1978) Some general characteristics of interrogative systems. In Greenberg, J.H., Ferguson, C.A., and Moravcsik (eds.), *Universals of Human Language*, Vol. 4: *Syntax*, Stanford:Stanford University Press, 211-48.

Acoustic and Articulatory Correlates of P-center Perception

Kenneth de Jong

1. Introduction

The term, p-center, has been applied to two different types of data. On one hand, its coiners (Morton, Marcus and Frankish, 1976) spoke of a syllable's perceptual moment of occurrence. That is, the p-center is the point in time that subjects use to align syllables in a regular rhythm. To determine relative p-center locations, Marcus (1981) had listeners insert a target syllable half-way between regular repetitions of a base token. As has been often demonstrated, listeners do not align the onset of the target to occur halfway between the bases' onsets, but rather they seem to use some point within the syllable.

Morton, et al. (1976) note that the p-center can also be thought of as a production center, and refer to earlier work (Rapp, 1971; Allen, 1972) which had speakers produce various syllables rhythmically (with or without a metronome to help them). In these experiments, speakers again do not produce isochronous onsets, but rather seem to align some point within the syllable to get rhythmic regularity. Thus, the production and perception data are (broadly speaking) similar.

Similarly, there are two different types of theoretical accounts for p-centers. Pompino-Marschall (1989) has recently developed a detailed account of some of his perceptual p-center data which is entirely stated in terms of the timing of energy onsets and offsets in a set of auditory filters. In his modeling, the temporal location of the thresholds in each of the filters is subjected to a weighted averaging to determine a p-center location. Rhythmic productions, then, would be productions designed to create the pattern of auditory stimulation which will appear isochronous according to his perceptual model.

Fowler, by contrast, proposed (1979, 1983) that p-centers are associated directly with gestural events in speech production. Specifically, she suggested that listeners in p-center experiments are locating the temporal onset of the vowel gesture. Because of the complexity of the articulatory to acoustic mapping, the acoustic correlates of these gestural events are not straightforwardly reflected in the acoustic signal. Thus, the p-center and other speech isochrony studies which used acoustic criteria failed to locate isochronous events in rhythmic speech. In her account, speakers simply execute rhythmic productions by producing isochronous vowel gesture onsets.

Although the perceptual and production aspects of the p-center are closely bound together, making separation of these two types of accounts extremely difficult, the accounts can and must be evaluated in terms of how well they integrate p-center data into either a general theory of production, on one hand, or a general theory of perception, on the other. The present paper will focus on evaluating a gestural account of p-centers.

Most of the p-center data are reasonably well suited to the sort of gestural account proposed by Fowler, except for one class of results -- the effect of consonant codas on p-center location. Marcus (1981), for example, found that inserting an additional period of silence during the stop gap for [t] in the word *eight* tended to shift the p-center later in the syllable. Cooper, Whalen, and Fowler (1988) achieved similar results by increasing the duration of later portions of the vowel.

However, there are two barriers to interpreting these (and other) results with respect to a gestural account. First, the articulatory behavior associated with a given acoustic stimulus is a matter of conjecture. Second, temporal aspects of the p-center stimuli are typically controlled either by using synthetic speech, or by digitally editing a naturally spoken base. This is also a problem, at least, because it is difficult to determine what articulatory behavior would have produced the edited tokens.

In addition, digital editing may cause problems because it seems to weaken the psychological link between the auditory stimuli and the articulations that would have produced them. During a pilot experiment to the experiment presented below, listeners were presented with digitally vowel-stretched and vowel-shortened tokens at various inter-stimulus intervals. The stimuli were edited using the same techniques described in Marcus (1981) and in other works. At the short ISI's the stimuli tended to 'lose cohesion.' That is, the high frequency and low frequency components separated into two percepts -- yielding a kind of duplex perception. This effect is especially troublesome, because one subject gave different responses, depending on which percept he attended to. The important point here is that unedited tokens in later experiments did not lose cohesion. Listeners who had not participated in the pilot reported no loss of cohesion when listening to natural unedited stimuli; the truth of their report was especially evident in the curious looks they gave me when I tried to describe the effect to them. Thus, it seems listeners find it easier to integrate acoustic information when it faithfully reflects a producable syllable or word.

Bearing in mind these difficulties, the experiment presented here used a perceptual adjustment paradigm similar to that in Marcus (1981). To determine accurately the articulatory behavior associated with the stimuli, the stimuli were extracted from a corpus of X-ray microbeam data. In addition, to avoid the problems of stimuli loosing cohesion after many repetitions, the tokens were not digitally edited, rather temporal variation was introduced by varying the prosodic pattern within which the target syllables were uttered.

2. Methods.

The present experiment uses a perceptual adjustment paradigm in which listeners were presented with a base token alternating with a target token as is illustrated in Figure 1. During each trial, the subjects were presented with 6 syllables, 3 alternations of a base and the target tokens. The subjects then could move the targets with respect to the base until they were satisfied that the syllables occurred in a regular rhythm ('like soldier's marching'). The subjects used a mouse and a series of boxes to select from a range of rhythmic possibilities which differed from one another in 3 ms increments. By clicking on boxes closer to the edge of the screen, the subjects either added or subtracted a greater period of time to the first ISI -- between the base and target. The same amount of time was either subtracted or added to the second ISI (between the target and the base), so that the interval of time from each presentation of the base was fixed within each trial. The final intervals between base and target were recorded as the subject responses.

To set the initial ISI and determine the range of rhythmic possibilities which would be necessary, a pilot experiment was conducted which used stimuli with digitally vowel stretched and shrunken stimuli. Vowel duration was manipulated by repeating or excising pitch periods in the center of the vowel at zero crossings. The main function of this pilot was to determine the effect of different interstimulus intervals. Results of this experiment from three subjects indicate that perceptual adjustments occurred with consistency at 100, 300, and 500 ms initial ISI's. Subjective impressions were that the perceptual adjustments were more difficult at the longer ISI's. Longer ISI's also introduced more fatigue into the task due to the longer amount of time absorbed by each trial. At the shortest ISI, the

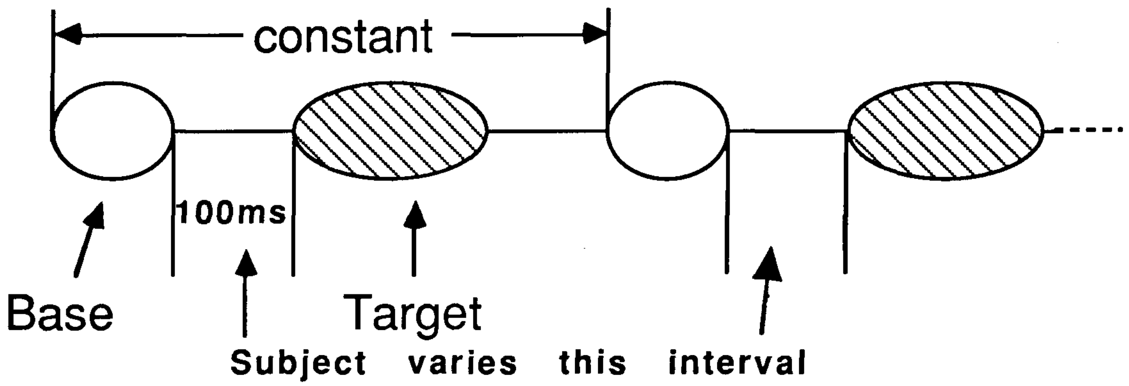


Figure 1. An illustration of the perceptual adjustment method. Subjects are presented with an alternating base and target. Subjects adjust the relative timing of the target with respect to the repeating base.

streaming effect mentioned above was most likely to occur. Since unedited tokens were resistant to the streaming effect, the main experiment was conducted with initial presentations at a fixed interval of 100 ms between each token.

The target stimuli consisted of 12 renditions of the word, *toast* , and 12 repetitions of the word, *totes,* by a Northern Mid-western male speaker who was an undergraduate at the University of Wisconsin. Temporal variability was introduced into the stimuli by varying the prosodic conditions within which the stimuli were uttered. Thus, stimuli differed in how stressed they were. Figure 2 shows the range of variation in the tokens. The duration of the stimuli ranged over approximately 200 ms -- somewhat more than for previous studies, such as Marcus (1981) and Cooper, et al. (1988) whose stimuli continua extended from 60 to 150 ms. From each set of 12 tokens one of the tokens of median duration was selected as a base. Each of the twelve target stimuli (including the one selected for the base) was then paired with the base 3 times for a total of 72 trials.

Note that, since the target stimuli were of different duration than the base, the points within each stimulus which occurred isochronously in the initial presentation were the temporal mid-points. Thus, under the reasonable null hypothesis that listeners will align the acoustic onsets of the syllables to occur isochronously, listeners will have to actively introduce a greater period of silence before tokens of longer duration. Also, since the base was of median duration, half of the expected target responses would occur on each half of the screen. In addition, there is also an identity condition where the base and the target are identical.

The perception subjects were the author and three graduate students in linguistics who were naive to the precise purposes of the experiment.

To have a direct measure of articulatory activity associated with each stimulus, they were extracted from a corpus of X-ray microbeam data. Thus, each acoustic record has time-aligned articulatory trajectories. Included were three points on the tongue (dorsal, mid, and tip), the upper and lower lip, and two points on the jaw. The articulatory data and

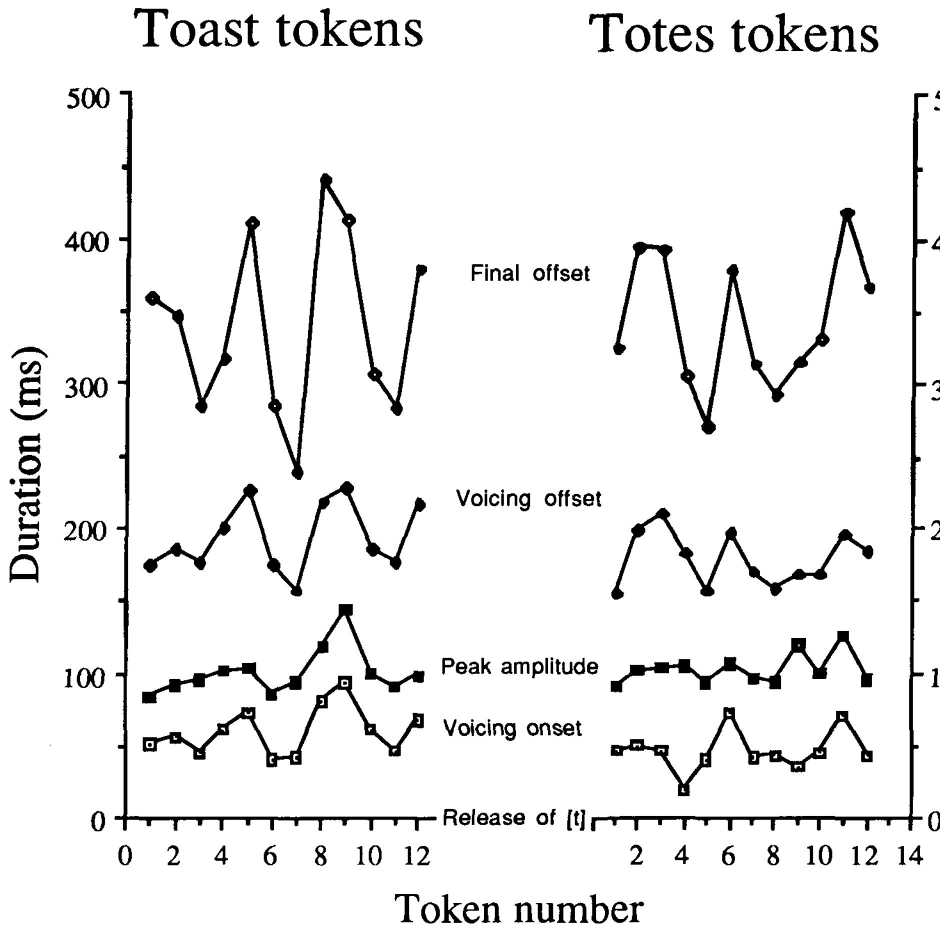


Figure 2. Durations of target stimuli. Token number 1 of each word is used as a base.

elicitation procedures are described more fully in deJong (1991). Tokens were excised from their carrier sentence by means of digital editing. The onset and offset of each acoustic record occurred in periods of silence just before the release of the [t], and in a period of silence which occurred consistently before the onset of the following syllable. Finally, the gain was set in each token so that the peak amplitudes were roughly equal in all of the tokens.

3. Analysis and Results.

To determine whether the listeners responded in a consistent way to the different tokens, their responses were subjected to a two way ANOVA with subject and token as factors. For *toast*, there was a strong main effect of token (F11,144) = 31.76; p < 0.01) and of subject (F(3,144) = 21.61; p < 0.01). The interaction was weaker but significant (F(33,144) = 2.77; p < 0.01). Patterns were similar for *totes* (Subject, f(3,144) = 12.16; token, F(11,144) = 30.07; sub. X tok., f(33,144) = 3.36; all p < 0.01). There was a general consistency in the responses to each token across the subjects, even though the subjects differed in their absolute amount of adjustment.

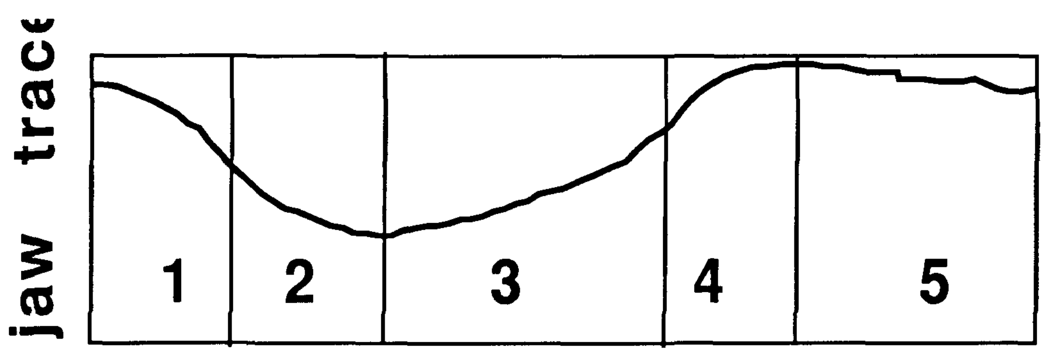


Figure 3. An illustrative trace of jaw movement segmented into five intervals. Events marked by vertical lines are (from left to right) the jaw max during [t], the minimum downward velocity, the jaw min during the vowel, the maximum upward velocity, and the jaw max during the coda consonants. Thus, for the event, jaw minimum during vowel, interevent differences = [3(base) + 4(base) + 5 (base) + 1(target) + 2 (target)] - [3(target) + 4(target) + 5(target) +1(base) + 2(base)].

Assuming listeners were paying attention to a particular acoustic or articulatory event in their alignment of the tokens, one should be able to predict their responses by measuring the difference in inter-event intervals. Any anisochronies in the location of an event will be compensated for in the listener's responses, yielding a strong correlation between inter-event differences and perceptual adjustment. Five articulatory events were extracted from jaw movement trajectories and tongue tip movement trajectories -- the maximum position during the initial [t], the minimum during [o], the maximum during the consonant coda, and points of peak velocity in transition between these points. These points are included schematically in Figure 3. Three events were taken from tongue dorsum trajectories -- the point of extreme retraction during the [o], and velocity maxima into and out of this location.

Four acoustic events were included also for comparison -- the release of the [t], the onset of voicing for the vowel, the point of peak amplitude during the vowel, and the offset of voicing for the vowel. Interevent anisochronies were calculated as the time following an event in the base and the time preceding the event in the target minus the time following the event in the target and the time preceding the event in the base.

The results of regressing event interval differences and the listener’s perceptual adjustments for *toast* are summarized in Figure 4. First, if subjects are attending to vowel gesture timing in particular, the dorsal events should be better at predicting p-center judgments than the jaw or tongue tip. Since the vowels are non-high and occur in an alveolar consonantal context, the dorsum indexes the vowel gesture more clearly than the tip and jaw movements do. However, if listeners are depending on the shape of the acoustic amplitude envelope, one would expect jaw or tip events to be more effective, since they index the size of the radiating orifice of the vocal tract.

As is expected if listeners are localizing the vowel gestures in making their p-center adjustments, the event which performs the best is the velocity peak of the dorsum going toward the vowel position. It accounts for about 93% of the variance in average response to the tokens. In addition, this articulatory event performs better than the acoustic events,

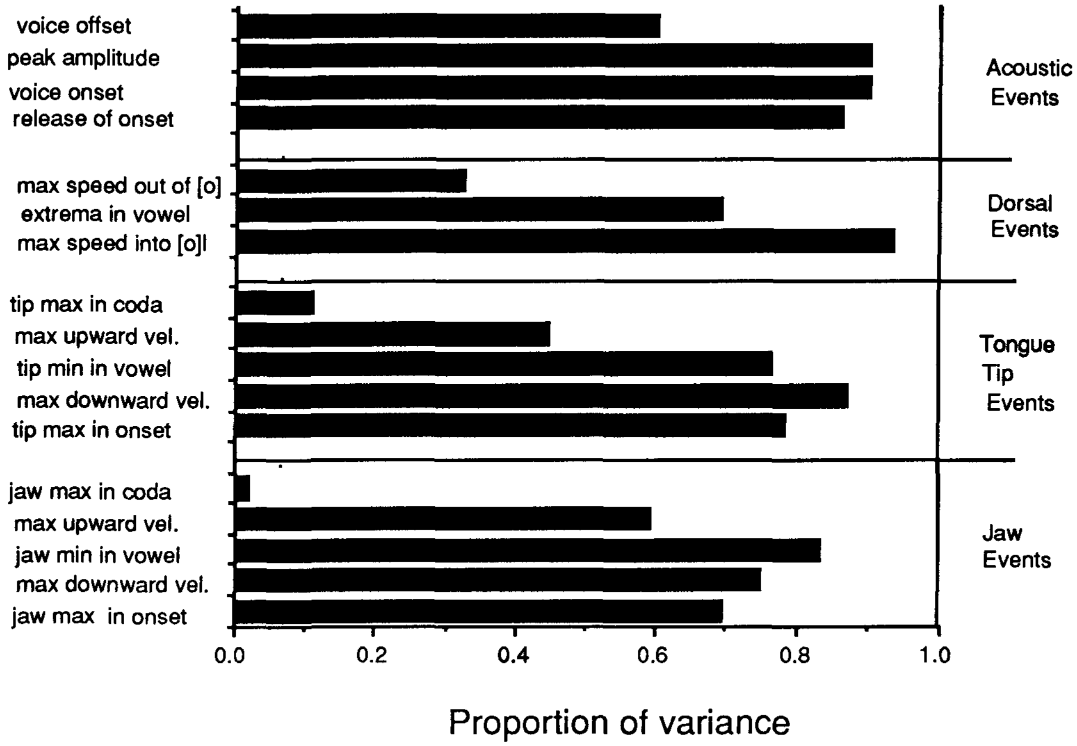


Figure 4. Pearson R-squared values for regressions between interevent anisochronies and average perceptual adjustments.

even though the temporal resolution of the acoustic measurements is 100 times greater. The present results, then, suggest a vowel gesture account of p-centers is appropriate.

However, two aspects of the present data introduce complications. First, comparing across sets of events shows that events which occur near the onset of the syllable all perform quite well, regardless of whether they are defined acoustically or articulatorily. Table 1 summarizes the results for the other subjects and for *totes* as well. Looking across subjects, it is often the case that acoustic events performed better than articulatory events, especially for subjects 3 and 4. Also, the patterns for *totes* are somewhat different than for *toast*. Here, acoustic predictors perform better, but more notable, the tip minimum predictor performs well across all of the subjects, suggesting that listeners are paying more attention to the overall acoustic envelope. Thus, one cannot rule out an acoustic account here.

Second, looking at the residual variance (that not explained by the isochronous positioning of dorsal velocity peaks), one finds additional factors influencing p-center location. In a second analysis of the data, average responses were subtracted from those predicted by dorsal peak velocity isochrony to get a deviation from predicted response. This deviation then was subjected to a stepwise regression against the intervals of time between the events used in the previous analyses. The results are summarized in Table 2. Two things are of note. First, the first factor included in every model accounts for much greater than half of the variance in the deviation from articulatory isochrony. Second, the duration of the coda consonants is always one of the first two factors included in the equation, usually the first. Note that other important durations also occurred toward the end of the syllable. Thus, longer durations in the second half of the syllable tended to shift the p-center location later with respect to the point of dorsal peak velocity.

Table 1. Pearson R-squared values for regressions between interevent anisochronies and perceptual adjustments.

TOAST

Events	Subjects				
	1	2	3	4	Avg..
Jaw max in [t]	0.811	0.730	0.471	0.279	0.693
Jaw min in [o]	0.757	0.784	0.664	0.601	0.830
Jaw max in coda	0.109	0.069	0.000	0.068	0.019
Jaw vel. min	0.841	0.753	0.479	0.405	0.754
Jaw vel. max	0.409	0.444	0.059	0.691	0.591
tip max in [t]	0.837	0.738	0.658	0.341	0.782
tip min in [o]	0.660	0.701	0.704	0.517	0.764
tip max in coda	0.093	0.109	0.097	0.086	0.113
tip vel. min	0.907	0.876	0.634	0.476	0.870
tip vel. max	0.347	0.432	0.398	0.356	0.445
dorsal extremum	0.686	0.665	0.582	0.345	0.354
dors. speed into [o]	0.968	0.891	0.727	0.507	0.932
dors. speed out [o]	0.193	0.213	0.387	0.418	0.324
burst of [t]	0.911	0.864	0.597	0.486	0.858
voice onset, [o]	0.832	0.771	0.779	0.636	0.896
peak amplitude	0.792	0.765	0.822	0.650	0.898
voice offset, [o]	0.458	0.520	0.545	0.563	0.598

TOTES

Jaw max in [t]	0.805	0.329	0.715	0.295	0.702
Jaw min in [o]	0.907	0.626	0.816	0.348	0.854
Jaw max in coda	0.199	0.024	0.132	0.052	0.135
Jaw vel. min	0.928	0.588	0.821	0.287	0.839
Jaw vel. max	0.500	0.312	0.531	0.270	0.514
tip max in [t]	0.860	0.493	0.697	0.474	0.799
tip min in [o]	0.898	0.779	0.884	0.569	0.962
tip max in coda	0.032	0.098	0.049	0.079	0.063
tip vel. min	0.843	0.474	0.679	0.245	0.721
tip vel. max	0.226	0.099	0.234	0.045	0.198
dorsal extremum	0.687	0.243	0.492	0.147	0.516
dors. speed into [o]	0.796	0.268	0.671	0.126	0.614
dors. speed out [o]	0.277	0.144	0.300	0.118	0.272
burst of [t]	0.951	0.592	0.851	0.325	0.870
voice onset, [o]	0.566	0.858	0.555	0.520	0.702
peak amplitude	0.720	0.583	0.711	0.316	0.729
voice offset, [o]	0.350	0.041	0.464	0.042	0.292

Table 2. Summary of stepwise regression. Deviation from expected response against token partial durations.

TOAST		**TOTES**	
Variables Included (in rank order) (F > 4)	R-squared of entire model	Variables Included (in rank order) (F > 4)	R-squared of entire model
JAW			
1. Max up vel. - max in coda	0.900	1. After jaw max in coda	0.698
2. After jaw max in coda	0.942	2. Max up vel. - max in coda	0.878
3. Min in [o] - max up vel.	0.983	3. Onset - max down vel.	0.974
		4. Max down vel. - min in [o]	0.976
TONGUE TIP			
1. After tip max in coda	0.748	1. After tip max in coda	0.786
2. Max up vel. - max in coda	0.917	2. Min in [o] - max up vel.	0.895
3. Min in [o] - max up vel.	0.965	3. Max up vel. - max in coda	0.996
4. Max down vel. - min in [o]	0.991		
5. File onset - max down vel.	0.995		
ACOUSTIC			
1. After [o] voicing (coda)	0.952	1. After [o] voicing (coda)	0.890
2. Peak ampl. - voice offset	0.980	2. Peak ampl. - voice offset	0.953
3. Voice onset time	0.994	3. Voice onset - peak ampl.	0.979
		4. Voice onset time	0.987

4. Discussion and Conclusion.

To summarize, then, the present experiment seeks to replicate earlier perceptual p-center studies using naturally varied acoustic stimuli with articulatory recordings. Results are consistent with a vowel gesture account of p-centers in that the timing of tongue dorsum velocity peaks predict perceptual adjustments better than acoustic markers do for one of the stimuli words.

However the results are not entirely consistent across listeners. Thus, an acoustic account cannot be entirely ruled out here. What can be said is simply that two of the subjects, and the average taken across the subjects point toward a dorsal event as the best predictor of p-center location.

Also, the results are not consistent from word to word. Adjustment of tokens of the word, *toast,* was well predicted by the dorsal velocity peak location, while adjustments of *totes* were better predicted by acoustic events and by the tongue tip minima. It is possible that this disparity in results is due to an oddity of the *totes* tokens. Three of the subjects noted that the *totes* presentations tended to loose cohesion, similar to the digitally edited tokens in the pilot experiment. This was never true of the *toast* tokens. The difference may be due to differences in the spectral makeup of the tokens. Having a voiceless stop immediately following the vowel, *totes* typically has a relatively short vowel and rapid transitions between the vowel and neighboring consonants. This rapid acoustic change may tend to defeat the ability of the auditory processing system to integrate the acoustic signal into a perceptual unit. Similarly, the difference in the tokens may be due to the strength of the hearer's representation of the two lexical items. The relatively common occurrence of *toast* may aid the hearer in interpreting the various acoustic components as signs of a single complex articulatory event. Regardless of the reason, the difference in the two sets of tokens in presentational cohesiveness may explain the difference in the results for the two sets of stimuli.

Finally, even among the *toast* tokens, perceptual adjustments deviated systematically from those necessary for dorsal velocity peak isochrony. Tokens with long coda consonants tended to have later p-center locations than those with short codas. This may be due to a couple of different factors. First, it may indicate an averaging of event locations throughout the syllable. This is the type of explanation presented in Pompino-Marschall (1990), as well as in Cooper, et al. (1988), for the effect of coda duration manipulation p-center location which they found.

Given the present paradigm, there is a second possible explanation -- that subjects tend to undershoot the amount of adjustment needed for complete isochrony of events. If subjects tend, for some reason, to be conservative in the magnitude of the adjustment they perform, the longer tokens will exhibit a p-center location closer to the center of the token. The apparent effect will be one of the extended coda duration "drawing" the p-center location toward the longer codas.

The present work suggests that the experimental paradigm should be examined more closely for potential artifacts on the p-center data. To tease out these issues, a variety of paradigms should be used. For example, it might be useful to ask listeners to locate anisochronous sequences in a series of trials. This paradigm would give a range within which perceptual isochrony is be free to vary, and thus, would show whether the effect of consonant coda duration could plausibly be seen as an artifact of the present paradigm.

One general methodological conclusion can also be drawn. If experimental studies are to effectively evaluate gestural accounts of p-center location, the stimuli used must be kept as natural as possible. Given the present state of our knowledge of the dynamics of natural speech production, it is difficult to assess which aspects of the speech signal might cue relevant information about rhythmically important aspects of speech production. Similarly, the nature of the psychological link between production and perception which plays a role in a gestural account of p-center adjustments needs further investigation. The streaming effects which occurred on occasion in the experiments reported here suggest that this psychological link may be quite fragile, and care must be exercised to keep it in tact during experimental sessions.

Acknowledgments.

I would like to thank Keith Johnson for extensive aid in developing the experiment running software, as well as for extensive discussion of the research reported here. Also thanks go to Victoria Anderson, Barbara Blankenship, and Richard Wright for allowing themselves to be subjected to the P-center paradigm. Work supported by the NIDCD under grant number T32DC00029.

References.

Allen, G. (1972), The location of rhythmic stress beats in English: an experimental study, I, *Language and Speech*, 15:72-100.

Cooper, A.M., Whalen, D.H. and Fowler, C.A. (1986), P-centers are unaffected by phonetic categorization, *Perception and Psychophysics*, 39:187-196.

Cooper, A.M., Whalen, D.H. and Fowler, C.A. (1988), The syllable's rhyme affects its P-center as a unit, *Journal of Phonetics*, 16:231-241.

deJong, K.J. (1991), The articulation of consonant-induced vowel duration changes in English, *Phonetica*, 48:1-17.

Fowler, C.A. (1979), "Perceptual centers" in speech production and perception, *Perception and Psychophysics,* 25:375-388.

Fowler, C.A. (1983), Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet, *Journal of Experimental Psychology: General,* 102(3):386-412.

Marcus, S.M. (1981), Acoustic determinants of perceptual center (P-center) location, *Perception and Psychophysics*, 30: 247-256.

Morton, J., Marcus, S.M. and Frankish, C. (1976), Perceptual centers (P-centers), *Psychological Review*, 83:405-408.

Nadler, R.D., Abbs, J.H., and Fujimura, O. (1987), Speech movement research using the new x-ray microbeam system, In *Proceedings of the 11th International Congress of Phonetic Sciences,* Vol. 1, pp. 221-224.

Pompino-Marschall, B. (1989), On the psychoacoustic nature of the P-center phenomenon, *Journal of Phonetics*, 17:175-192.

Rapp, K. (1971), A study of syllable-timing, *Speech Transmission Laboratory, Quarterly Progress and Status Report* (Stockholm), 1971(1):14-19.

Tuller, B. and Fowler, C.A. (1980), Some articulatory correlates of perceptual isochrony, *Perception and Psychophysics*, 30:247-283.

Fronted velars, palatalized velars, and palatals

Patricia Keating

Aditi Lahiri

Max-Planck-Institute for Psycholinguistics
Nijmegen, The Netherlands

1. Introduction

It has long been observed that velar consonants in front vowel environments tend to be articulated at a more forward position along the palate than velars in back vowel environments (e.g. Sapir 1921:52). Sapir compared the articulatory positions of the English /k/s in *kin* vs. *cool*, and textbooks since have offered similar comparisons (e.g. Heffner 1950:191; Ladefoged 1975:49ff). Many experimental studies of articulation have provided support for this basic claim. Nonetheless, most sources are not very precise about the segment that results from such contextual velar fronting. Is the segment still a velar, despite its being fronted? Or is it in fact fronted all the way to being a palatal? If a palatal, is it the same kind of segment as other consonants classed as palatals? In either case, is its articulation like that of contrastively palatalized segments? (The term "palatalized" is used here only to refer to surface phonemic contrasts, not to contextual fronting. Following the new IPA conventions, we will symbolize a palatalized velar as [kʲ] and a fronted velar as [k̟]).

In one sense, the first two of these questions have a straightforward definitional answer. In traditional terminology, the "palate" refers to the hard palate, and any sound made on the hard palate is "palatal". On this view, any instance of /k/, /g/, etc. articulated on the hard palate is "palatal", and only instances articulated on the soft palate are "velar". (See, for example, Recasens 1990 for this anatomical answer -- he refers to fronted velars as "back palatals" -- but Gorecka 1989 for a different, phonological, division of articulatory regions.) Since the hard palate is a fairly large portion of the roof of the mouth, palatal articulations may encompass many variants. Thus it may be the case that fronted velars are palatal in this broad usage, but systematically different from other kinds of palatals. The third question above, whether fronted velars or palatalized velars are like other palatals, is then still of interest.

These descriptive questions bear on issues of feature representation. If a feature system is to have some way of representing velar fronting, it should be known whether a fronted velar is the same thing as some other segment type, such as a palatal or a palatalized velar. The system of Chomsky & Halle (1968) forced the strong hypothesis that extremely fronted velars can be categorized with both palatals and palatalized velars, because there is only one set of feature values available for all three of these (supposed) segment types. Under this proposal, the tongue body features High, Low, and Back are used to describe vowels as well as a variety of consonants, including velars and palatals. Velars are high, back segments, and palatals are high, front segments. In (1) it is shown how values of these features are used to represent velars, palatals, and related segments.

(1)

VOWELS:	[i], [I]	[u], [U]
CONSONANTS:	palatals palatalization	velars velarization
high	+	+
low	-	-
back	-	+

Under this proposal, velars differ in place from front vowels like [i] in their value for Back. Fronting of a velar by a front vowel can be represented in feature terms by replacing the [+back] value of the consonant with the [-back] value of the vowel. The fronting is then directly expressed in terms of the feature describing location of the tongue body with reference to the palate. The same representation is used for palatalized velars, which are velars on which a high front vowel articulation is superposed: they are [-back], rather than [+back]. Note, however, that this combination of feature values is the same as that for palatals. Under this analysis, then, fronted velars, palatalized velars and palatals are all represented featurally as the same thing[1]. This aspect of the SPE representations, which is grounded in the traditional definition of palatal sounds as made on the hard palate, is quite intentional. The addition of a vowel type of articulation to a consonant articulated by the tongue body is seen as shifting the primary place of articulation of the consonant towards that of the vowel, in this case from velar to palatal. That such place shifts under fronting or palatalization are an automatic outcome of the SPE multiple use of the features High, Low, and Back was taken to be an argument for the feature system.

Another argument presented for this aspect of the SPE system is that languages do not contrast palatals, fronted velars or palatalized velars with one another. There are three potential paired contrasts of these three phonetic categories. The first is palatals vs. fronted velars. As we will see below, languages with palatals do contrast them with fronted velars before front vowels. The second potential contrast is palatals vs. palatalized velars; though these two sound-types are phonetically distinct, it will be seen that we consider this case equivalent to the first. The third is fronted vs. palatalized velars. As far as we can tell, it is true that no language contrasts these segment types before a phonetic high front vowel [i], and our data will argue against such a contrast before the mid front vowel in Russian. We will claim that fronted and palatalized velar articulations are phonetically too similar to contrast.

Since SPE, feature systems and representations have changed quite a bit, particularly with respect to place of articulation. Nonetheless, essentially the same issues remain, since we still do not know how many different types of segments need to be represented. If anything, recent proposals about feature geometry only sharpen our need to understand the relevant articulatory differences. For example, under many proposals since Sagey (1986), a sharp distinction is drawn between dorsal, or tongue body,

[1] Another phonological sound type that would also be represented by the feature values for palatals would be "front velars" which contrast with back velars. For example, Yanyuwa (Kirton & Charlie 1978; see also Ladefoged & Maddieson 1986:20) contrasts front velars ("palatovelars") with back velars ("dorsovelars") (and also with laminal postalveolar coronals -- "alveo-palatals") before /a/ and /u/. These two velars neutralize to the front velar before /i/ (which is the only front vowel in this 3-vowel language). Other nearby languages are reported to have the same contrast. (Kirton & Charlie mention Garawa (Furby 1974) and Djingili (Chadwick 1975). As neither source discusses distribution or neutralization facts, we cannot say whether the velars contrast before /i/.) Whether these front velars differ phonetically from sound types made on the hard palate in other languages remains to be determined.

articulations, and coronal, or tongue blade, articulations. We can then ask whether any of the segment types under discussion -- palatals, fronted velars, palatalized velars -- are phonetically coronal or dorsal or both.

Our main point, then, is to clarify the phonetic differences between certain segment types, in a way that is relevant to any version of feature theory. For that reason, we will not present formal proposals about how these sounds might be represented featurally. In particular, since there is currently no agreement as to how the phonetic dimension of tongue body frontness/backness is to be encoded, we will not rely on specific features. Rather we will use general articulatory terms to specify information which might be taken into account under any theory of phonetic representation. Specific proposals can be found in Gorecka (1989), Keating (1988, 1991), Clements (1991), and Lahiri & Evers (1991), inter alia.

In this paper we focus on stop consonants in four languages, though other data will be mentioned as well. Czech and Hungarian were selected as languages with contrasting palatal and velar stops. In both of these languages, the palatals alternate with anterior coronals such as /t/, suggesting that the palatals might also be coronal. At the same time, the palatals of these languages are thought to differ, with the Hungarian ones more backed and therefore more prototypically palatal (P. Ladefoged, p.c.). In addition, we wanted to know whether the Czech and Hungarian velars are fronted before front vowels as in other languages. Russian was selected for its contrasting palatalized and non-palatalized (velarized) velar stops. English was included for its allophonic velar fronting.

For each of these languages, articulatory data from the literature, and our own acoustic analyses, are used to compare the consonant types. The articulatory data show that all of them are articulated on the hard palate and thus are all "palatals" in this definitional sense. However, the palatals of both Czech and Hungarian differ from the fronted velars and from some of the palatalized velars in having a clear coronal-postalveolar component. The acoustic data also support this distinction.

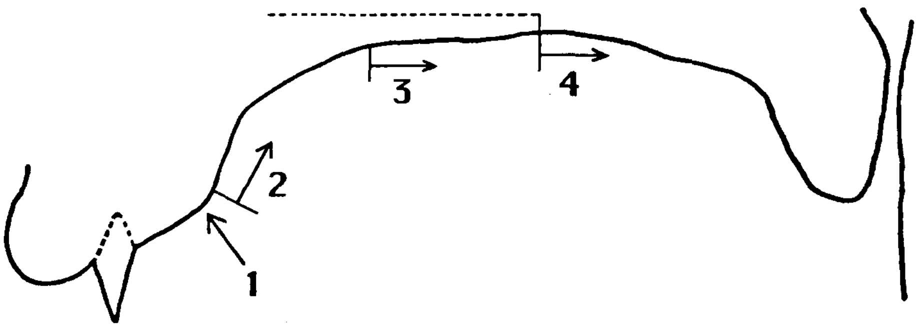
2. Articulatory comparisons

2.1. Method and terminology

The data to be considered consist primarily of published tracings of midsagittal X-rays. Sources were located largely by examining the listings in Dart (1987). In each case the available tracings were inspected to determine where the greatest constriction was formed, and by what part of the tongue, and also the general position of the blade of the tongue. When a source indicates the division between the hard and soft palates on the roof of the mouth, constrictions are described relative to that point. In addition, if palatograms and linguograms are included in the source, they were also examined for information about any contact between active and passive articulators. This study is thus similar in scope and method to Recasens (1990) and Keating (1991).

A few words are in order about limitations on inferences from such sources of data. With respect to published X-ray tracings, there are three major difficulties. First, tracings are difficult to make and therefore not entirely reliable, especially when viewed only in reproduction. Second, the tracing is taken either from a still X-ray, in which case the speech sample is artificial, or from a cine X-ray, in which case it may not show the desired moment in a rapidly changing speech event. Third, it may not be clear whether the tracing shows central or lateral contact. With respect to static palatograms, problems arise with consonants in /i/ and /e/ contexts, since these vowels themselves may have extensive palatal contact. The static palatogram shows the combined contacts of all the segments in the speech sample. The palatograms are then valid only with respect to a stop consonant's central occlusion.

In describing place of articulation, it will be extremely useful to refer to an anatomical distinction for which there are no standard terms. The hard palate extends up and back from the alveolar ridge to the roof of the mouth. For many speakers, the hard palate can be divided into two parts with different orientations: one, a vertical or diagonal sloping section behind the alveolar ridge, the other, a more horizontal section forming (part of) the roof of the mouth. The first, sloping, part is roughly the same as the part with rugae, or ridges. Where possible, then, we will refer to the corner of the alveolar ridge, meaning the posterior part of the ridge where the upward slope of the hard palate begins; to the diagonal of the hard palate, meaning the basically upward-sloping part of the hard palate; and to the roof of the hard palate, meaning the topmost part adjacent to the soft palate. (See Catford 1988:93 for a slightly different division.) Figure 1 illustrates these.



1. Corner of
Alveolar Ridge
2. Diagonal of Hard Palate
3. Roof of Hard Palate
4. Soft Palate

Figure 1. Schematic of relevant anatomical landmarks and terms.

With respect to active articulators, we distinguish between the tongue blade, and the tongue body or dorsum[2] . Following Keating (1991), we take the blade to include, conservatively, the tip of the tongue and the first 2 or 3 cm beyond the tip (roughly, the part of the tongue, in front of the anterior genioglossus, which at mid-line feels less firm to the touch). Some investigators define the blade as the movable end of the tongue not attached to the floor of the mouth, or roughly 1 cm. The reason to prefer a definition giving a longer blade is that clearly-"coronal" consonants are often formed more than a cm behind the tip. Stops made at the corner of the alveolar ridge are typically made 2-3 cm back on the tongue. Furthermore, Dart (1991) shows that even alveolars may be articulated this far back, when they are made with the blade of the tongue. Since such consonants must be "coronal" for the term to be of any useto linguists, then operationally the blade must go back at least that far. Such an operational definition also has a physiological basis. A length of 2-3 cm corresponds to the back of Stone (1990)'s "anterior" tongue segment. Stone identifies four functionally-independent segments of the tongue; our "blade" includes the first two of these, corresponding to the tip and the

[2] In anatomical usage dorsum refers only to the cover of the tongue. (Hence Ohman's 1966 distinction between "dorsal" - for consonants - and "vowel" channels of articulation.) We follow current linguistic practice in using dorsum to refer to the entire body, for both vowels and consonants.

rest of the blade. Thus we expect this 2-3 cm segment of the tongue to function as an articulator independent of the rest of the tongue body.

Finally, we will often have occasion to discuss lateral tongue contact. By this we mean contact between the sides of the tongue and the sides of the palate, usually along the teeth. The length of lateral contact in the front- back dimension can be distinguished from the amount of lateral contact in the side-to-side dimension. It is important to note that such lateral contact is not what is involved in the production of "lateral" segments. Lateral segments involve escape of air around the lowered side(s) of the tongue, whereas lateral contact, with the tongue sides raised, precludes such airflow.

In this connection it is important to review how terms for consonantal places of articulation are used here. There are several non-anterior coronal consonant categories in the IPA. Of these, palato-alveolars and retroflexes are made in about the same area at or behind the corner of the alveolar ridge, with the palato-alveolars laminal and the retroflexes apical or sublaminal. In this system English [ʃ] is palato-alveolar, though it is often called palatal in American usage. IPA palatals are distinct from either of these, with the IPA chart suggesting that their place of articulation is further back; their lowered tongue tip position is also characteristic. A wide range of consonant types -- stops [c] and [ɟ], fricatives [ç] and [ʝ], the corresponding affricates, nasal [ɲ], lateral [ʎ], and glides [j] and [ɥ] -- is recognized. In addition, the IPA alphabet -- though not the main consonant chart -- contains alveolopalatal fricatives [ɕ] and [ʑ]; affricates [ʨ] and [ʥ] are also commonly transcribed. Thus there is a distinction between palatal and alveolopalatal fricatives and affricates. For our purposes this amounts to the claim that the "palatal" stops of Czech and Hungarian are different in place from the "alveolopalatal" affricates of, for example, Polish, Mandarin, and Serbo-Croatian (see Keating 1991).

No other alveolopalatal consonant types are recognized in the IPA system. It seems quite likely, however, that some languages, including languages with alveolopalatal obstruents, would have sonorants that are more alveolopalatal than palatal; see Recasens (1990). Then the IPA symbols for palatal sonorants must be used for these alveolopalatals as well, creating some ambiguities. We intend our discussion here to hold of palatals alone, excluding alveolopalatals, palato-alveolars, and retroflexes. Another symbol ambiguity that arises with the palatals is the topic of this paper. Palatal symbols such as [c] and [ç] are often used for front velars articulated on the hard palate, instead of IPA [k̟] and [x̟]. We will clarify this distinction at length below.

2.2 Czech.

Several sources on Czech phonetics include X-rays of oral and nasal palatal stops. However, it appears that several of these sources rely on the same original study[3] ; in all there are probably at most 3 or 4 different speakers represented.

The X-ray tracings of oral palatal stops in Daneš, Hála, Jedlička, & Romportl (1954)[4] are typical of those available for Czech. Figure 2 is derived from one of these. These tracings show long contact from the corner of the alveolar ridge back onto the roof,

[3] Various reports from Hála's group are apparently all based on Polland & Hála (1926a,b). These early X-rays are stills; in some of the sources, some of the articulatory profiles shown are not from X-rays at all, but are reconstructed from palatograms (as in Hála 1923). In Hála 1962, the X-rays and palatograms are probably from different sources. In another source, Pacesová 1969, some figures are from Hála sources and some are apparently from her own cine film.

[4] This source is based on Polland & Hála (1926) and therefore shows tracings from still X-rays, along with labiograms, linguograms, and palatograms. The hard-soft palate boundary is shown. No information is given about the speech items.

though not covering all of the hard palate. Palatograms show wide contact all along the sides, but a stop occlusion only at the front of the palate, probably on the corner and part of the diagonal. It covers about the same total area as the area of an alveolar plus the area of a palato-alveolar combined. Linguograms show that this occlusion is made with the blade, not the tip, of the tongue. They also show lateral contact extending far back on the tongue. That is, the palatograms and linguograms make clear that the long contact seen in the X-rays is most likely lateral contact only, not central contact (though see Straka 1965:157 for the possibility of long central contact). The tip of the tongue is down and the whole tongue is pulled forward, with the pharynx quite wide as a result. As no information is given about the speech items recorded, the contribution of the vowel context to the overall tongue position cannot be determined. The nasal stop is similar to the oral ones, but with a shorter occlusion which does not extend as far forward on the palate or tongue, and lateral contact that, in the X-ray tracing, does not extend as far back.

Since the contacts for velars shown in these sources are made only on the soft palate, there is no overlap at all between palatals and velars in this sample. The occlusions for velars are slightly longer than those for palatals.

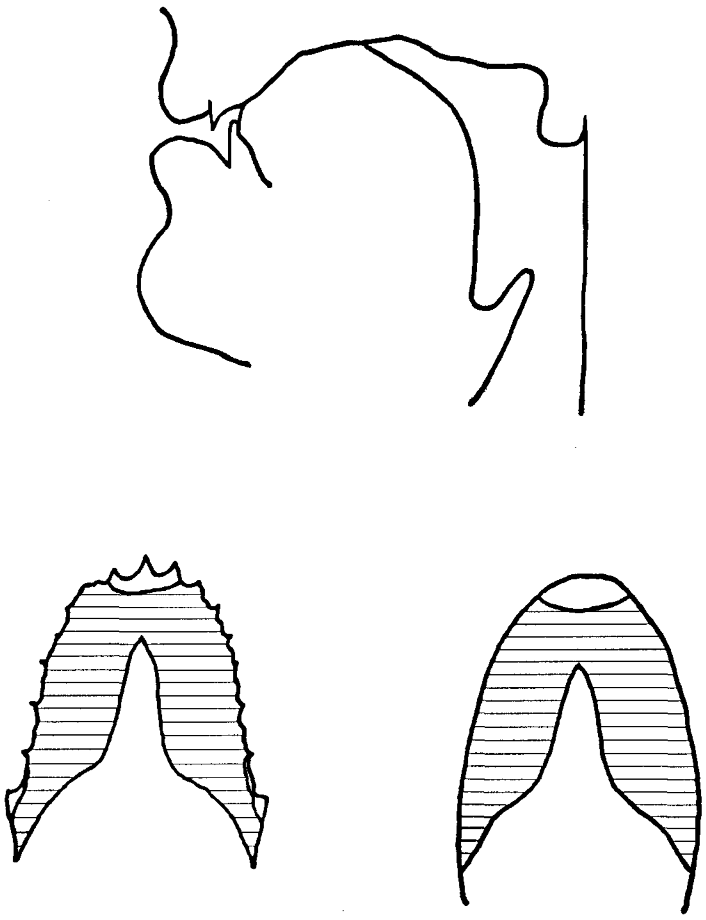


Figure 2. Articulatory data on palatal stops in Czech, after Daneš et al. (1954). Top: tracing of sagittal X-ray, showing lateral tongue - palate contact. Bottom left: tracing of palatogram, contacted area shaded. Bottom right: tracing of linguogram, contacted area shaded.

Based on these data, the Czech palatal stop looks like a long coronal stop (extending over anterior and nonanterior areas) combined with the lateral contact of a [j]. It is therefore useful to compare these palatal stops with the Czech palatal glide /j/. Palatal glides in general show lateral contact only, basically a long narrowing along the hard palate. In Czech the profile of the tongue seen in the X-ray tracings is very similar for the stops and the glide, but the glide of course does not show the central contact seen with the stops. Correspondingly, the palatograms show lateral contact for the glide which extends as far forward along the teeth but which is slightly less extensive from side to side, indicating that the tongue body is less raised. (This contact for the glide is at the same time much more extensive than for the non-palatal coronal consonants.) This difference suggests that with the stops, the goal of stop occlusion leads to extreme tongue raising. Otherwise, the stops are quite comparable to the glide in tongue body position. In some palatograms, the greatest constriction in the glide is at about the corner of the alveolar ridge; others show no point of greatest constriction. Straka (1964, 1965) provides interesting palatographic data on the relation of the Czech palatal stops and glide. Straka (1965:157) shows that as the stop /c/ is pronounced more energetically, the lateral contact becomes more extensive so that the central occlusion becomes longer, and that more of the hard palate is contacted. Straka (1964:96) shows that for the glide /j/, an emphatic pronunciation results in so much lateral contact that there is almost a central occlusion, and this near-occlusion is in the same location as for the stop occlusion. This finding supports the parallel treatment of the Czech stops and glide as both consisting of a coronal (laminal postalveolar) component together with a dorsal component. It might be said that for the stops, the coronal component dominates the articulation, while for the glide the dorsal component dominates, but that in both cases the secondary component becomes stronger in emphatic pronunciation.

2.3. Hungarian

Bolla discusses X-rays of Hungarian in various sources; we refer here primarily to Bolla (1980) because there he gives continuous tracings of the tongue's shape and location, rather than schematics based on a few points, as in some of his other publications[5] . The palatals (and velars as well) were all before/after a low vowel, except one palatal /c/ before /u/; thus all might be somewhat backed in their place of articulation. Figure 3, based on Bolla, shows a Hungarian palatal stop articulation. The voiced stop and nasal of Hungarian look similar, with contact largely on the diagonal nearer the roof. This contact is somewhat longer than that for velars. The tracing of the voiceless stop shows no occlusion at all; the near-occlusion shown is on the diagonal of the palate. The palatograms and linguograms for all three show occlusions over a large area of the hard palate, but not using the first cm or two of the tongue blade. The central contact is not as long as the lateral contact, but it extends quite far back on the tongue. The 1981 schematics- over-time treatment shows that the palatals have their first contact more to the back of the palate and slide it forward onto or towards the diagonal; if this is correct, then the palatograms show a composite of total contact areas during the articulation. Indeed, the central contact for the occlusion is longer in the palatogram than in the linguogram.

None of these palatals overlap at all in place of articulation with the palato-alveolars or velars shown by Bolla. The consonants transcribed as palato-alveolars (fricatives and affricates) are apical and postalveolar, made in front of the palatals; they

[5] The data in Bolla (1980) are palatograms, linguograms, tracings of frames from a cine X-ray, and other kinds of data; the tracings shown are from the "pure phase of articulation" of the segments. Sometimes the side and center of the tongue are distinguished. The hard-soft palate boundary is not shown. Original frames from five points in time are also reproduced in miniature. Several words were recorded for each segment. The palatograms do not necessarily correspond to the X-ray tracings with which they are paired.

show less side-to-side lateral contact than do the palatals. The velars are entirely on the roof, with their central contact on the soft palate. The palatograms and linguograms of the palatals and velars indicate that the contacts are in quite different regions. In terms of the overall shape of the front of the tongue, as seen in the X-ray tracings, for the palatals the tongue is bunched forward so that the blade comes to the lower teeth and the diagonal. For the velars, the tongue is also bunched up, but is not thrust forward to the teeth/diagonal; its shape is instead rather symmetrical.

The glide /j/ in Hungarian is not quite like the palatal stops. Bolla's X- ray tracing shows a glide which is rather back, on the roof, and has no obvious blade involvement. The palatogram shows no single point of constriction, and the linguogram shows a consistent lateral contact along the tongue. The glide and stops are similar in terms of their rearward lateral contact, but the glide lacks the forward central constriction. In these respects the glide is like the Hungarian palatal fricative, which also has only lateral contact. As in other languages in the literature, the Hungarian palatal fricatives have longer lateral contact than do the stops -- from the diagonal over the entire hard palate, and even onto the soft palate. A wide passage remains open along the center of the tongue, as for [i] vowels. That is, the glide and the fricative constrictions extend further back than does the stop, and the stop is not strongly coronal.

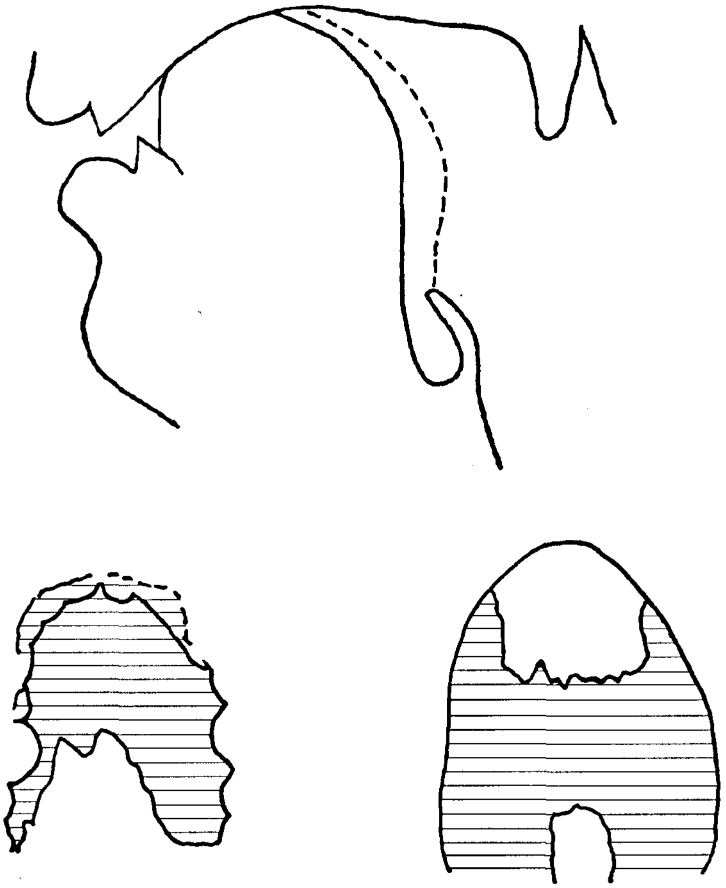


Figure 3. Articulatory data on palatal stops in Hungarian, after Bolla (1980). Top: tracing of sagitaal X-ray. Bottom left: tracing of palatogram, contacted area shaded. Bottom right: tracing of linguogram, contacted area shaded.

Although Icelandic is not one of our primary languages of investigation here, Petursson's (1974) X-ray study of that language provides additional information about palatals not available from the studies of Czech and Hungarian. In this study, superposed frames from different time points in palatal stops before front vowels clarify the time course of the lateral articulation (which, it will be recalled, is what X-rays primarily reflect for palatals). These figures indicate that neither the contact nor the release is made all at once, but instead often proceeds from small to large to small contact, with a more fronted contact at the release. It is certainly plausible that the same holds in Czech and Hungarian; the back-to-front movement of contact has already been mentioned for Hungarian. At the same time, it must be noted that Recasens found that the alveolopalatal nasal of Catalan is released from front to back (alveolar contact released first).

In Hungarian, then, a palatal stop has a long occlusion on the hard palate, together with additional rearward lateral contact. The stop occlusion probably slides during closure, but its location lies between that of the other postalveolar and velar consonants. The palatal glide and fricative have lateral contact all along the tongue but the tongue shows no particular point of constriction. The Hungarian palatals are thus somewhat different from those in the Czech sources. In both languages the tongue is raised and pulled forward with the tip behind the lower teeth, so that there is extensive lateral contact between tongue and palate for palatal stops and glides. (Czech has no palatal fricatives to compare with Hungarian.) The palatal stops in the two languages differ in where central contact is made -- it is more forward, and definitely coronal, in Czech. The Hungarian stops are only marginally coronal even under our definition of the blade. The palatal glides in the two languages differ in the X-ray tracings, with the Hungarian glide showing less constriction along the diagonal, but in the linguograms the Hungarian glide shows more contact. From these sources, however, it is impossible to say whether the palatals of the two languages are systematically different, since we do not know what vowel contexts the Czech palatals were produced in. If, as seems likely, they were largely front vowel contexts, then we would expect the Czech palatals to differ from the Hungarian ones, produced in back vowel contexts. We will return to this point in Section 3.

2.4. Russian

Several published sources provide data on Russian velars, usually in comparisons of palatalized with non-palatalized velars. However, minimal pairs are not often given, since isolated native words do not provide them. Palatalized velars occur mainly before front vowels, and non-palatalized velars mainly before back vowels. (According to Jones & Ward 1969, palatalized /k^j/ occurs before back vowels only in borrowings, and non-palatalized /k/ occurs before front vowel phonemes /i/ and /e/ only in combinations of the preposition /k/ with words beginning with those vowels. In these cases the phoneme /i/ is realized as a retracted allophone. Thus non-palatalized /k/ is not found at all before a phonetic [i], and is found before /e/ only across a word boundary.) Therefore, sources tend to give palatalized velars in a front vowel context, and non-palatalized velars in a back vowel context; other sources give no information about vowel context. Nonetheless, minimal pairs are most helpful for our purposes, and some studies use minimal nonsense syllables. Sources showing minimal pairs include Koneczna & Zawadowski (1956), Skalozub (1963), and Fant (1960)[6] . Figure 4 shows a palatalized vs. a non-palatalized velar before /a/ based on Skalozub (1963).

Skalozub (1963) compares velars between /a/ vowels for three speakers. For two of the speakers (S's figures 100, 102), the /k/ and /kj/ are extremely different in place of articulation, with no overlap at all and contact on soft vs. hard palates. For the other speaker (S's figure 101), the /k/ and /k^j/ are much more similar; the non-palatalized /k/ is clearly on the soft palate, while the palatalized /k^j/ is just in front of it, at the border of the two palates. However, this speaker does have a clear distinction between /g/ and /gj/ (S's figure 105). A greater difference between /g/ and /g^j/ than between /k/ and /k^j/ seems to be a general pattern (S's figures 104-107), and is due to the fact that /g/ is more back than is /k/. (The same can be seen in the palatograms reproduced in Koneczna & Zawadowski (1956).) Bulanin (1970) reproduces one pair of figures from Skalozub and describes the articulation of the non-palatalized consonants as being on the soft palate near the border between the two palates, and of the palatalized consonants as being on the hard palate near the border between the two palates. He also notes that different points on the tongue form these articulations. However, in both kinds of velars, the contour of the tongue is

[6] Koneczna & Zawadowski (1956) shows tracings from still X-rays along with palatagrams from other sources. Four speakers produced nonsense and real words, but tracings of velars are limited to one speaker. The hard-soft palate boundary is shown; the sides and center of the tongue are distinguished. Skalozub (1963; data reproduced also in Skalozub 1966, Bulanin 1970, and possibly Akishina & Baranovskaja 1980) shows originals and tracings of still X- rays, palatagrams and odontograms (lower teeth contact), of nonsense items for 4 speakers. Fant (1960) presents still X-rays for which the entire surface of the tongue has been highlighted and in which the sides and center of the tongue are distinguished; one speaker produced monosyllables.

rounded, with the blade not raised to the palate; odontograms (lower teeth contact) show contact between the tongue (tip) and the lower teeth.

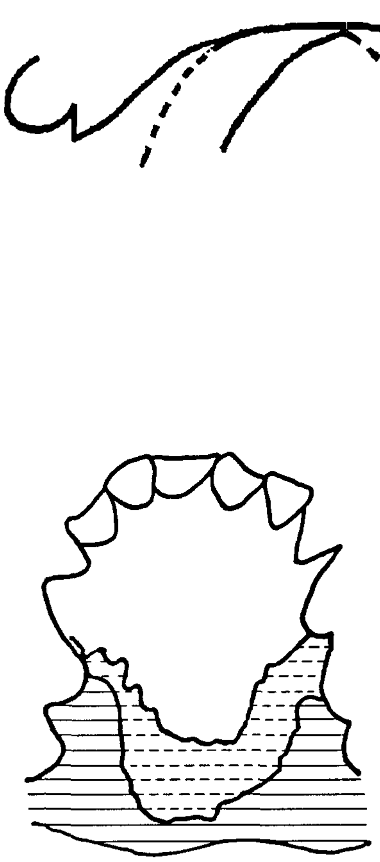


Figure 4. Articulatory data on palatalized and nonpalatalized velar stops in Russian, after Skalozub (1963). Top: superimposed tracings of sagittal X-rays of nonpalatalized (plain line) and palatalized (dashed line) stops. Bottom: superimposed tracings of palatograms of nonpalatalized (plain-line shading) and palatalized (dashed-line shading) stops.

Koneczna & Zawadowski (1956) show /ka/, /k^{j}a/, /ga/, and /g^{j}a/ for one speaker, plus /ka/ and /k^{j}i/ for another speaker, and palatograms, also probably not minimal pairs, from still another speaker. In the minimal pairs, the palatalized velars are well forward on the roof of the hard palate, while the non-palatalized velars span the hard-soft palate boundary. Also, in the palatalized velars the tongue is somewhat raised in front of the constriction. The constriction after stop release is probably quite long, like that for [x^{j}].

Fant (1960:186) gives very small tracings of /ka/ and /k^{j}a/. The non-palatalized velar has the longest contact seen in any source for any oral velar; it appears to cover the entire soft palate including the uvula[7] . The palatalized velar has a much shorter contact (though one which is still longer than in any other source), probably on both the hard and soft palates. (The boundary between them is not shown, and the palate is quite domed; the contact is along the dome.)

[7] Fant (p.c.) attributes this feature to the style of speech used for the still X-rays.

Other sources for Russian velars, which do not show minimal pairs, include Oliverius (1974) and Bolla (1981, 1982). Dem'janenko (1966) and Matusevic & Ljubimova (1964) present data from several different publications. From sources such as Oliverius (1974) and Bolla (1982), as well as from Koneczna & Zawadowski (1956), the articulation of a palatalized velar in a front vowel context can be examined. In the latter source, the X-ray contact seen for /k^j/ is longer before /i/ than before /a/. Since one palatogram shows long central, as well as lateral, contact, the X-ray might also be reflecting both of these. (Recall that in front vowel contexts, static palatograms will probably show the lateral contact associated with the vowel, and are reliable only with respect to the central stop contact.) The X-ray tracings in Oliverius show contact entirely on the hard palate; the accompanying palatogram shows that both central and lateral contact are more fronted than for the plain velar in back vowel contexts; that is, the whole tongue, except for the tip, is moved forward, as for the vowel. Bolla's X-ray tracings are point schematics, and therefore not precise with respect to contact, but the accompanying palatograms and linguograms are useful, and show an occlusion rather far back on the tongue, and probably on the soft palate -- that is, the least far forward palatalized velars seen in the sources.

The same sources generally also show tracings of palatalized vs. non-palatalized labials and coronals. Such tracings indicate that palatalization quite generally involves a forward raising of the entire tongue blade, body, and root, with the body approaching the roof of the hard palate. Palatalized velars can be seen to have this same basic articulation, where fronting of the tongue body necessarily affects the location of the consonantal constriction.

Russian also has a glide /j/, and the relation between its articulation and palatalization is a key question. Skalozub's X-ray tracing and palatogram of /j/ before /a/ shows that it is much like the vowel [i] but with greater narrowing at the diagonal/roof border, along with an expanded pharynx, and greater side-to-side lateral contact -- thus, presumably, a more raised and fronted tongue body and blade. Fant's X-ray tracing shows long palatal contact, which is surely lateral, not central, and the blade clearly raised up. Bolla's palatogram of /j/ before /u/ shows an even lateral contact all along the molars. Thus some but not all of these tokens of /j/ appear to have a coronal component. At the same time, however, the X-ray sources by and large do not show blade raising in /k^j/ as they do for /j/. Also, the pattern of lateral contact for the palatalized velars in the sources is more variable, since the vowel context is variable. In an /a/ context they show either no lateral contact in front of the occlusion (Skalozub) or lateral contact only along the molars (Oliverius), but in /i/ contexts the palatograms show (as expected) a great deal more lateral contact, almost as much as for /j/. In the /i/ context, of course, the palatograms confound the consonant and vowel articulations. That is, there is no clear evidence for extensive lateral contact, or for a coronal articulation, in the palatalized velars, unlike for /j/.

Taking into account all of these sources, the general pattern in Russian is that the contact for palatalized velars is just in front of that for non- palatalized velars. The non-palatalized velar is usually articulated toward the front of the soft palate, and the palatalized velar is usually articulated toward the back of the hard palate. Both the lateral and the central contacts are shifted forward for the palatalized velar. Furthermore, the entire tongue is shifted forward, giving a larger pharyngeal cavity for the palatalized velar. The contact for the palatalized velar is usually the same length in the sagittal dimension as the contact for the non-palatalized velar, but the lateral contact may be longer. Palatalized velars involve a fronted tongue, but not generally a raised tongue blade.

A recent electro-palatographic (EPG) study of Irish consonants (Ni Chasaide & Fealy 1991) provides data on another contrast between palatalized and non-palatalized velars. Not much can be said about the difference in overall contact for Irish /gʲ/ vs. /g/ because the EPG pseudo-palate does not extend far enough back on the speaker's palate to record much /g/ contact. What can be seen is that at closure onset both /g/ and /gʲ/ are fronted in an /i/ context, compared to /ɑ/ and /u/ contexts, and that for /gʲ/ the front of the tongue is also raised. However, this fronting in the /i/ context is not seen just prior to release of closure. This suggests that between /i/s both velars reach their places of articulation at closure onset, and that in other vowel contexts the velar constriction moves forward during the closure interval. This is consistent with the idea that secondary articulations tend to be stronger at release than at closure onset.

2.5. English

Surprisingly little data on velar fronting is available for English. MacNeilage & DeClerk (1969) collected emg and X-ray data for /g/ before the vowels /i,u,æ, and ɔ/. They show one figure from the X-ray data, combining tracings for the four contexts. However, these are from 100 msec before the /g/ release, which is before the stop closure; this means that stop contact cannot be seen in the figure. Judging from this figure, it appears that /g/ will have its contact in similar locations before /u,æ,ɔ/, and more than a cm in front of that before /i/. However, unpublished X-ray microbeam data made available by Mary Beckman suggests a more forward release before /æ/ than before /a/. Figure 5 shows our estimates of contact for /g/ before /i/ vs. /u/. Similarly, Ladefoged (1982:58) shows palatograms for [k̟i] and [ku], along with inferred sagittal sections; however, it must be borne in mind that static palatograms confound information from vowel and consonant.

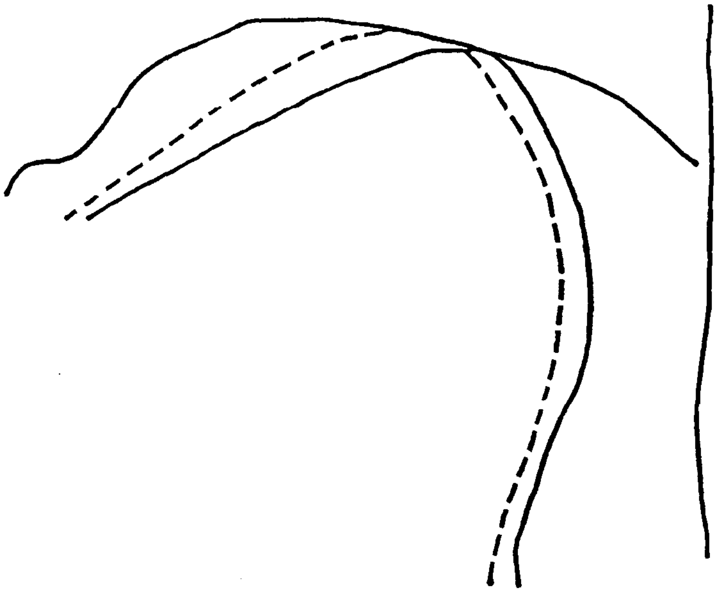


Figure 5. Estimated superimposed tracings of sagittal X-rays of velar stops in English before /u/ (plain line) vs. /i/ (dashed line), estimated after data in MacNeilage & DeClerk (1969).

Another source of data, but not of X-ray tracings, is Houde's (1967) comparison of /g/ between /i/, /ɑ/, and /u/[8] . While he does not reproduce tracings of the tongue as a whole, he does reproduce, and discuss, several figures which show just the tongue dorsum and palate at different points in the VCVCV sequences, as well as measurements of pellet location. Again, the contact locations vary along the palate over about 8 mm, with /igi/ having the most forward location. The contact for /igi/ does not overlap with those for /ɑgɑ/ and /ugu/, whereas these two backer contacts do overlap in location. In these three sequences, where the velars are between two like vowels, there are two distinct /g/ movements seen, one vertical and one horizontal. First, tongue movements from vowels to velars show mainly a raising motion: the tongue retains the front-back position from the vowel, and raises to form the closure. Thus the location of the initial palatal contact for /g/ varies according to the vowel context. Second, there is a forward and backward horizontal movement of the tongue, which Houde calls a "palatal closure-related perturbation". During the closure, the body moves forward along the palate, and after the release it moves downward and backward in the oral cavity. The horizontal movement component is larger (up to about 6 mm) when the context vowels are low. Houde is unsure whether this component is actively controlled by the speaker, or is passive[9]. Regardless, the effect is that the location of the tongue body for /g/ varies throughout its duration.

In VCVCV sequences in which the vowels are not the same, an additional movement component is seen during /g/. There is a vowel-to-vowel gesture whose direction depends on the frontness/backness of the vowels, so that it can be either front to back, or back to front. This gesture is executed simultaneously with the vertical vowel-to-velar gesture, so that the two are effectively added together. Houde refers to these two gestures as the "target- directed component" of the velar movement; these are components directed towards the velar target and the vowel target. The vowel-to-vowel movement is seen both as the tongue is raised toward the palate, and during the closure interval.

The vowel-to-vowel gesture from a front to a back vowel is obscured during /g/, however, because in those cases its effect is countered by the independent forward perturbation during stop closure. The two horizontal components -- forward for the /g/, backward for the second vowel -- can simply cancel each other out. On the other hand, when the vowel-to-vowel gesture is from back to front, the two horizontal components are in the same direction, and when added together result in a very pronounced forward movement of the tongue. The largest changes in contact location for /g/, up to 1 cm, are seen in these cases.

Houde's analysis of observed movements of the tongue body for velars is thus that they are the sum of three components: the vowel-to-vowel gesture (if any), the vertical velar gesture, and the closure-related forward/backward perturbation.

What is important here for our purposes is that these data indicate that velar fronting, at least in English, is a more or less continuous process during closure. The forward movement of the tongue is distributed over time. It is not the case that the entire

[8] In this study, pellets were attached to the midline of the tongue body, and their positions were measured over time from the frames of a cine film (100 frames per second). The overall tongue contour was also visible. VCVCV utterances combined the vowels /i,ɑ,u/ with /g/ or with /b/, under different rate and stress conditions.

[9] Ohala (1983) interprets the closure-induced fronting of /g/ in Houde's data as an active voicing-facilitating mechanism of back-cavity expansion. Houde also considers this interpretation but finds it less convincing, and in the case of /b/ argues directly that the expansion is passive, not active.

velar contact is equally fronted before a front vowel; rather, it becomes more front during the closure, so that by the moment of release it is noticeably fronted. In a similar vein, Ohman (1966) claims that in Swedish /ygɑ/ the position for the /g/ slides from front to back during contact, but no data or detailed discussion are presented. (Houde cites this result as agreeing with his own.)

Such articulatory gradience has been hypothesized to be a characteristic of surface underspecification by Keating (1988, 1990). On this view, the fact that the backness of velar contact varies during closure according to vowel context is taken to show that the velar itself is relatively unspecified on this dimension. That is, the velars of English and Swedish would be specified as being articulated by the raised tongue dorsum, but the tongue's position would not be specified as back or front. The location of the velar contact would be determined by the phonetic context and thus would vary over the hard and soft palates as for vowels.

We should note that this kind of underspecification of tongue body backness is a property of most consonants, not just of velars. For example, the X-ray tracings of /t/ before different vowels in Perkell (1969) show marked variation in tongue body position during consonant closure. Labials probably show even more variation in tongue body position across vowel contexts. That is, context effects on tongue body position are probably quite general. What is special about velars in this regard is that the tongue body is the primary articulator, so that variation in tongue body position is also variation in the stop constriction location. For velars, variation in tongue body location affects the size of the resonating cavity in front of the constriction, and thus affects the audible characteristics of the consonant noise. For labials and coronals, tongue body movement does not affect the size of the front cavity, and therefore does not affect the release acoustics as much.

Many sources in the literature, on a variety of languages other than English, provide examples of velars in various vowel contexts, usually /i a u/. Besides Ohman (1967), these include Wierzchowska (1980) for Polish; Dukelski (1960) for Rumanian; Miletič (1960) for Serbian; Sovijärvi (1963) for Finnish; Warnant (1956) for Wallon; Chlumsky (1938), Straka (1965), Simon (1967) and Dem'janenko (1966) for French; Bothorel (1982) for Breton. A few sources show velars in five different vowel contexts (e.g. Dukelski, Miletič). These data, with more vowel contexts, show that the central contact for velars varies in location according to the frontness of the contextual vowel. The more front the vowel, the more front the velar. It would be helpful to know whether these fronted velars consistently have lateral contact in front of the central occlusion, but such information cannot be gleaned from X-rays and static palatography. Similarly, it cannot be determined from the available data whether the point of contact changed during the occlusion interval.

2.6. Discussion and Comparisons

All of the sound types discussed above -- palatals, palatalized velars, and fronted velars -- are articulated on the hard palate. Thus all of them can be said to be "palatal" in the traditional sense (see Recasens 1990). However, the palatals of Czech and Hungarian have a much more forward articulation on the hard palate, and contact more of its surface. It also makes sense to reserve the term "palatal" for this sort of articulation, as is often done in more phonemically-oriented descriptions of languages.

Fronting of velars between or before front vowels appears to be a robust general trend across the languages surveyed here. Although details differ across languages, back vowel contexts generally give rise to velar articulations on the soft palate, while front vowel contexts generally give rise to velar articulations on the hard palate. The precise location of the contact varies with vowel frontness. Crucially, this influence of vowel on consonant appears to be temporally gradient. In English and Swedish, when two different vowels flank the velar consonant, then the velar contact moves between the locations of the two vowels during its closure interval. We expect that this is more generally the case across languages.

Following Keating's surface underspecification approach (Keating 1988, 1990), we hypothesize that fronting of velars is not indicative of a specified feature value for frontness (e.g. [-back] or Coronal) on the velar. Instead, velars, like other consonants, are generally unspecified for the relevant feature, with a range ("window") of possible tongue positions that is about the same as the front-back dimension of the vowel space. Observed fronting arises from phonetic implementation of the specified vowel-unspecified velar-specified vowel sequence, as described by Houde. This analysis is necessarily tentative, since the data come only from sagittal X-rays. We hope that further relevant data on the time course of overall velar contact will emerge via other techniques.

A corollary of this account of velar fronting is that there is no sense in which back velars are basic and front velars special; all are equally dependent on context. That is, it would appear that there is no reason we should speak of velar "fronting" (but not "backing") in the first place. We can think of two reasons why fronting might appear more salient than backing to introspective linguists. The more obvious reason is the closure-related fronting for velars described by Houde, a fronting that is enhanced by vowel- to-vowel fronting but potentially canceled by vowel-to-vowel backing. The less obvious reason is that front velars might have a special kind of release due to their lateral contact. Non-low front vowels, and especially [i], generally have lateral contact, and introspection suggests that this contact is probably coarticulated with an adjacent velar as well. If this is so, the release of the velar consonant could be central only, with the lateral contact held from the consonant into the vowel. Such a release will entail a smaller-than-expected front cavity (because of the long central channel in front of the occlusion) plus a good chance of affrication (because of the small opening). This sort of coarticulated release would be a special characteristic of [ki] (and presumably [ci] as well). This idea is completely speculative on our part, since evidence from X-rays and static palatography does not bear on this point.

The palatalized velars of Russian are generally like very fronted velars. They have similar points and lengths of occlusion involving similar parts of the tongue. Their X-ray profiles are like those of velars before [i] in other languages. On the other hand, the palatalized velars before [a] seem not to show the pattern of extensive lateral contact seen in, e.g., [j]. That is, we think that the palatalized velars before most vowels have an occlusion location like that of velars before /i/ in other languages, but without the lateral contact that usually accompanies such occlusion. We suspect, but cannot show from available data, that Russian [k^{j}i] does have extensive lateral contact during the consonant as well as the vowel, and that it shares with [ki] in other languages a central-only release. In this respect, then, [k^{j}i] would be different from the other palatalized velars of Russian.

Palatal stops -- by which we mean stops of the type discussed here for Czech and Hungarian -- are consistently articulated on the hard palate with both the blade and the body of the tongue. The contact seen in X-rays is very long (about three times that seen for velars), but that contact is mainly lateral contact extending back from the stop occlusion onto the tongue body. Comparison of the palatograms and linguograms for Czech and Hungarian indicates that the stop occlusion might be somewhat backer and longer in Hungarian than in Czech, extending further back on the hard palate and involving more of the tongue body and less of the blade. The lateral contact in the two languages' palatals appears quite similar in the palatograms, but perhaps less extensive in Hungarian according to the X-rays. We have raised the possibility that these differences are due to different vowel contexts in the two language samples. Nonetheless, the palatals of both languages have a more forward articulation than for even palatalized or fronted velars. The stop occlusion is much further forward on the hard palate (on the diagonal rather than on the roof), and the lateral contact is much more extensive. For velars, the blade at most contributes lateral contact in front of the occlusion, while the body forms the occlusion. For palatals the blade forms the occlusion while the body contributes lateral contact behind the occlusion.

These palatals of Hungarian are more back than those of Czech, but they are still more front than Russian palatalized velars, and they have the extensive lateral contact (extensive in two dimensions) associated with palatals but not palatalized velars.

Thus we have seen that palatals stops, but not fronted or palatalized velars, have a coronal component. This coronal component is more pronounced in Czech than in Hungarian, but it is present in both languages. Palatals are also clearly postalveolar (non-anterior) and laminal (distributed). Palatals also have extensive lateral contact, due to tongue body raising and fronting in the hard palate region; this amounts to saying that they involve palatalization. Therefore palatals must also have the features of palatalization, at least in a fully specified surface representation. This second articulatory component is not just a passive consequence of the primary postalveolar articulation, since other postalveolar sounds do not have it (Keating 1991). The presence of two different articulations by different active articulators makes the palatals "complex segments" in the sense of Sagey (1986) and subsequent work. In this characterization, then, we disagree with Recasens (1990), who claims that adjacent parts of the tongue cannot be separately controlled and that therefore segments such as palatals cannot be coronal and dorsal at once. We consider palatalization in coronals, including this palatalization component in palatals, to be an instance of such complex articulatory structure.

Recasens has two main differences with us. First, he objects to the proposal that long continuous articulations produce "complex" segments. At one level this issue is terminological. Phonologists call "complex" those segments involving two different active articulators, such as (parts of) the blade and body of the tongue. Whether those articulators are spatially contiguous, or whether it is easy to deploy them simultaneously, are not criteria for the phonological complexity of segments. As opposed to this particular notion of complex segments, Recasens probably intended a more physical phonetic sense of segment complexity, based on the difficulty of executing combinations of articulations. There is certainly a real issue here, whether the physical articulatory demands of segments have phonological consequences, but it is not the issue being addressed here.

Second, Recasens distinguishes more regions along the hard palate and then shows that more phonetic distinctions can be made among sounds articulated there; he wants phonetic descriptions to maintain these distinctions. We have no quarrel with this. Our description here is a broader one, such as might be relevant to a phonology. We focus on larger differences in the active and passive articulators in place of articulation, so that e.g. differences attributable to consonant manner, such as stop vs. continuant, are collapsed in our description.

3. Acoustic comparisons

3.1. General considerations

We turn now to the acoustic characterization of the three consonant types. The first question of interest is whether the fronted velars are acoustically distinct from the palatals. The second question of interest is whether the fronted velars are acoustically distinct from the palatalized velars. Our acoustic data are spectra of the stop consonant release.

3.2. Method

Acoustic data were collected from one male speaker of each of the languages under consideration (Hungarian, Czech, Russian, English). The Hungarian sample is the one used by Blumstein (1986). Palatals and velars were recorded for Hungarian and Czech; plain and palatalized velars for Russian; and velars for English. The consonants are voiceless in all cases; though there is more aspiration in English than in the other three languages, all are sufficiently aspirated that there is no voicing source at the release. Five following vowel contexts were used, broadly /i e a o u/. Different numbers of repetitions were available for each language: five in Czech and Hungarian, four in Russian, and three in English. The Russian speaker recorded was one who had no objections to producing the combination plain /k/ plus /i/ (phonetically [kɨ]) and plain /k/ plus /e/, as if it were the first syllable of the the preposition /k/ followed by a vowel-initial word. The speech samples were digitized at 10 kHz and filtered at 4.8 kHz, and autocorrelation LPC spectra were computed with a 25.6 ms full-Hamming window and 14 coefficients. One spectrum was computed for the consonant release by centering the window at the burst onset (so that half the window , or 13.2 ms, spans the burst, as in Blumstein and Stevens 1979); another was computed for the vowel onset by positioning the left edge of the window at the beginning of the first full glottal pulses. These two spectra for each CV were then displayed together; relative amplitude is preserved within a CV token. The data are presented in Figures 6-9 as superposed tracings of the 2 spectra for all the repetitions: plain velars in Figures 6-8; palatalized velars in Figure 8, and palatals in Figure 9.

Spectra can be characterized as compact, with a main spectral peak at mid (to low) frequencies, or as diffuse, with energy more evenly distributed (Jakobson et al. 1963). The dominant peak in a compact spectrum (or the highest-energy peak in a diffuse spectrum) can be further characterized in terms of its frequency location, and in terms of its correspondence to resonances in, e.g., a following vowel (Fant 1960). The consonant spectrum can also be characterized relative to a following vowel in terms of overall distribution of energy, and changes in that distribution (Kewley-Port 1983; Lahiri et al. 1984; Stevens & Keyser 1989, Halle & Stevens 1990). Our discussion of the data focuses on the first two of these[10] .

3.3. Results

[10] Another way of looking at spectra is discussed by Stevens & Keyser (1989), in terms of F2 and the feature Back. Within a segment, an F2 raised near F3, possibly merging with it auditorily, is the main acoustic correlate of the feature value [-back]. The feature Back is used to represent, roughly, the difference between soft palate vs. hard palate tongue body articulations. Therefore we would expect all but the non-fronted velars to show this kind of raised F2 merged with F3. However, this acoustic correlate cannot be applied to consonant release spectra, since these show primarily the resonances of the cavity in front of the constriction, and either F2 or F3 will be a back cavity resonance in the velar region.

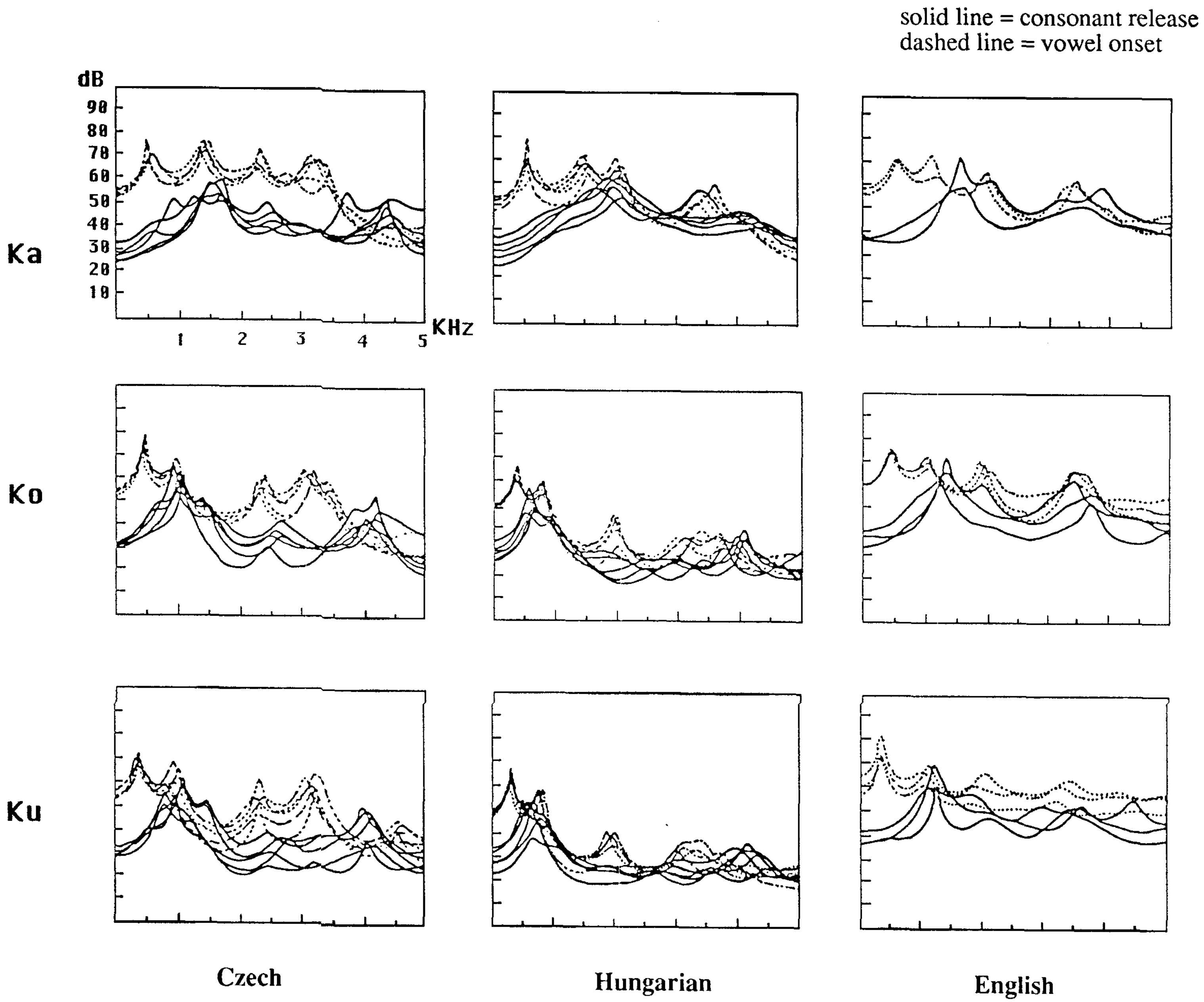


Figure 6. LPC spectra for stop release burst (solid lines) and vowel onset (dashed lines) of velars before **back vowels** in Czech,Hungarian, and English. Individual spectra for each token are superimposed.

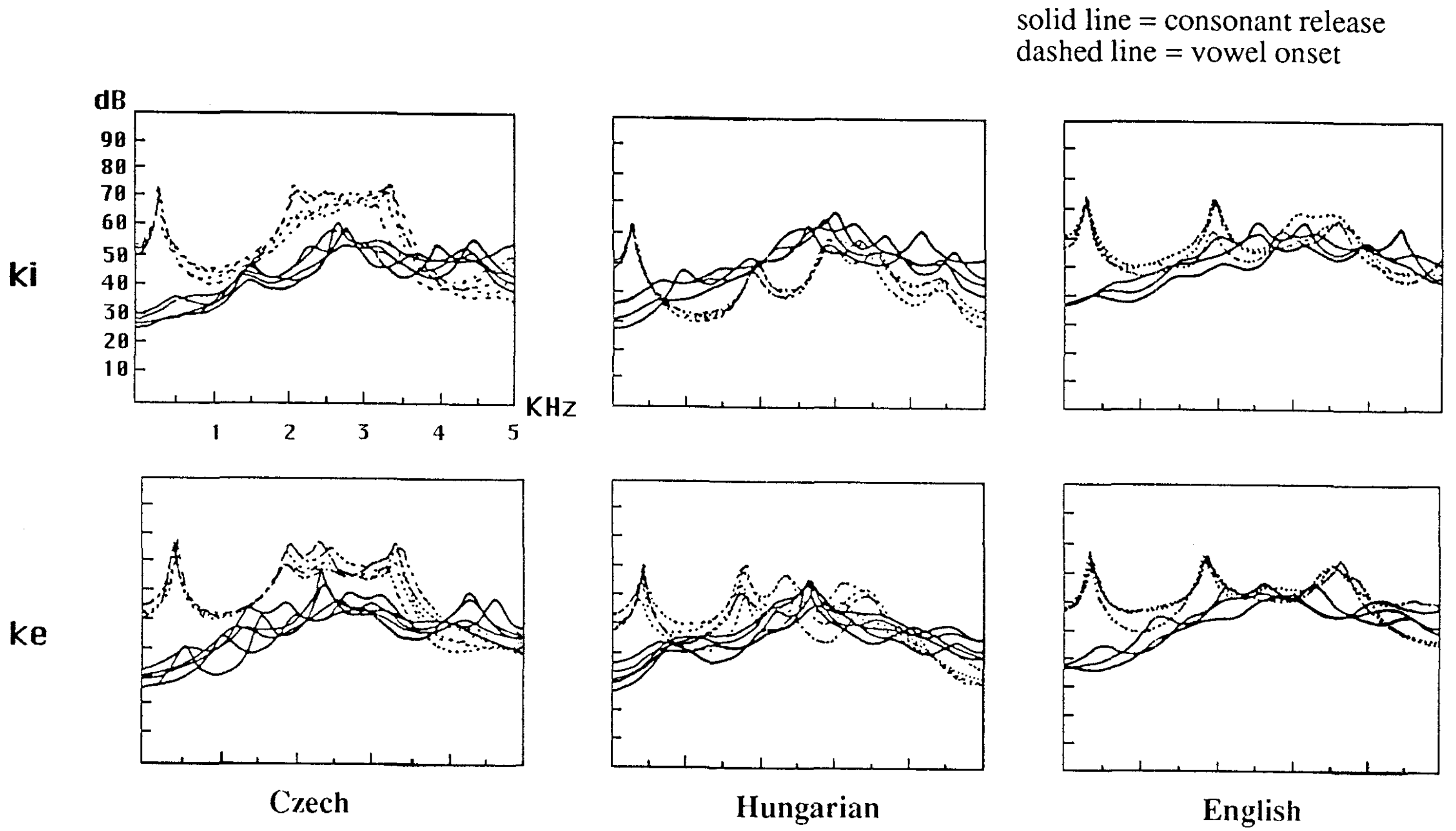
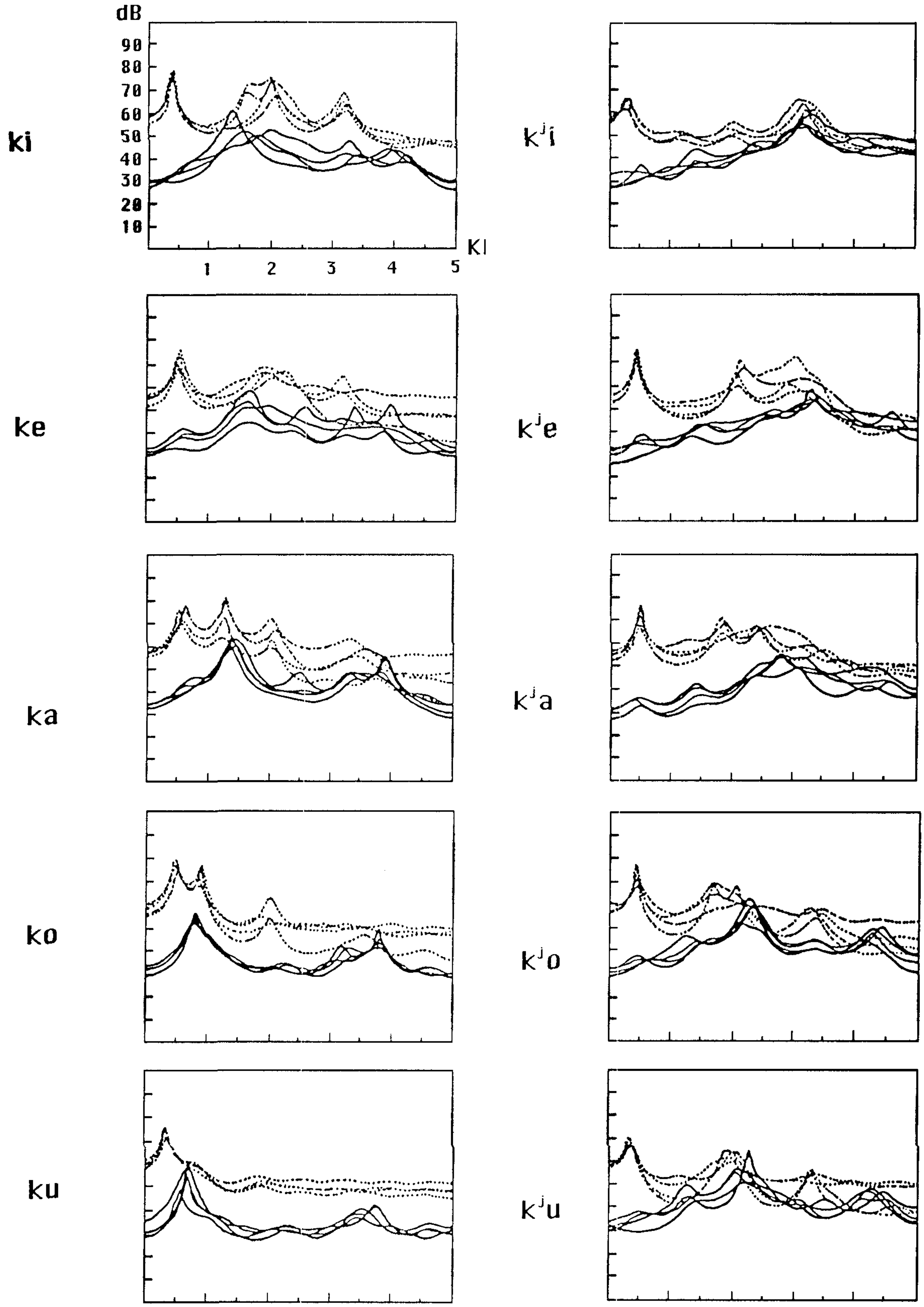


Figure 7. LPC spectra for stop release burst (solid lines) and vowel onset (dashed lines) of velars before front vowels in Czech, Hungarian and English. Individual spectra for each token are superimposed.



solid line = consonant release
dashed line = vowel onset

Figure 8. LPC spectra for stop release burst (solid lines) and vowel onset (dashed lines) of velars before five vowels in Russian. Left panel, nonpalatalized stops; right panel, palatalized stops. Individual spectra for each token are superimposed.

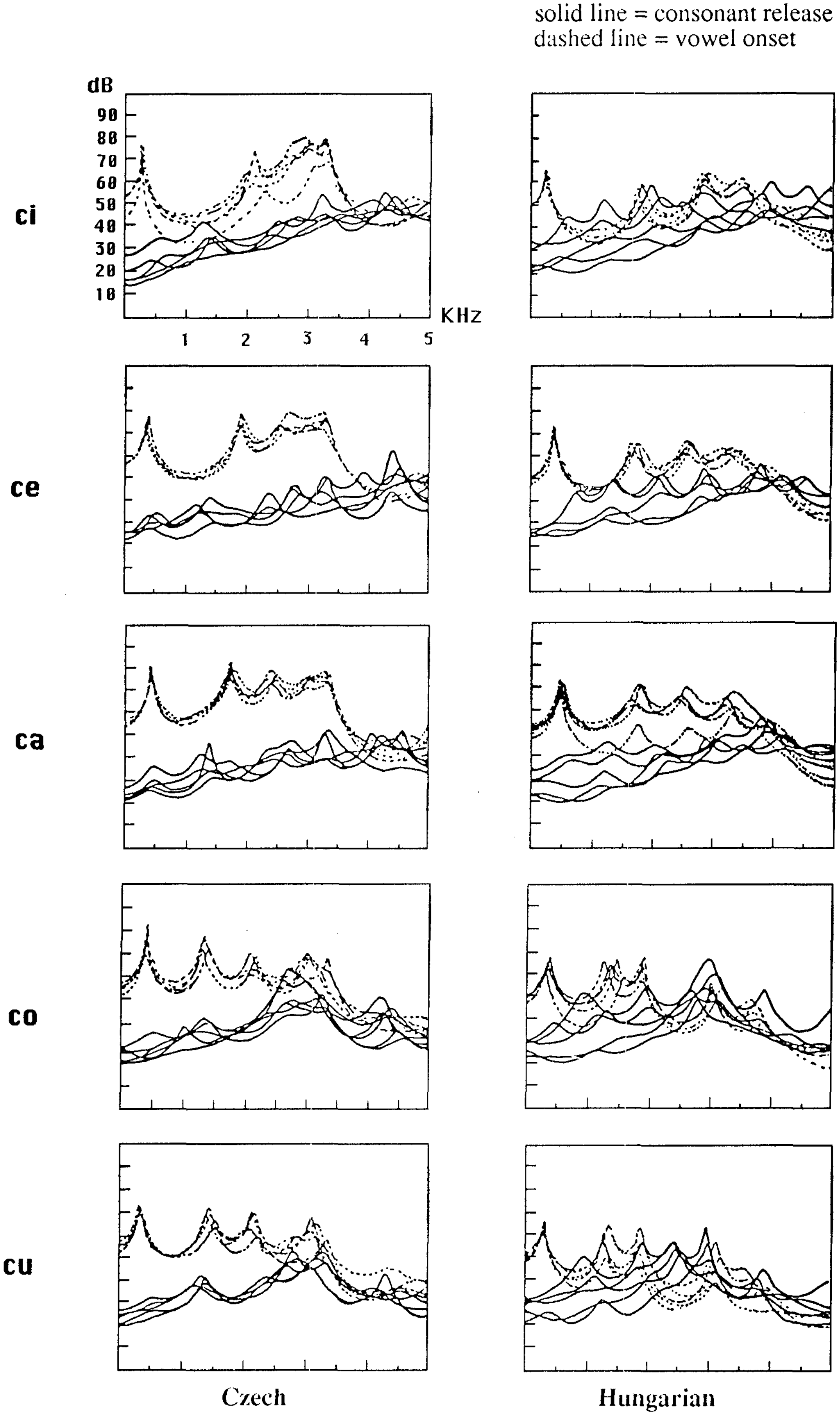


Figure 9. LPC spectra for stop release burst (solid lines) and vowel onset (dashed lines) of palatals before five vowels in Czech (left panel) and Hungarian (right panel).

The data for non-palatalized velars can be seen in Figures 6 and 7. In our data, as in many previous studies, the spectrum at release of a velar typically is compact, with a prominent peak at about the second or third formant frequency of, and at a similar intensity as, the following vowel. The strongest peak stands out clearly from any other peaks, dominating the spectrum. For example, in Figure 6 the English velar spectra have 2 peaks, one at 4000 Hz and another more prominent one in the region of the vowel's F2. The spectrum around this peak is often, but not always, low and flat.

The main peak in a consonant release spectrum is a front cavity resonance whose frequency value largely depends on the following vowel. Thus the velar spectra differ according to the vowel context. This is true not only in terms of absolute frequency values, but also in terms of the correspondence between the consonant peak and formants of the following vowel. Velars in front vowel contexts, seen in Figure 7, have their strongest peaks higher than the vowel's F2 but lower than the vowel's F4. (Russian patterns differently and will be discussed below.) In Czech the peaks in /ki/ and /ke/ fall at around 2800 and 2500 Hz respectively, and often correspond to the vowel's F3. In Hungarian peaks at 3000 and 2700 Hz sometimes lie above the vowel's F3. Velars before /i/ and /e/ in English generally have peaks at or below F3. Comparing Figures 6 and 7, in these three data sets the velars before front vowels differ from the velars before back vowels in this respect. In back vowel contexts, the strongest consonant peak usually lies at or below F2 of the vowel, while in front vowel contexts it is consistently higher. The difference is less in English than in Czech or Hungarian, but on average it holds.

In this respect, our data on velars are comparable to results of previous studies (e.g. Zue 1976, Sereno & Lieberman 1987 on English). In particular, discussions of velars in Swedish by Fant (1973) and Danish by Fischer-Jørgensen (1954) make similar points about the way in which the major peak in the consonant's release spectrum depends on the adjacent vowel's higher formants, and the way in which the compactness of the spectrum varies across vowel contexts. Minifie (1973) also discusses the fact that the release spectrum for velars will vary along with the contextually-determined place of constriction.

A further effect of the front vowel contexts is that the consonant spectrum is less compact. Particularly for Czech /ki/ and /ke/, and Hungarian /ki/, the most prominent peak is still in the mid-frequency range, but this peak does not dominate the spectrum, since surrounding peaks are also at high levels. Also, because the main velar peak is at a higher frequency in Hungarian, it is nearer to the secondary peak above it.This gives the fronted velars in the three languages somewhat different overall spectra, with a more prominent main peak in Hungarian and Czech, and a flatter spectrum around the peak in English.

Russian differs from the other languages with respect to velar fronting. Data for the non-palatalized velars are presented in the left panel of Figure 8. Recall that phonemic /i/ is phonetically [ɨ] here. The only non-palatalized Russian velar in a phonetic front vowel context is /ke/. This velar behaves just like non-palatalized velars before central or back vowels. All of the non-palatalized velars have their strongest peak at or below the vowel's F2. Thus it appears that the Russian non-palatalized velar is not fronted before the mid front vowel the way the velars of the other languages in Fig. 7 are.

The data for Russian palatalized velars can be seen in the right panel of Figure 8. These are like the fronted velars in having a higher-frequency consonant peak; in fact, the peak is even higher for the palatalized velars. At the same time, the palatalized velars are like plain, backed velars in having decidedly compact spectra, consistently flat around the main peak. Fant (1960) notes that all Russian palatalized consonants have compact spectra. Vowel context effects can also be seen with palatalized velars. Before /i/ and /e/, these velars pattern like some of the Hungarian and Czech palatals, with their strongest peak at or above the vowel's F4. (The Russian peaks are generally lower, but there is overlap in the distributions.) Before /a/, /o/, and /u/ Russian palatalized velars pattern like the fronted velars in Czech and Hungarian, with their strongest peaks at or above the vowel's F3. That is, palatalized velars always have a fairly high frequency peak, but it is higher -- in terms of frequency and formant correspondence -- before front vowels than back.

Data for palatals can be seen in Figure 9. Before unrounded vowels (front vowels and /a/), palatals show spectra which slope up to a highest peak at 3-4 kHz or even higher. These spectra do not always have a single dominant peak: there may be a few peaks of similar amplitude which together dominate the spectrum in a single broad region. For these palatals, the energy of the release spectrum is strongly concentrated at higher frequencies. We consider these spectra marginally compact. Before back rounded vowels, the consonant spectra are quite compact. The main peak for /cu/ and /co/ is usually in the range of 2500-3000 Hz, which corresponds roughly to F4 of the following vowel.

In sum, plain velars before back vowels have compact spectrum whose main peak is low in frequency and lies near F2 of the vowel. Plain velars before front vowels have a less-compact spectrum whose main peak is higher in frequency and lies near F3 of the vowel. Palatalized velars have a compact spectrum whose main peak is mid to high in frequency and lies near F3 (back vowel contexts) or F4 (front vowel contexts) of the vowel. Palatals before back vowels have a compact spectrum whose main peak is at a mid frequency and lies near F4 of the vowel. Palatals before front vowels have a somewhat more diffuse spectrum whose strongest peak is high in frequency and lies near F4 of the vowel. Thus the front velars are distinguished from the palatals by spectral shape and location of main peak. Palatalized velars are distinguished from the palatals in back vowel contexts by the location of the consonant peak, and in front vowel contexts by spectral shape: the palatalized velars are compact while the palatals are more diffuse. Thus, when release spectra are considered in relation to vowel-onset spectra, all of the phonetic types under discussion can be distinguished.

3.4. Discussion

In Hungarian and Czech, velars are in contrast with palatals in all vowel contexts, and thus it is not surprising that there are clear acoustic differences between them. Furthermore, the velars before front vowels are clearly different from the velars before back vowels. Not only is the consonant peak higher in frequency, but it corresponds to a higher formant of the vowel. This difference between velars in front vs. back vowel contexts is greater in Czech and Hungarian than in English. That is, Hungarian and Czech show acoustic evidence of more velar fronting than English. This is perhaps surprising in that in front vowel contexts the fronted velars contrast with palatals; we might have expected the velars to be somewhat less fronted to keep the contrast clear. However, the fact that the spectra do indicate velar fronting in surface contrast with palatals shows clearly that fronted velars cannot be equated with, and represented as, palatals.

In Russian, /k/ before /e/ is not a fronted velar. This result is consistent with the view that Russian non-palatalized velars have a secondary articulation of their own, namely, velarization, which consists of a mandatory backed tongue body. This articulatory requirement prevents the velars from assuming a fronted position in a front context. This finding clarifies how it is that velars can be velarized, a seeming tautology: velarization of velars is the absence of velar fronting.

The Russian palatalized velars before front vowels are clearly more fronted at their release than are velars before front vowels in the other languages, especially English. In none of these three languages does the velar's peak correspond to the vowel's F4. At the point of release, then, the palatalized velars are apparently more front than are the fronted velars before [i], especially in English. This difference may be related to the claim that only the palatalized velars have a positive feature specification for tongue backness: this results in a clear fronting of the tongue body which reaches a more extreme position.

Because palatalized velars are always fronted, the difference between palatalized and non-palatalized velars will be largest before back vowels. In back vowel contexts, the distinction can be cast in terms of whether the main spectral peak of the consonant release is at the vowel's F3 (in which case the consonant is palatalized) or F2 (in which case the consonant is not palatalized). In front vowel contexts, the distinction would be at best subtle. Presumably because of this difficulty, Russian non-palatalized velars do not occur before front vowels within words.

4. General Discussion and Conclusions

4.1. Comparison of articulatory and acoustic data

In general terms, the more fronted a constriction in the vocal tract, the smaller the front cavity, and the higher the frequency of its main spectral peak -- back velars having their main peak at the frequency of the vowel's (lowered) F2, palatals having their main peak at the frequency of the vowel's F4. This relation is orderly in our data, with front velars falling between back velars and palatals. There is, however, one apparent exception -- palatalized velars before front vowels have their main spectral peak at the vowel's F4, like the palatals. From this we can infer that the front cavities of palatalized velars in front vowel contexts are quite small, yet none of the articulatory data suggest very fronted occlusions. Most likely, the front cavity of these velars is decreased in size by lateral contact in front of the occlusion, which is seen in some though not all of the articulatory sources. That is, the sides of the tongue blade may be forming a kind of secondary constriction in front of the stop occlusion. Thus palatalized velars before front vowels may be weakly coronal in articulation, accounting for their acoustic similarity to the coronal palatals.

At the same time, it must be stressed that the palatalized velar spectra with peaks near F4 are distinct from the palatal spectra. In particular, the palatalized velar spectra are more compact, with palatals before front vowels generally having more than one high-energy peak in a high-frequency band. It seems clear from the articulatory data that the primary articulatory constrictions are much more front for palatals than for palatalized velars, so it is to be expected that they will be acoustically distinct. What is not clear from the articulatory data is why they are not even more distinct acoustically.

Another question raised by the articulatory data was the possibility that the Czech palatals might be systematically more front than the Hungarian palatals. The acoustic data, where vowel contexts are as matched as possible across the languages, show no obvious difference in this respect. There are noticeable differences between the two panels of spectra in Figure 9, but these have to do with the vowels, not with the consonants or with the relation between consonants and vowels. This result lends weight to the alternative suggestion, that the articulatory data for Czech came from fronter vowel contexts.

4.2. Summary

We have compared articulatory and acoustic data on back velars, contextually fronted velars, palatalized velars, and palatals to determine whether all of these consonant categories can be phonetically distinguished. The data suggest that all of them differ. First, the phonemic palatals have occlusions which are distinctly more fronted than the others, and which are made with the tongue blade as well as the tongue body. Neither fronted nor palatalized velars have such a forward occlusion, though palatalized velars, especially before [i], may have quite fronted lateral contact. Acoustically, the palatalized velars show evidence of strong fronting. From these results we conclude that palatals are coronal, front velars are not, and palatalized velars may be weakly so.

Second, fronted and palatalized velars are both fronted along the palate, but in different ways. Palatalized velars appear to be more fronted at release than contextually fronted velars. Furthermore, contextual fronting appears to be a continuous, gradient effect of context: the constriction location moves continuously during the closure, from a position more dependent on the preceding segment to a position more dependent on the following segment. We also saw that Russian non-palatalized velars fail to show evidence of contextual fronting. From these results we conclude that velar stops in Russian have specifications for tongue body frontness or backness, but that velars in other languages, where palatalization is not contrastive, are not specified for front vs. back tongue positions, even in surface forms. Rather, their fronting results from phonetic implementation.

One of our initial questions was whether contextually fronted velars, articulated on the palate, should be (featurally) represented like the palatals of Czech and Hungarian. We showed that such sounds contrast on the surface in Czech and Hungarian, and so cannot be identified. Furthermore, the hypothesis that emerges from this study is that contextual fronting of the tongue body during consonants should not be represented featurally at all, and thus the question is answered negatively. Another question was whether contrastively front velars, such as the palatalized velars of Russian and also other cases (as in footnote 1) should be represented like the Czech and Hungarian palatals. Again the answer is negative, in this case because of clear articulatory and acoustic differences.

ACKNOWLEDGEMENTS

The first author acknowledges the support of grant BNS 84-18580 from the NSF, and from the Max-Planck-Institute for Psycholinguistics in Fall 1988. The second author acknowledges the support of Brown University, where the audio recordings for Czech and Hungarian were made. Both authors thank Henning Reetz for his software for acoustic analysis; and Chris Golston, Larry Hyman, Ian Maddieson, Maureen Stone, Daniel Recasens, Aaron Shryock and Michael Shalev for helpful comments and/or assistance in ms. preparation.

REFERENCES

Akishina, A. A. Baranovskaja, S. A. *Russkaja fonetika* (Izdatel'stvo Russkij Jazyk, Moscow, 1980).

Blumstein, S. E. On acoustic invariance in speech. In J.S. Perkell & D.H. Klatt (eds.), *Invariance and variability in speech processes*, pp. 178-93 (Lawrence Erlbaum Assoc, Hillsdale NJ, 1986).

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

Bolla, K. A phonetic conspectus of Hungarian. The articulatory and acoustic features of Hungarian speech-sounds. *Magyar Fonetikai Fuzetek* (Hungarian Papers in Phonetics) 6 (1980).

Bolla, K. The articulation of Hungarian long consonants. *Magyar Fonetikai Fuzetek* (Hungarian Papers in Phonetics) 7:7-55 (1981a).

Bolla, K. *A conspectus of Russian speech sounds* (Bohlan, Cologne and Vienna, 1981b).

Bolla, K. A phonetic conspectus of Russian: The articulatory and acoustic features of Russian speech-sounds. *Magyar Fonetikai Fuzetek* (Hungarian Papers in Phonetics) 11 (1982).

Botherel, A. *Etude phonétique et phonologique du breton parlé à Argol* (Finistère-sud) (Université Lille III, Lille, 1982).

Bulanin, L. *Fonetika sovremennogo russkogo jazyka* (Izdatel'stvo Vyššaja Škola, Moscow, 1970).

Catford, J.C. *A practical introduction to phonetics* (Clarendon Press, Oxford, 1988).

Chadwick, N. A descriptive study of the Djingili language. *Australian Aboriginal Studies, Regional and Research Studies No. 2*. Australian Institute of Aboriginal Studies, Canberra (1975).

Chlumsky, J. *Radiographies des voyelles et des semivoyelles françaises* (Czech Academy, Prague, 1938).

Chomsky, N. & M. Halle. *The sound pattern of English* (Harper & Row, New York, 1968).

Clements, N. Place of articulation in consonants and vowels: a unified theory. In B. Laks & A. Rialland (eds.), *L'architecture et la geometrie des representations phonologiques* (Editions du C.N.R.S., Paris, 1991).

Daneš, F., B. Hála, A. Jedlička, & M. Romportl. *O mluvem slove* (Státnî Pedagogické Nakladatelství, Prague, 1954)

Dart, S. A bibliography of x-ray studies of speech. *UCLA Working Papers in Phonetics* 66: 1-97 (1987)

Dart,S. Artculatory and acoustic properties of apical and laminal articulations .*UCLA Working Papers in Phonetics* 79 (1991)

Dem'janenko, M.J. *Vstupnyi fonetyko-hrafichnyi kurs francuz'koi movi* (Laboratory of Experimental Phonetics, Kiev, 1966).

Dukelski, N.I. Cercetare foneticâ experimentalâ asupra palatalizârii şi a labializârii consoanelor romîneşti. *Fonetica sį Dialectologie* 2: 7-45 (1960).

Fant, G. *Acoustic theory of speech production* (Mouton, The Hague, 1960).

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

Fischer-Jørgensen. Acoustic analysis of stop consonants. *Miscellanea Phonetica* 2:42-59 (1954).

Furby, C. E. Garawa phonology. *Papers in Australian linguistics* 7: 1-11 (1974). Pacific Linguistics, Series A, No. 37 (Australian National University, Canberra).

Gorecka, A. Phonology of articulation. Ph.D. dissertation, MIT (1989).

Hála, B. K. popisu prazšké výslovnosti (Studie z experientální fonetiky). *Rozpravy Ceské Akademie Ved a Umeni* 3(56), Prague (1923).

Hála, B. Uvedení do fonetiky češtiny: *Na obecne fonetickém základe* (Nakladatelství Ceskoslovenske Akademie Ved, Prague, 1962).

Halle, M. & K. N. Stevens. The postalveolar fricatives of Polish. Ms., MIT, 1989.

R.-M. S. Heffner. *General Phonetics* (U. Wisconsin Press, Madison, 1950).

R. A. Houde. A Study of Tongue Body Motion During Selected Speech Sounds. U. Michigan Ph.D. dissertation, 1967.

Jakobson, R., G. Fant, & M. Halle. *Preliminaries to speech analysis* (MIT Press, Cambridge, MA, 1951, fourth printing 1963).

Jones, D. & D. Ward. *The phonetics of Russian* (Cambridge U. Press, Cambridge, 1969).

Keating, P. Palatals as complex segments: X-ray evidence. *UCLA Working Papers in Phonetics* 69: 77-91 (1988).

Keating, P. The window model of coarticulation: articulatory evidence. In J. Kingston & M. Beckman (eds.), *Papers in Laboratory Phonology I: Between the grammar and the physics of speech*, pp. 451-470 (Cambridge U. Press, Cambridge, 1990).

Keating, P. Coronal places of articulation. In C. Paradis & J.-F. Prunet (eds.), *The special status of coronals* (Phonetics and Phonology 2), pp. 29-48 (Academic Press, San Diego, 1991).

Kewley-Port, D. Time-varying features as correlates of place of articulation in stop consonants. *J. Acoust. Soc. Am.* 73:322-335 (1983).

Kirton, J. & B. Charlie. Seven articulatory positions in Yanyuwa consonants. *Pacific Linguistics*, Series A, No. 51 (Australian National University, Canberra, 1978).

Koneczna, H. & W. Zawadowski. Obrazy rentgenograficzne głosek rosyjskich. *Prace językoznawcze*, 9 (Panstowowe Wydawnictwo Naukowe, Warsaw, 1956).

Ladefoged, P. *A course in phonetics* (Harcourt, Brace, Jovanovich, New York, first edition 1975, second edition 1982).

Ladefoged, P. & I. Maddieson. Some of the sounds of the world's languages. *UCLA Working Papers in Phonetics* 64 (1986).

Lahiri, A., L. Gewirth, & S. Blumstein. A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. *J. Acoust. Soc. Am.* 76:391-404 (1984).

Lahiri, A. & V. Evers. Palatalization and coronality. In C. Paradis & J.-F. Prunet (eds.), *The special status of coronals* (Phonetics and Phonology 2), pp. 79-100(Academic Press, San Diego, 1991).

MacNeilage, P. F. & J. L. DeClerk. On the motor control of coarticulation in CVC monosyllables. *J. Acoust. Soc. Am.* 45:1217-1233 (1969).

Matusevič, M. I., & N. A. Ljubimova. Artikuljatsija russkix zvukov pod udareniem na osnove rentgenografičeskix dannyx. Voprosy Fonetiki. Učenye zapiski 325. *Serija filologičeskix* nauk 69:37-44 (1964).

Miletič, B. *Osnovi fonetike srpskog jezika* (Nau¢na Kniga, Belgrade, 1960).

Minifie, F.D. Speech Acoustics. In F. D. Minifie, T. J. Hixon & F. Williams (eds.), *Normal aspects of speech, hearing, and language*, pp. 235-284 (Prentice- Hall, Englewood Cliffs NJ, 1973).

Ni Chasaide, Ailbhe and Geraldine Fealy. Articulatory and acoustic measurements of coarticulation in Irish (Gaelic) stops. *Proceedings of the XIIth International Congress of Phonetic Sciences* 5:30-33 1991

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

Öhman, S. Coarticulation in VCV utterances: spectrographic measurements. *J. Acoust. Soc. Am.* 39:151-168 (1966).

Öhman, S. Numerical model of coarticulation. *J. Acoust. Soc. Am.* 41:310-320 (1967).

Oliverius, Z. F. *Fonetika Russkogo Jazyka* (Státní Pedagogické Nakladatelství, Prague, 1974).

Pacesova, J. *Fonetika a ortoepic češtiny* (Nákladem JAMU, Brno, 1969).

Perkell, J. S. *Physiology of speech production: Results and implications of a quantitative cineradiographic study* (MIT Press, Cambridge MA, 1969).

Pétursson, M. Les articulations de l'Islandais à la lumière de la radiocinématographie. *Société de Linguistique de Paris* 68 (Klincksieck, Paris, 1974).

Polland, B. & B. Hála. *Les radiographies de l'articulation des sons tchèques* (Prague, 1926a).

B. Polland & B. Hála. *Artikulaca českých zvukû v rentgenových obrazéch.* (NakladatelstvíČeske Akademie Ved, Prague, 1926b).

Recasens, D. The articulatory characteristics of palatal consonants. *J. Phonetics* 18: 267-280 (1990).

Sagey, E. W. The representation of features and relations in non-linear phonology. Ph.D. dissertation, MIT, 1986.

Sapir, E. *Language* (Harcourt, Brace and World, New York, 1921).

Sereno, J. & P. Lieberman. Developmental aspects of lingual coarticulation. *J. Phonetics* 15: 247-257 (1987).

Simon, P. Les consonnes françaises. *Biblioteque française et romane*, Serie A, No. 14 (Klincksieck, Paris, 1967).

Skalozub, L. G. *Palatogrammy i rentogenogrammy soglasnyx fonem russkogo literaturnogo jazyka* (Izdatel'stvo Kievskogo Universiteta, Kiev, 1963).

Skalozub, L. G. *Uprazhenija po fonetike russkogo jazyka.* (Laboratory of Experimental Phonetics, Kiev University, 1963).

Sovijärvi, A. *Suomen Kielen Aännekuvasto* (K.J. Gummerus Osakeyhtiö, Jyväskylässä, 1963).

Stevens, K.N. & S.J. Keyser. Primary features and their enhancement in consonants. *Language* 65: 81-106 (1989).

Stone, M. A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data. *J. Acoust. Soc. Am.* 87: 2207-2217 (1990).

Straka, G. L'évolution phonétique du latin au fran_ais sous l'effet de l'énergie et de la faiblesse articulatoires. *Travaux de linguistique et de littérature* 2(1):17-98 (1964). Centre de Philologie et de Littératures Romanes, Strasbourg University.

Straka, G. Naissance et disparition des consonnes palatales dans l'évolution du latin au français. *Travaux de linguistique et de littérature* 3(1):117-167 (1965). Centre de Philologie et de Littératures Romanes, Strasbourg University.

Warnant, L. *La constitution phonique du mot wallon* (Société d'édition Les Belles Lettres, Paris, 1956).

Wierzchowska, B. *Fonetyka i fonologia języka polskiego* (Ossolineum, Warsaw, 1980).

Zue, V. Acoustic characteristics of stop consonants: a controlled study. Ph.D. dissertation, MIT (1976); distributed by Indiana U. Linguistics Club.

Phonetic interpretation of voiceless nasals

Peter Ladefoged
Phonetics Laboratory, UCLA
and
P. Bhaskararao
Department of Linguistics, Deccan College, Pune 411006, India

Many languages in South East Asia have voiceless nasal consonants that contrast with their voiced counterparts. What has not been reported previously is that there are two distinct types of voiceless nasals. We will begin by considering the more well known type, found in languages such as Burmese. These voiceless nasals are usually said to have an open glottis for most of the articulation, but some voicing for the period just before the stricture is broken (Ladefoged 1971:11). They thus have two parts. The first is necessary to distinguish them from their voiced counterparts. The second distinguishes one voiceless nasal from another, by making the place of articulation more apparent; the voiced offglide from the nasal into the vowel has formant transitions that are characteristic of each place of articulation. Despite the small voiced portion, in most phonological treatments of these sounds, a voiceless nasal consonant is considered not as a sequence of two items, a voiceless segment followed by a voiced segment, but as a single voiceless unit with a low level phonetic rule inserting the voicing towards the end. Burmese is a textbook example, with forms as shown in Table 1 (from Ladefoged 1982:273).

Table 1. Burmese contrasts involving voiced and voiceless nasals (from Ladefoged, 1982:273 Tones are marked as: [ǎ] rising (traditionally “level”), [â] falling (“heavy”); and [a̰] “creaky”.

mâ̰	'lift up'	m̥â̰	'from'
nǎ	'pain'	n̥ǎ	'nose'
ɲǎ	'right'	ɲ̊ǎ	'considerate'
ŋâ	'fish'	ŋ̊â	'borrow'

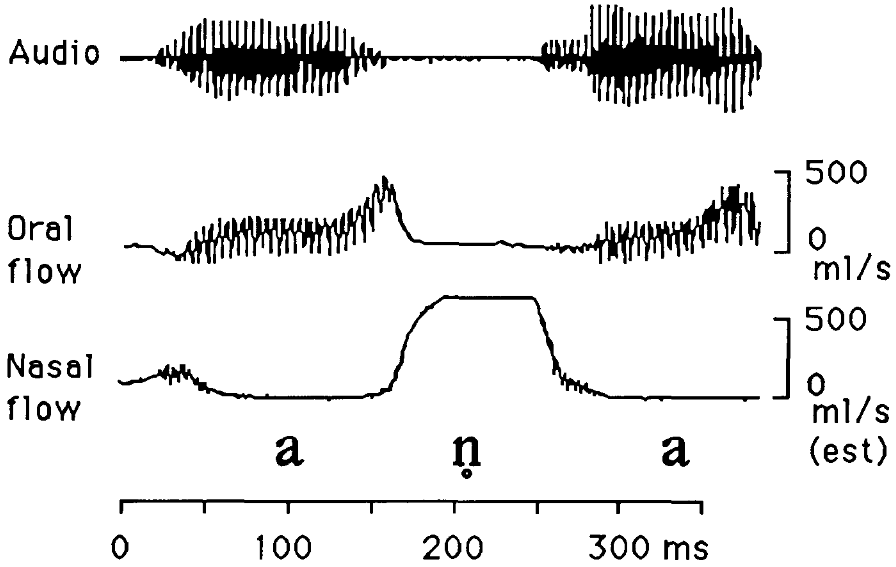


Figure 1. Aerodynamic records of the Burmese word [(a) n̥a] ‘nose’ that are typical of five of the six Burmese speakers.

Using instrumentation as described in Ladefoged (1991), we recorded the oral and nasal airflow during the pronunciation of these Burmese words when spoken in the sentence [ŋa ___ ko ye ne te] "I write ___". A typical recording of a word beginning with a voiceless nasal consonant is illustrated in Figure 1. There is considerable nasal airflow during the initial consonant in this word. During the latter part of this airflow voicing begins, so that there is an interval with a voiced nasal before the vowel.

Six Burmese speakers, three men and three women, all from Rangoon, Myanmar, were recorded in this way. Five of them had records very similar to that shown in Figure 1 for all four voiceless nasals. The sixth speaker was also similar in that he had a short period of voicing during the nasal consonant before the vowel; but in his case, for all four voiceless nasals, there was also a small release of air from the mouth before voicing commenced, as illustrated in Figure 2.

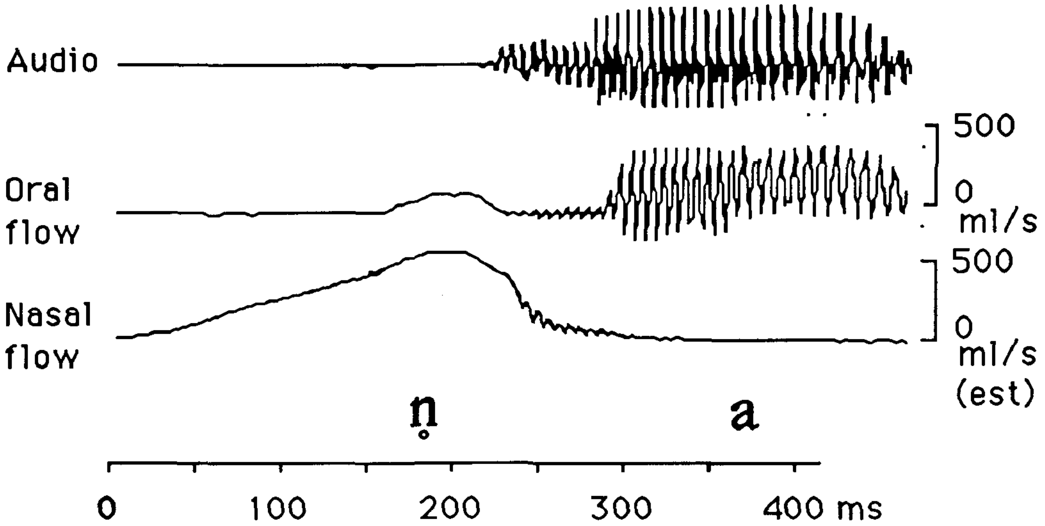


Figure 2. Aerodynamic records of the Burmese word [n̥a] 'nose' as produced by one of the six speakers.

Recordings made while speaking into an airflow mask are inevitably somewhat distorted. Accordingly we made a separate set of high quality audio recordings of the same speakers saying the same words in the frame sentence. Measurements were made of each of the parts of the voiceless nasals, as shown in Table 2. The part labeled "- voicing" was measured from the articulatory closure until the start of the voicing. This interval was often not entirely voiceless, as the voicing from the previous vowel (of the frame sentence) often continued for a few periods at the beginning of the closure. The data for the three male speakers are based on the audio recordings. It was not possible to segment the portions of the voiceless nasals reliably for the female speakers using these recordings, as they all used a more breathy voice quality during the voiceless nasal, making it more difficult to decide exactly where in this segment regular voicing began. For these speakers, the measurements reported in the table are based on the aerodynamic recordings.

Table 2. Durations (in ms) of the two parts, [- voice] and [+voice], of the voiceless nasals for six speakers of Burmese.

	bilabial		**dental**		**palatal**		**velar**		**MEAN**	
Voice	-	+	-	+	-	+	-	+	-	+

male 1	153	60	148	75	155	33	170	54	157	56
male 2	123	55	122	24	124	32	132	42	125	38
male 3	137	49	120	40	156	40	145	32	140	40
male mean	138	55	130	46	145	35	149	43	140	45
% voiced	28%		26%		19%		22%		24%	
female 1	220	32	174	90	192	43	?	?	195	55
female 2	226	22	176	51	203	42	?	?	202	38
female 3	142	22	144	64	152	67	?	?	146	51
female mean	196	25	165	68	182	51			181	48
% voiced	11%		29%		22%				21%	

It may be seen that in the last part of the articulatory closure there is substantial voicing, often amounting to almost a quarter of the duration of the segment.

We also investigated another language that has voiceless nasals similar to those in Burmese. Mizo, also known as Lushai (e.g. by Bright 1957, Burling 1957, Weidert 1975), is a Tibeto-Burman language spoken in the state of Mizoram in North Eastern India. It has contrasts as shown in Table 3.

Table 3. Mizo contrasts involving voiced and voiceless nasals.

mai^{1}	'pumpkin'	m̥ei^{1}	'face'
na^{3}	'pain'	n̥e^{3}	'leaf'
ŋei^{3}	'to be fed up with'	ŋ̊ei^{3}	'to fast'

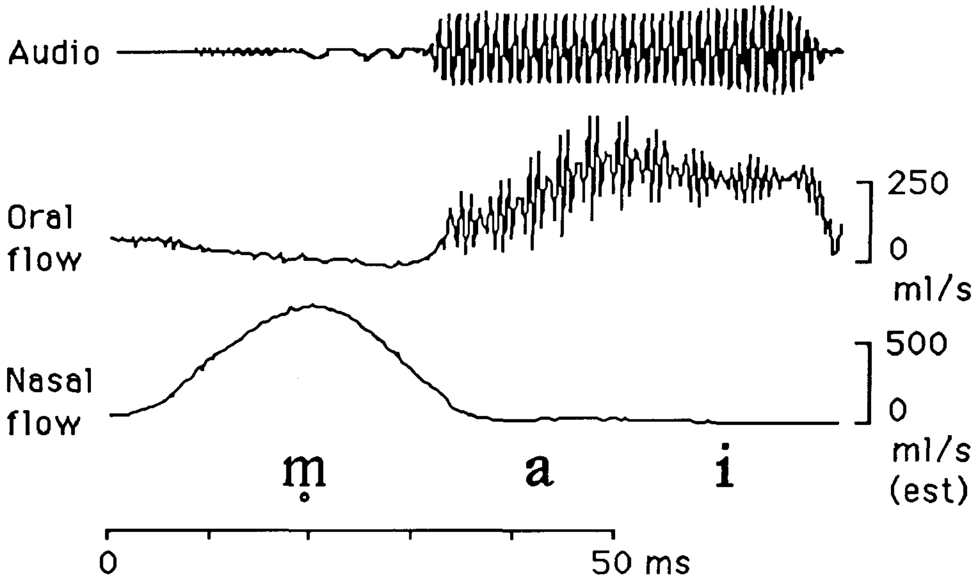


Figure 3. Aerodynamic records of the Mizo word [m̥ai] 'face'.

We recorded the speech of three female speakers of the Hmar dialect of Mizo. For the first of the three speakers we recorded the oral and nasal airflow during the pronunciation of these words when spoken in the sentence [dɔŋgin ___ tiʔasawi] "Dawngin said ___". For the other two speakers the frame sentence was [sawi rɔʔ ___] "Say ___". The aerodynamic records are similar to those for Burmese, except that, for these three speakers, the rise and fall of nasal airflow was less rapid, as shown in Figure 3.

We had no difficulty segmenting the parts of the voiceless nasal for these three female speakers, as they did not use a breathy voiced articulation. The time relations during the last part of the voiceless nasal can be seen clearly in Figure 4, which shows another token of the same word as in Figure 3, recorded without the airflow mask, and thus with a more accurate representation of the sound wave. There are five vibrations of the vocal cords during the characteristic nasal waveform, before the more complicated waveform for the vowel begins. This pattern was observed for all three voiceless nasals in every utterance by all three of Mizo speakers. Because of the poor choice of frame sentence for the first speaker, it was not always possible to measure the length of the nasal articulation. But, considering all three speakers, the mean voiced period at the end of the voiceless nasal was 44 ms which is similar to that found in Burmese.

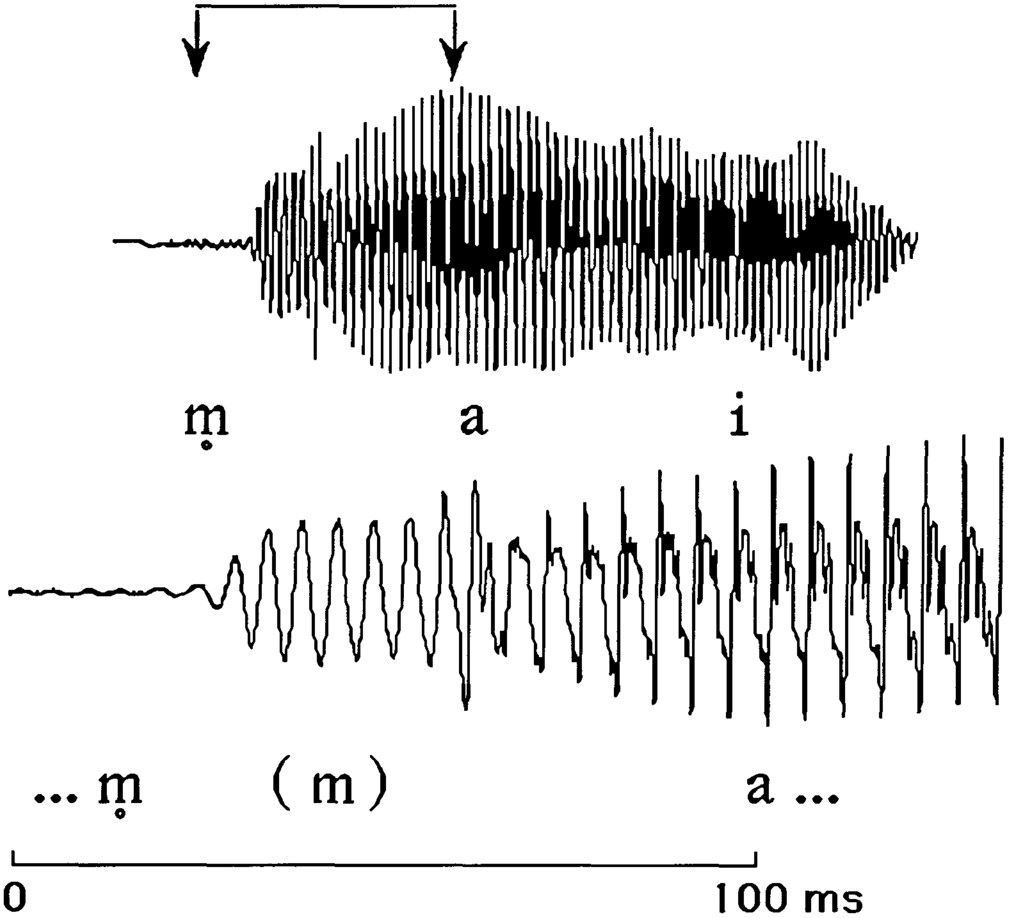


Figure 4. The waveform of another token of the Mizo word [m̥ai]. The upper part of the figure shows the whole word, and the lower part shows an expanded version of the portion of the waveform marked by arrows in the upper part of the diagram.

It might seem, then, that these short voiced portions do not constitute additional full segments, and could be regarded as simply part of a transition universally required by voiceless nasals. But this is not the case. We will now consider a very different kind of voiceless nasal used in Angami, another Tibeto-Burman language, spoken in the State of Nagaland, in North Eastern India. Examples of the contrasts among Angami voiced and voiceless nasals are given in Table 4.

Table 4. Angami contrasts involving voiced and voiceless nasals, based on Chase (forthcoming). Tones are marked as: [a^1] high, [a^2] mid high, [a^3] mid low, and [a^4] low.

me^1	'mouth'	m̥e^4	'to blow'
ne^1	'to push'	n̥e^4	'to blow one's nose'
ɲie^3	'thousand'	ɲ̥ie^4	'to paste'

We have recorded a total of 9 speakers of this language. Most of our work was concerned with the Khonoma dialect, which is spoken by about 4,000 people; six of our speakers used this dialect, and only three were first language speakers of standard Angami. The Khonoma dialect is distinct from standard Angami in many respects, but it uses the same articulatory mechanism for the voiceless nasals. In both forms of Angami there is no voiced portion towards the end of the voiceless nasal consonant. Instead, before the voicing for the vowel begins, the oral occlusion is released while air is still flowing out through the nose. The auditory impression is that there is an epenthetic voiceless plosive after the voiceless nasal and before the vowel. But this is an incorrect description as a plosive involves a complete stoppage of the air, which is then released orally. In Angami voiceless nasals there is continuous nasal airflow, which even persists into the following vowel.

The structure of these unusual voiceless nasals may be seen from the aerodynamic records in Figure 5, which shows examples of each of the three voiceless nasals extracted from the frame sentence. Significant moments in time are marked with arrows in the top example, At time (1) the articulators (in this case, the lips) close, and after a few vibrations of the vocal cords voicing ceases. The line indicating the oral airflow slopes slightly upwards from this point, probably because the lips are being pushed forwards into the mouthpiece. At time (2) the articulators open and there is a rapid flow of air from the mouth. At the same time 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. At time (3) voicing starts, probably with somewhat breathy vibrations, as there is a high rate of airflow through the mouth, as well as through the nose. If we take it that the vowel begins at this point, then we must consider at least the first part of it to be nasalized.

A similar sequence of events may be seen in the records for the other two voiceless nasals in this language. The oral airflow on the release of the alveolar closure (at the equivalent of time (2) in the middle set of records) is particularly strong. It even causes some artifacts on the audio record which was made via a microphone held just outside the oral mask. The nasal airflow drops at this moment in time, but it still remains at about 500 ml/s. The voiceless palatal nasal at the bottom of the figure shows a far less sharp release of the oral air. The nasal airflow also drops more slowly, and we may conclude that there is a voiceless palatal fricative accompanying this voiceless nasal.

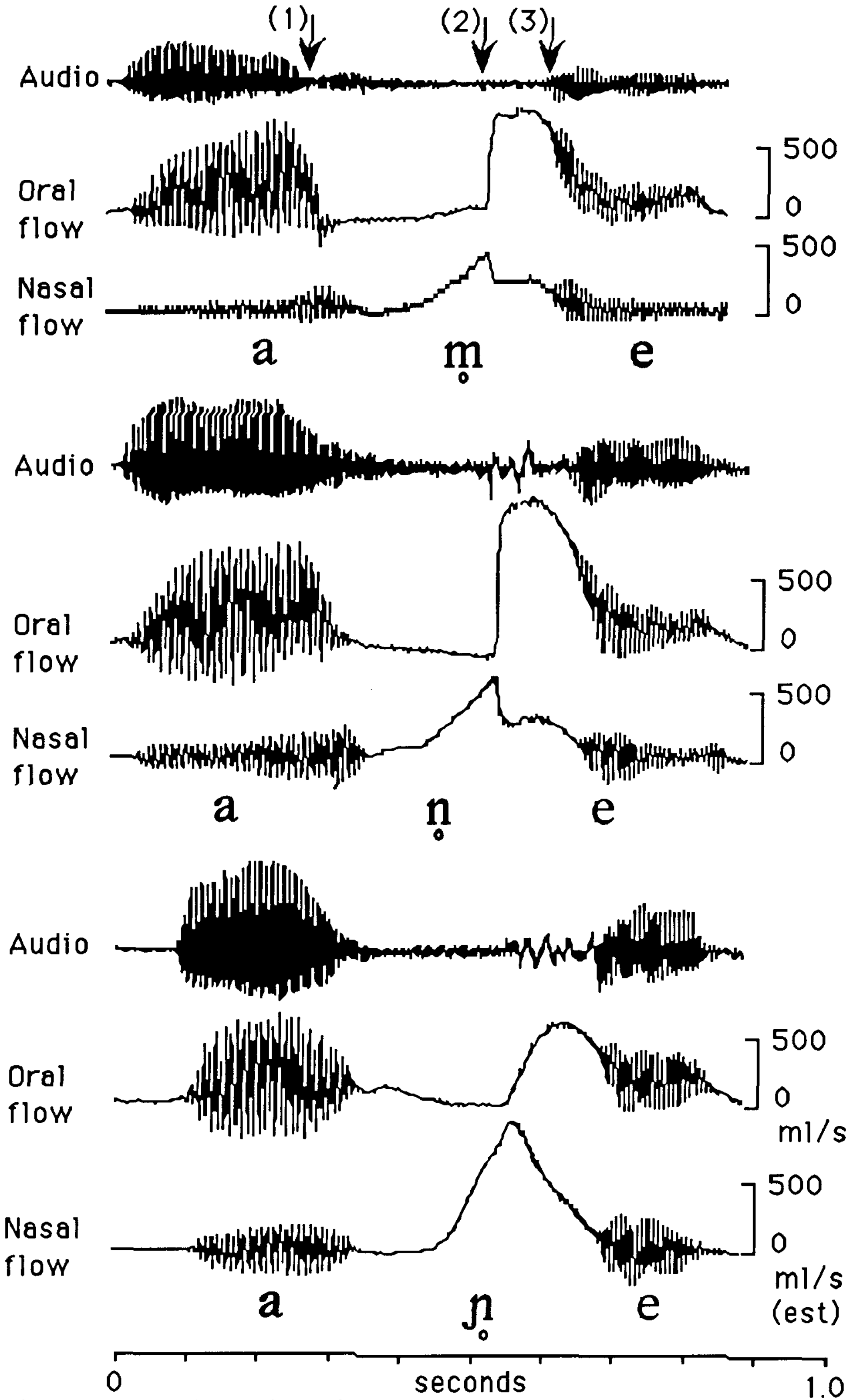


Figure 5. Angami aerodynamic records. See text for explanation.

These patterns were consistent across all repetitions for all of the nine speakers of Angami that we have recorded. Oral airflow began a little over half way through the voiceless section. Unlike the Burmese and Mizo 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. There were notable differences among the Angami voiceless nasals. The oral airflow was usually greatest during the alveolar nasal, and almost as great during the bilabial nasal. The palatal nasal usually had both a slower increase of nasal airflow, and a slower decrease.

The durations of the different parts of the voiceless nasals for the six speakers of the Khonoma dialect of Angami are shown in Table 5. All these measurements were made on the aerodynamic records. As in the case of the Burmese measurements, the nasal consonant was considered to begin when the oral closure occurred, although there were often a few periods of voicing extending into this portion of the sound. There is a certain amount of variation in all these durations due to overall differences in speed among the speakers, the different degrees of emphasis which the speakers placed on the word, and their degree of accomodation to the experimental task. Nevertheless it is clear that in by far the majority of cases a considerable portion of the consonant has oral as well as nasal voiceless airflow.

Table 5. Durations in ms of the parts of the voiceless nasals for six speakers of the Khonoma dialect of Angami. Several sets of data were obtained for some of these speakers.

	bilabial		**dental**		**palatal**		**mean**	
Speaker	**nasal**	**oral**	**nasal**	**oral**	**nasal**	**oral**	**nasal**	**oral**
1 (a)	150	162	91	117	97	230	113	170
(b)	96	130	159	101	141	182	132	138
(c)	137	121	189	137	146	177	157	145
2 (a)	135	119	105	207	155	139	132	155
(b)	110	93	143	79	87	131	113	101
(c)	105	145	175	124	115	158	132	142
3 (a)	206	45	260	85	213	185	226	105
(b)	253	91	302	95	282	88	279	91
4	248	89	227	107	221	126	232	107
5	73	44	184	22	139	26	132	31
6	146	57	182	49	116	102	318	69
mean	151	100	183	102	156	140	179	114
% voiceless oral		40%		36%		47%		39%

These voiceless nasals pose some interesting problems in relation to the phonetic interpretation of phonological units. Unlike their counterparts in Burmese and Mizo, which are essentially voiceless nasals followed by a small portion of voiced nasal, each of these sounds can be described as a voiceless nasal for only part of its duration. The other part consists of a sound that is best regarded as a weak oral release followed by a nasalized voiceless semivowel. There are thus two distinct types of voiceless nasal: (1) in Burmese and Mizo; (2) in Angami. Both types have a similar beginning, but they differ in their endings. There could also be languages with voiceless nasals that differ from these in their beginning, being voiced for the first half and voiceless thereafter. In any case, it is evident that the phonetic interpretation of voiceless nasals cannot be made in terms of universal rules that apply in all languages. These sounds provide one more demonstration that we cannot describe the phonetic substance of the languages of the world in terms of a universal set of phonetic interpretation rules.

References

Bright, W. (1957). Alternations in Lushai., *Indian Linguistics* 18, 101-110.

Burling, R. (1957). Lushai Phonemics, *Indian Linguistics* 17, 148-175.

Chase, N. (forthcoming) *A descriptive analysis of the Khwünomia dialect of Angami.* Ph.D., University of Poona.

Henderson, E. (1948). Notes on the syllable structure of Lushai. *BSOAS*, 12, 713-725.

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

Ladefoged, P. (1982). *A Course in Phonetics* (Second edition ed.). New York: Harcourt, Brace, Jovanovich.

Ladefoged, P. (1991). Instrumental phonetic fieldwork: techniques and results. In *XII International Congress of Phonetic Sciences*, 4 (pp. 126-129). Aix-en-Provence: Université de Provence.

Ravindran, N. (1974). *Angami phonetic reader.* Mysore: Central Institute of Indian Languages.

Weidert, A. (1975). *Componential analysis of Lushai phonology.* Amsterdam: John Benjamins.

Another View of Endangered Languages

Peter Ladefoged

[To be submitted to *Language* as a commentary on a paper by Ken Hale, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, Laverne Masayesva Jeanne, & Nora C. England on endangered languages. Lg. 68. 1-42.]

Language seldom publishes opinion pieces, such as that of Hale, Krauss, Watahomigie, Yamamoto, Craig, Jeanne, & England (1992), on endangered languages. I have nothing but praise for the work that these linguists do. But language preservation and maintenance is a multi-faceted topic on which different opinions are possible. The views expressed in these papers are contrary to those held by many responsible linguists, and would not be appropriate in some of the African countries in which I have worked in the last few years. Tanzania, for example, is a country which is striving for unity, and the spreading of Swahili is regarded as a major force in this endeavor. Tribalism is seen as a threat to the development of the nation, and it would not be acting responsibly to do anything which might seem, at least superfically, to aid in its preservation.

Hale et al (1992) write from the perspective of linguists who have worked in particular cultures; but the attitudes of the speakers of the languages that they describe are far from universal. As they indicate, in many communities the language is regarded as sacred — literally God given. Linguists working in such communities should obviously respect the opinions of the speakers, and honor their wishes. The speakers are giving access to something that is sacred to them, and it should be treasured accordingly. But not everyone holds this view. The half a dozen speakers of Angami (Tibeto-Burman) with whom I worked earlier this year had a different attitude. They regarded it as an intellectually valid pursuit for me to take an interest in their language. Admittedly, they were all high school or college educated students, who had a similar intellectual interest in my language. They might therefore be regarded as part of an elite, with views that were only those of the elite. But I do not think this is so. The profane, as opposed to sacred, view of language is widely shared, even among those who are certainly not part of the socio-economically elite. Many of the people with whom I have worked in undeveloped parts of India and Africa regard being a language consultant as just another job, and a reasonably high status one at that. They have no problem with satisfying my intellectual curiosity. They in no way regard their work as prostituting something that is holy. Instead they are pleased with the honored status of being teachers. Furthermore, it pays better than alternative occupations, such as picking tea or digging yams, and it is much less hard work.

Even among those for whom language is a vital part of the sacred way of life, the attitude towards linguists is not always that outlined in Hale et al (1992). The Toda, speakers of a Dravidian language in the Nilgiri Hills of Southern India, have a series of songs which are an important part of their religious life (Emeneau, 1984). They eagerly welcome linguists who wish to assist them in recording their language. They also realize that with less than 1,000 speakers they are unlikely to remain a distinct entity. Many of the younger people want to honor their ancestors, but also to be part of a modern India. They have accepted that, in their view, the cost of doing this is giving up the use of their language in their daily life. Surely, this is a view to which they are entitled, and it would not be the action of a responsible linguist to persuade them to do otherwise. In the circumstances of my fieldwork it would also have been somewhat hypocritical. I was native language in their own home, so that their child could be brought up as a native language speaker of English. This choice, and any choices that the Toda might make, are clearly their prerogatives.

So now let me challenge directly the assumption of these papers that different languages, and even different cultures, always ought to be preserved. It is paternalistic of linguists to assume that they know what is best for the community. One can be a responsible linguist and yet regard the loss of a particular language, or even a whole group of languages, as far from a 'catastrophic destruction' (Hale et al, 1992:7). Statements such as 'just as the extinction of any animal species diminishes our world, so does the extinction of any language' (Hale et al, 1992:8) are appeals to our emotions, not to our reason. The case for studying endangered languages is very strong on linguistic grounds. It is often enormously strong on humanitarian grounds as well. But it would be self-serving of linguists to pretend that this is always the case. We must be wary of arguments based on political considerations. Of course I am no more in favor of genocide or repression of minorities than I am of people dying of tuberculosis or starving through ignorance. We should always be sensitive to the concerns of the people whose language we are studying. But we should not assume that we know what is best for them.

We may also note that human societies are not like animal species. The world is remarkably resilient in the preservation of diversity; different cultures are always dying while new ones arise. They may not be based on ethnicity or language, but the differences remain. Societies will always produce sub-groups as varied as computer nerds, valley girls and drug pushers, who think and behave in different ways. In the popular view the world is becoming more homogeneous, but that may be because we are not seeing the new differences that are arising. Consider two groups of Bushmen, the Zhuǀoãsi and the !Xóõ, who speak mutually unintelligible languages belonging to different sub-groups of the Khoisan family, but otherwise behave in very similar ways. Are these two groups more culturally diverse than Apalachian coalminers, Iowa farmers and Beverly Hills lawyers? As a linguist, I am of course saddened by the vast amount of linguistic and cultural knowledge that is disappearing, and I am delighted that NSF has sponsored our UCLA research, in which we try to record for posterity the phonetic structures of some of the languages that will not be around much longer. But it is not for me to assess the virtues of programs for language preservation versus those of competitive programs for tuberculosis eradication, which may also need government funds.

In this changing world, the task of the linguist is to lay out the facts concerning a given linguistic situation. The approach that I would advocate is exemplified in our study of language use and teaching in Uganda (Ladefoged, Glick and Criper 1971). With the full cooperation of the then (more or less) duly elected government (this was immediately before the time of Idi Amin), we assembled data on the linguistic situation. We tried to determine the linguistic similarities and mutual intelligibilty of some of the major languages spoken in Uganda (about 16 Bantu, 5 Western Nilotic, 4 Eastern Nilotic, 2 Central Sudanic and 4 non-Ugandan). We found that about 39% of the people could hold a conversation in Luganda (the largest single language), 35% in Swahili, and 21% in English. We noted that Radio Uganda put out programs in 16 Ugandan languages (some of them mutually intelligible), plus Swahili and English, and that there were literacy campaigns in 20 languages. Six Ugandan languages were used in schools. We summarized all our data so that the government could assess the linguistic situation. We did not try to determine the costs of making changes or of maintaining the *status quo*, in either monetary or human terms. It would have been presumptuous of us to weigh the loss of a language against the burdens facing Uganda. We tried to behave like responsible linguists with professional detachment.

Last summer I was working on Dahalo, a rapidly dying Cushitic language, spoken by a few hundred people in a rural district of Kenya. I asked one of our consultants whether his teen-aged sons spoke Dahalo. 'No,' he said. 'They can still hear it, but they cannot speak it. They speak only Swahili.' He was smiling when he said it, and did not seem to regret it. He was proud that his sons had been to school, and knew things that he did not. Who am I to say that he was wrong?

References

Emeneau, Murray. B. 1984. Toda Grammar and Texts. Philadelphia: American Philosophical Society.

Hale, Ken, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, Laverne Masayesva Jeanne, & Nora C. England. 1992. Endangered languages. Lg. 68. 1-42.

Ladefoged, Peter, Ruth Glick, & Clive Criper. 1971. Language in Uganda. Nairobi, Kenya: Oxford University Press.

What do we need to know in order to do spoken language research?

Peter Ladefoged

[Invited paper for the International Conference on Spoken Language Processing,
Banff, Alberta, Canada, October 12-16, 1992]

ABSTRACT

Nobody can know everything about spoken language processing. Whatever aspect of the field we work in, we have to get help from experts in other fields. For example, for my work on the phonetic structures of dying languages, I rely on the talents of linguists, anthropologists, sound recording engineers, physiologists, psychologists and signal processing engineers to support my own phonetic knowledge. We are fortunate in that by coming to a congress such as this we can all profit and get a little help from our friends.

In the days of Leonardo da Vinci, a large library contained only a few thousand volumes. Nowadays there are tens of thousands of books that deal with speech in one way or another. Nobody, not even Leonardo, could know all that there is to know about speech research. The only way in which one can work in this field is to have good friends. We all have to rely on other people to fill in the gaps — the vast holes — in our knowledge. Any scientist today is part of a team that cannot hope to build a bridge into the future without a lot of help. This is clearly so in my case.

Much of my life is spent trying to describe how the sounds of one language differ from another. There is also some lofty goal, such as trying to develop a theory of the nature of spoken language; but that is usually in the back of my mind. In the forefront is the everyday business of describing some particular piece of spoken language. I would guess that the majority of us at this Congress could say the same thing: we spend most of our working day trying to describe some particular piece of spoken language. I concentrate on linguistic contrasts — meaningful differences — between one sound and another. One way or another, this is still true of most of us, the major exception being those who study pathologically deviant forms of speech. The rest of us are dealing with the meaningful elements of speech. In the case of some scientists, for instance communication engineers, the intent may be not to describe the meaningful elements but to process them, or to change them into other forms and transmit them. But a transformation involves a tacit description of what is to be transformed and what is not. So, in some sense we are nearly all doing the same thing, albeit in ways very different from one another.

Of course, the different aspects of spoken language that we choose to describe lead to our tasks — our daily lives — being very different from one another. But my general theme, that we have to rely on a great deal of help from many different people, is still true. My current project is the description of the sounds of dying languages. It has been estimated that about half the world's 6,000 languages will become extinct within the coming century [2]. This is a vast amount of cultural knowledge that is disappearing, and I am delighted that NSF has sponsored our UCLA research, in which we try to record for posterity the phonetic structures of some of the languages that will not be around much longer. Our first step in doing this is to identify suitable languages for investigation. This involves the cooperation of other linguists, local representatives of speakers of these languages, government officials, missionaries, and anybody who can give us advice on the viability of a particular language. We try to select languages that are phonetically interesting, for which we rely on our own expertise; but we could not do without the help of other linguists who know the particular languages. It is not practicable to go out and record the phonetic structures of a language without the assistance of someone who knows the phonology, and can give us access to a large lexicon so that we can select suitable contrastive forms.

Even more important than the help we receive from knowledgeable linguists is the help that we must have from the speakers of the language. This is somewhat outside the theme of this paper, in which I am trying to show how much we rely on other scientists in our own and neighboring fields; but it is obviously worth noting that one cannot do work on any language — dying or not — without the cooperation of its speakers. We have found that this is usually easy to obtain; most people are interested in their own language and are only too willing to share its mysteries with others. On other occasions, however, particularly when dealing with some Native American Indians, it may be more grudgingly given. For some people, language is sacred, literally god-given, and not to be casually shared with outsiders.

Having decided which language to record, our next task is to go out and do it. This, too, involves knowledge gained from other fields. Any fieldwork demands some of the skills of the anthropologist, in that one must know how to make observations within the local culture. To begin with, we must know how to choose suitable speakers who are truly representative of the population. One of the problems that I face is that local leaders expect me to record the old men and women who know the wisdom of the tribe. But these people often have weak and quavery voices; and they may not have any teeth. So, although they may be valuable in showing how people used to speak, and reliable in their control of the syntax and vocabulary of an older generation, they are not so useful in providing formant frequencies indicative of the vowels of current speech.

We must also make sure that the people we record are speaking in a normal way; we must be able to observe the culture without disturbing it. I regret that I often have to give up on this. The exigencies of my work are such that it is impossible to spend enough time to be able to record all the phonetic structures of a language spoken in a completely natural way. I would like, for example, to be able to record all the vowels of a language in normal conversational speech. But it is just not possible to wait for each of 10 speakers to say words containing similar consonants but each of 10 different vowels, all within the context of a friendly chat. So I have to structure my observations in some way. Any advice on this topic is welcome.

These are not problems that apply to fieldwork phoneticians alone. Virtually all of us work with recorded data provided by a sample of speakers. One of the most cited early works in acoustic phonetics [8], is flawed because the authors did not pay proper attention to choosing suitable speakers. Their data do not provide a reliable account of the vowels of what they call General American English. Their speakers consisted largely of people around Bell Telephone Laboratories, New Jersey, some of whom were not even born in the United States. They cannot be considered to be a sample which is representative of any particular population.

Nor are many more recent studies without fault. The TIMIT database, sponsored by the National Institute of Standards and Technology, made a more serious attempt to control the dialects of the speakers, both by reporting the speakers' own assessments of their dialect, and by building in so-called calibration sentences that allow investigators to observe enough features to be able to place each speaker into a dialect category regardless of the speaker's own judgment. But the TIMIT database is sometimes not so successful with regard to the second point mentioned above, namely, observing the culture without disturbing it. Many of their speakers were not very good readers. As I noted, I do not know how best to record speakers talking in a natural way. I see that there are several papers on databases at this congress, and I look forward to learning something from them. And I hope that the anthropologists and sociologists amongst us will keep us all honest in these fields.

Recording a spoken language, whether in a laboratory or the outback of Australia, also demands some instrumental skills. Nowadays a good phonetician will use a DAT (digital) recorder, so as to make recordings of the highest possible quality. Regrettably many of us regard this as a comparatively simple procedure, requiring no particular skill–which is why so many bad recordings get made. We should be consulting our engineering colleagues to ensure that we are indeed using the highest quality, noise canceling, directional microphones, and the best available recording system.

We should also consult with our colleagues in physiology. We need to know about current work in speech production for many reasons, ranging from the straightforward practical help that we can get by using instrumental aids that tell us what the speaker is doing, to a deeper understanding of the nature of the units of speech that we are trying to record. My own phonetic fieldwork now relies extensively on physiological data. I cannot go out into the field with the more elaborate techniques for studying speech production described in some of the papers at this congress. There is no portable x-ray microbeam we can take into the Kalahari desert; and even techniques for studying articulatory movement such as electromyography and magnetometry are difficult to use in circumstances that require lightweight apparatus that can be dropped and bounced around in transit, and then operated without an electrical supply from a public utility. One of the new methods of studying tongue movements, electropalatography, has been successfully used in the field by Butcher (personal communication), but this technique has its limitations in that it requires the production of a special artificial palate for each speaker. Each palate costs $500 and takes about 4 weeks to produce. It would usually be too expensive and too time consuming to set up, if one is trying to record half a dozen representatives of a language in a single field trip.

Nevertheless, phoneticians are missing many opportunities if they think of fieldwork as simply involving tape recording. One can learn a lot about different places of articulation from static palatography, a nineteenth century technique in which one of the articulators is coated with a marking medium, a word is said, and then one can observe where the articulators have made contact. A video camera, which is our current technique for recording palatographic observations, will also provide useful data on labial articulations, as has been shown by my colleague Ian Maddieson [6]. Aerodynamic data is another staple of contemporary instrumental phonetic fieldwork. Butcher (personal communication) has produced some interesting studies of Australian aboriginal languages. Our recent aerodynamic studies of Sandawe, an East African click language, are described in Maddieson, Ladefoged and Sands (in press), and of Angami, a Tibeto-Burman language spoken in India, in Ladefoged and Bhaskararao [4]. In all these and many other cases, the ability to record and analyze physiological data has made a valuable contribution to the description of the sounds of the language.

I suspect that, whatever our corner of the field, similar considerations apply. Whatever kind of spoken language data one is examining, one's knowledge of why it is as it is would probably be improved if physiological data were also available. Everyone who is concerned with spoken language processing should at least be aware of what our colleagues in speech production are doing. That is why we have congresses like this, so that we can all get a little help from our friends.

Nor should any of us neglect the work of our psychologist colleagues who study the perception of speech. Linguistic phonetic fieldwork has not usually involved techniques of this kind. But there has been some notable work, such as that of Traill [9], who took audiometry equipment out into the field, tested the hearing of a group of Bushmen, and then reported the results of listening tests involving clicks synthesized by different rules on a Klatt synthesizer. There is probably not much demand for speech synthesis by rule in a hunter-gatherer economy, and certainly little need for automatic speech recognition in a society that does not use ATMs, or, for that matter, banks. But when the time comes, Traill's work will be there to help; and, more importantly, it has already provided us with useful information on auditory distinctions among clicks. We now know a little more about the perceptual phenomena that occur in the world's languages.

Other, less involved, psycho-acoustic experiments can be readily done in the field. We have a very portable system that permits subjects to find a match to a particular stimulus such as a vowel in their own language, using the protocol described by Johnson, Wright and Flemming [1]. The system can be run on any current Macintosh computer, including a power book, the only additional equipment needed being a pair of headphones. The subject's task is to use the mouse to find the best match out of 330 high quality synthesized vowels, each of which can be reproduced by clicking on one of 330 buttons arranged in a 15 x 22 matrix, corresponding to F1 and F2 values. This system has a much wider applicability than its use in our linguistic studies. It provides a way of describing a wide variety of vowels (not including nasalized vowels, rhotacized vowels, and other vowels with special considerations involved) in terms of a set of standard vowels. Subjects are usually in fairly good agreement on what constitutes the best match to a given vowel. It is possible to regard the system as an alternative to a phonetician's descriptions in terms of cardinal vowels. Using this system, phoneticians have a meaningful way of communicating with one another when they say, for example, that the mean match to the Mexican Spanish vowel [e] as in "mesa" is vowel #67, with certain formant frequencies. Perhaps those working in other aspects of spoken language processing, such as coding and ASR, might also use this same reference system, again making it possible for one part of the field to benefit from the work of another group.

Finally, while discussing perceptual psychology, I would like to appeal to this Congress and admit to a certain sense of frustration that I get when consulting some auditory psychologists concerning my practical needs. I want to know, when I describe the vowels of a language by plotting the formant frequencies, is it more appropriate to plot the formants on a mel scale, a bark scale, and ERB scale, or any other scale? None of the auditory experts seem able to agree. Some of them are distinctly unhelpful, by suggesting that I should not represent vowels in terms of formant frequencies at all; but they do not go on to say how I should represent them so that I can most usefully compare the vowels of one language with another. I suppose I will just have to go on with the ad hoc devices and intuitions I now use [3]. But if there is some general agreement on a better system, I would like to know about it.

Of course, the area of acoustic analysis is where most of us here overlap. There are many papers in this area that I as a linguistic phonetician have to take into account. Like most of us, I need to know about the latest analysis systems that our engineering colleagues are producing. Although we now routinely perform some analyses while we are out in the field and still have access to speakers, the bulk of the analysis of the fieldwork data is done on laboratory instrumentation when we return. We need to know the best ways of extracting descriptive parameters from our recordings. We also need some understanding of the engineering concepts involved in digital signal processing. We need to know why, for example, a 12th order LPC is appropriate when determining formants in data sampled at 10,000 Hz, and what is the best window to use when trying to measure pitch.

Sound spectrograms of one kind or another still provide a useful way of representing speech data in a visual form. Many of us from all different parts of the field are interested in what our colleague are saying about labeling spectrograms. They may be doing this for the purpose of creating large databases, but the problems they face are the same as we all have to consider when trying to interpret the facts about some particular piece of spoken language.

Moreover, databases themselves are becoming of more and more interest to many of us. We in phonetics have been using some kinds of databases for many years. Maddieson [4,7] showed how much we could learn by studying the segmental inventories of a carefully chosen sample of languages. More recently phoneticians have started making analyses of speech databases; two of my own colleagues, Patricia Keating and Dani Byrd, are among the many reporting at this Congress.

I believe that the use of large databases may completely transform phonetics, phonology, and perhaps even the whole of linguistics over the next few years. Since the advent of Noam Chomsky, the emphasis in linguistics has been on describing a speaker's competence — the mental structure of language — rather than the actual performance. Linguists, even phoneticians, have tended to ask "Could one say so and so," rather than to observe what percentage of people actually say the equivalent of "so and so." I am not meaning to imply that none of the recent advances in linguistics have been data driven. Many linguists are brilliant observers of what people do, and they have elaborated their theories so as to account for their observations. Nor am I meaning to imply that the mentalist approach to linguistic description is incorrect. Of course language can be usefully described as a set of rules in a speaker's mind. But that is not the only valid, nor even the only interesting, description of the social phenomena we call spoken language. Now that we are beginning to build up large databases, we can take a different approach. It would be foolish to go back to the American linguistics of the early 1950's, when descriptions of a language were supposed to encompass all and only that which was in a corpus. But it is equally foolish to continue sitting in an armchair and pontificating about the spoken language inside some imaginary speaker's head. I know that in saying these things at this Congress I am in some senses preaching to the converted. We are a fairly data driven lot. But let us make the breadth and excellence of our descriptions of spoken language so great that none of our more theory-bound colleagues will be able to disregard us.

I started off this paper by asking what do we need to know in order to work in spoken language processing, and I hope fairly rapidly demonstrated that there was no way any of us could know it all. Let me end on a more upbeat note. Suppose we ask instead, what does one need to know in order to do *good* work in spoken language processing. I have asked a number of well known people who are coming to this meeting whether there were any out of a list of topics connected with spoken language processing that they knew nothing about. Virtually everyone who answered admitted that they knew almost nothing about at least one of the listed topics of this congress. You would probably be surprised at the confessed ignorance of some of the major figures in the field. So obviously it is not necessary to know all about spoken language. You can do good work in the field knowing only your own little corner. But it is also true that the leading figures in the field do have at least some knowledge of many different parts of it. So here at this Congress is our opportunity to fill some of the gaps. Of course we will have the usual problems. In my case I hope I will be able to hear about new things such as magnetometer sensing systems, and new descriptions of tongue movements. But it seems that I will have to listen with only one ear to each; and my two ears will have to be in separate rooms. As there are many other cases like this, I expect I will often be listening with only half an ear. There is a wealth of material awaiting us.

References

[1] J. Johnson, R. Wright and E. Flemming. "Using the method of adjustment to study vowel spaces." *Journal of the Acoustical Society of America,* vol. 91, p. 2387, 1992.

[2] K. Hale, K. Michael , L. Watahomigie, A. Y. Yamamoto, C. Craig, L. M. Jeanne, and N. C. England. "Endangered languages." *Language*, pp. 68. 1-42, 1992.

[3] P. Ladefoged. *A Course in Phonetics* (Third edition ed.). New York: Harcourt, Brace, Jovanovich, 1992.

[4] P. Ladefoged and P. Bhaskararao. "Phonetic interpretation of voiceless nasals." this volume.

[5] I. Maddieson. *Patterns of Sounds*. Cambridge: Cambridge University Press 1984.

[6] I. Maddieson. "Revision of the IPA: Linguo-labials as a test case." *Journal of the International Phonetic Association,* vol. 17, pp. 26-30, 1987.

[7] I. Maddieson and K. Precoda. "Updating UPSID." *UCLA Working Papers in Phonetics*, vol. 74, pp. 104-111, 1990.

[8] G. E. Peterson and H. Barney. "Control methods used in a study of the vowels." *Journal of the Acoustical Society of America,* 24, 175-184,1952.

[9] A. Traill. "The perception of clicks in !Xóo." In D. Dwyer (Ed.), *Proceedings of the 23rd. African Languages Conference,* 1992.

Facilities for speech perception research at the UCLA phonetics lab

Keith Johnson & Henry Teheranizadeh

This document describes hardware and software facilities for on-line speech perception research which have recently been developed at the UCLA phonetics lab and is intended primarily for our future reference. However, we also hope that this description of some of the strategies we have used to develop a first rate facility at minimal cost will be useful to researchers at other institutions. Also, as with other software developed in the phonetics lab, the library of C-callable routines described here is considered to be in the public domain.

The first section describes the hardware configuration. The second section is an overview of the routines in PHONLAB (a library of C-callable routines for the IBM PC). The third section describes each routine in detail. The fourth section describes utility programs for preparing stimuli for on-line presentation.

1 Hardware setup

Figure 1 shows the hardware arrangement that we are currently using. One or two listeners are seated at stations in a sound booth. Each listening station is equiped with: (1) headphones (Sony MDR-V4), (2) a video display (Goldstar, Data Display Monitor), and (3) a response box. An IBM PC-AT clone is located in an adjacent room and is equiped with an A/D D/A board (Data Translation, DT2801A), analog audio hardware, and two video display cards (a Color Graphics Adapter and an Enhanced Graphics Adapter).

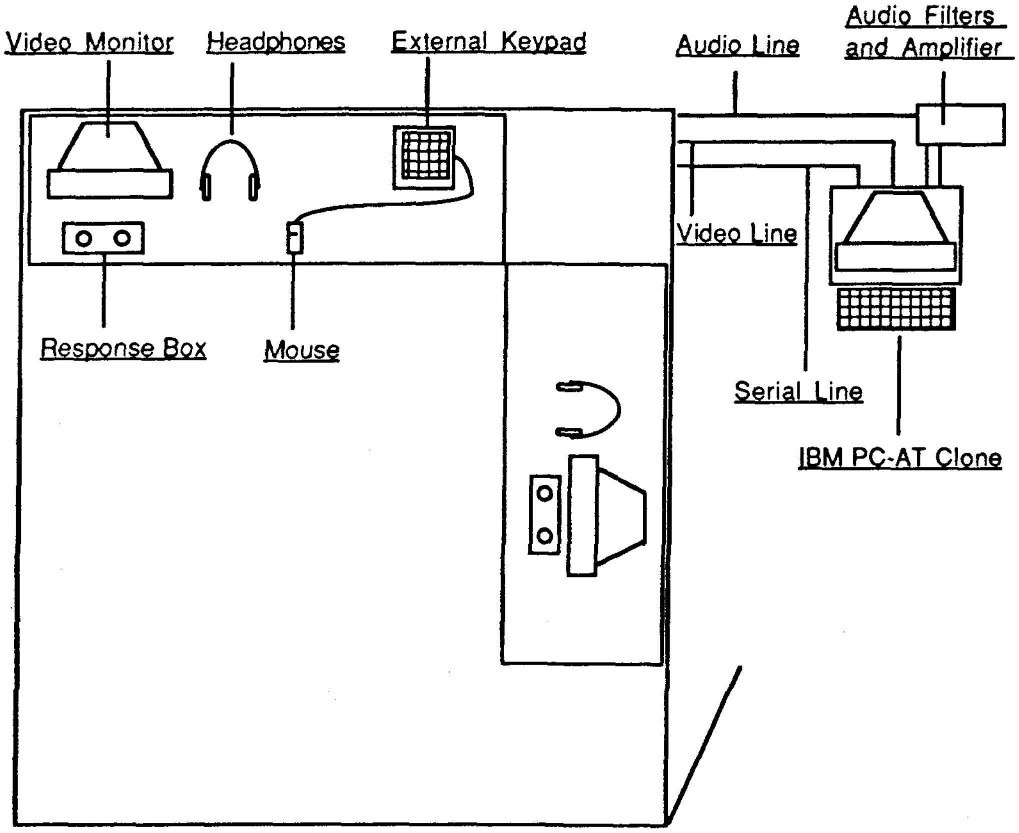


Figure 1. Schematic diagram of the speech perception hardware setup at the UCLA Phonetics Lab.

The analog signal to the listeners is provided by a digital to analog card (DT2801A) in the PC. The analog rack, adjacent to the PC, is configured for stereo presentation of auditory stimuli, although this capability is not as yet supported by the software. The signal is filtered at either 4.2 kHz or 8.4 kHz (Frequency Devices, 8 pole Butterworth active low pass filters) depending on the sampling rate of the digitized stimulus files. The signal is amplified (BGW Broadcast Power Amplifier, Model 85) and sent to the listeners' headphones and (in parallel) to a VU meter located on the front of the analog rack. The impedence of the amplifier signal is modified to match the combined impedence of the listeners' headphones.

The video signal to the listeners is provided by the composite video output of a standard Color Graphics Adapter (CGA). We use a CGA graphics card because it can be easily programmed for presentation of graphics (such as response "buttons" in method-of-adjustment experiments), and is also an inexpensive and effective method for driving more than one video display at a time. An Enhanced Graphics Adapter (EGA) serves as the primary display adapter for the PC and can be programmed to display ongoing reports to the experimenter during an experiment run.

The listeners' response boxes were fabricated in house. We used an Ortek extended keypad (supplied by Jameco Electronics) as the interface between the response boxes and the computer. The buttons on the response boxes (obtained from a pinball machine manufacturer) are wired to keys in the extended keypad. The keypad connects to a serial port in the computer, and software supplied with the keypad loads a memory-resident program which services hardware interrupts from the serial port (key presses). In addition to collecting listener responses via the response boxes, we also use a mouse (Mouse Systems PCMouse, or any MicroSoft compatible mouse) in some experiments. The mouse also connects to a serial port in the PC. We use a serial line T-switch (from Inmac) to choose between the external keypad and the mouse. There is also a T-switch near the PC to choose between using a mouse locally at the computer or to connect the sound booth to the serial port.

We use an IBM PC-AT for the on-line control of perception experiments. In addition to the special hardware described earlier, this machine has a 6MHz clock rate, a 40 Megabyte hard drive, math co-processor, 640 Kbytes of RAM, and a memory expansion card. The cost of the system from the ground up is less than $5000.

2 Software Overview

No one program can provide all the variations in experimental design which are likely to be needed in an active program of speech perception research, we therefore decided to develop a library of callable routines which handle most of the details involved in developing individualized experiment control programs. The routines are not as easy to use as a menu-driven experiment control system, but they give the user a much greater degree of flexibility in putting together perception experiments without having to worry about low-level details. Programs implementing several standard paradigms have also been developed, so it may not be necessary to do any programming (or only make minor revisions of an existing program) in many cases.

PHONLAB is a library of C-callable routines for the IBM PC which can be used to present auditory or visual stimuli, collect responses from several listeners at once, record response times for each response, and control the timing of various experimental events. The routines are written in C and have been compiled using the Microsoft C compiler (version 5.1). This section is an overview of the routines contained in PHONLAB. More detailed descriptions of the routines, including example calling sequences are listed in a later section.

2.1 Timing routines

A millisecond timer is implemented by manipulating the IBM PC's time of day clock.

Fast_timing() changes the time of day clock's interrupt rate and sets up a custom interrupt service routine which increments a global variable once every millisecond. This global variable (*g_msec*) is available to the user and to other routines in the PHONLAB library. *Slow_timing()* returns the computer system to normal. These routines are C translations of timing routines published by Brysbaert et al. (1989), who also tested the accuracy of the timer and discussed some considerations in the use of the system's time of day clock as a millisecond timer.

One higher-level routine is provided to time interstimulus and intertrial intervals. The routine *wait()* assumes that the millisecond timer has been started with *fast_timing()*. Upon calling *wait(n)*, program execution pauses n milliseconds.

2.2 Digital to Analog conversion

PHONLAB includes routines to play speech through a Data Translation DT2801A A/D-D/A board. The routines assume that the input speech files have CSpeech headers (Milenkovic & Read, 1990). Two different strategies are implemented with routines shown in the following table:

	Fast access/short files	Slower access/long files
Prepare a list of files for presentation	load_files()	open_files()
Play out a file from the prepared list	play()	l_play()
Release files	free_memory()	close_files()

For the fast access/short files strategy all of the files on the list are read into memory, for later D/A conversion. This means that when an experiment running program calls *play()* the file is already stored in memory and thus, no disk access is needed. The file begins to play almost immediately. However, this strategy is not feasible for experiments involving very long files, because the computer's memory may not be large enough to hold very many long files. The slow access/long files routines open the speech files and read their headers before being called upon to actually output the files, but the speech samples are not read from disk until the *l_play()* routine is called. This routine can play a file of indefinite length, but is somewhat slower to begin playing than is the *play()* routine. Note that the release routines, *free_memory()* or *close_files()* must be called at the end of any program using *load_files()* or *open_files()* or before calling *load_files()* or *open_files()* a second time in one program, otherwise the system memory or file access capabilities will be left in an unpredictable state. Also, note that the fast access routines and the slow access routines are incompatible with each other (i.e. the user must not attempt to play a file using *l_play()* if it was opened with *load_files()*).

The minimum interstimulus intervals possible on an IBM-AT 286 with a clock rate of 12MHz are shown below for the two types of play routines. Also, this table illustrates the effect of millisecond timing. The actual ISI value in an experiment running program will vary depending on the clock rate of the PC, and the number of operations during the interstimulus interval (e.g. response logging).

Minimum ISI's (in ms.)

	play	l_play
no timing	20	160
msec timing	22	180

One limitation of the routines described above has to do with the number of files which can be used in an experiment. For experiments involving more than 100 sound files it is necessary to open each file as it is needed. The routines involved are *set_play_buf()* to establish a DMA buffer for the playback routine, *open_file()* to open a particular sound file, *l_playwave()* to play out the file, *fclose()* from stdio, the compiler's standard input/output library, to close the file, and *free_play_buf()* to free the DMA buffer. We have used these routines in a method-of-adjustment study involving over 300 sound files. Of course the delay between auditory presentations is by necessity increased when each file must be located on the hard disk's directory before output can begin.

2.3 Response logging

Several routines are available for response logging when the button box hardware is used. *Get_listener_ids()* asks for identification numbers for each listener and sets up a table indicating the seating arrangement in the sound booth. This table is then used by *get_responses()* to detect responses and record button press and reaction time information for each listener. The routines *enable_timing()* and *disable_timing()* are used to turn on and off the button boxes. Call *enable_timing()* at the beginning of a response interval and *disable_timing()* at the end of the response interval.

One other use for the response box routines is that a voice activated switch may be attached to one of the external keypad keys and response times for naming experiments can be collected.

Response logging when the listener uses a mouse as the response device is not supported by any special routine in PHONLAB. We use the callable routines provided with the mouse to make the mouse cursor visible, query the state of the mouse, etc. Note also that there is no support for collecting reaction times for mouse button responses.

2.4 Visual stimulus presentation

There are two sets of video display routines. The first set assumes that the primary display adapter in the computer is a CGA card. For these routines the Microsoft graphics library has been used in producing some easy to use routines for visual stimulus presentation and video control. *Set_graphics()* sets the video mode for high resolution graphics, and *restore_graphics()* restores the display mode to the system default. The user should call *set_graphics()* at the beginning of a program which includes any of this first set of video display routines, and *restore_graphics()* at the end. *Center()* presents text in the center of the screen. *Labels()* puts button labels in the lower left and right corners of the screen. *Rt_feedback()* converts a time value to text and calls *center()* to print the value in the center of the screen. *T_scope()* puts text in the center of the screen for a certain interval of time. *T_scope()* and *center()* wait for a vertical retrace signal before presenting the text. This reduces the variance between the onset of a timed interval and the appearance of the text on the screen (see Brysbaert, et al., 1989).

The second set of routines assume that the primary display adapter is an EGA card and that a CGA card is the secondary adapter. These dual video routines also assume that the video monitors in the booth are conected to the CGA card. *Init2video()* is analogous to *set_graphics()* and should be called before any other dual video routines are used. Also *resrore2video()* should be called to return the computer to its default state before exiting the program. *Selectscreen()* is used to switch between the EGA and CGA adapters. When the global constant BOOTH is passed to *selectscreen()* all subsequent graphics commands write to the CGA and consequently affect the monitors in the booth and not the computer's consol monitor. (Note also that BOOTH should be selected before initializing the mouse because the mouse cursor display is initialized based on the currently selected video display adapter and mode.) When the global constant CONS is passed to *selectscreen()* further printing or display commands affect the experimenter's monitor but not the monitors in the booth. The display modes selected by *init2video()* are high resolution black and white graphics for the booth monitors, and 80X25 black and white text mode for the console. Consequently, dual video graphics routines in PHONLAB always select the CGA display adapter (booth screens) before putting text or graphics on the screen. *Clear_screen()* can be used to clear either the booth or console monitors. *Drawtext()* puts text on the booth screens. *Drawline()* puts lines on the booth screens. *Center2()*, *labels2()*, *rt_feedback2()*, and *t_scope2()* are dual video versions of the single video routines described above.

2.5 Randomization

A stimulus order randomization routine is included in PHONLAB. This routine, *randomize()*, sets the random number generator seed with the system time of day and then returns and array of numbers (from 0 to n-1) randomized without replacement.

3 PHONLAB routines

This section lists the user callable routines in PHONLAB in alphabetical order. To use these routines your program must include the header file <phonlab.h> and must be linked to phonlab.lib. See the program listing in the Appendix for an example.

3.1 center() - shows text at the center of the listeners' video monitors. The text may be stored in a character array or may be a text constant. The graphics routines in PHONLAB use video page swapping to set up video displays in memory before actually displaying them on screen. Center() writes the text in the center of the inactive page, waits for a vertical retrace signal from the computer (indicating that the screen is about to be refreshed) and switches the active page to show the new display.

```
short center(char *text, short active);
```

Input: char *text; a pointer to text.
short active; used to keep track of the current active video page.

Output: center() returns the current active page in video memory.

Example call: active=center("Press a button to continue",active);

3.2 center2() - analogous to center(), used in dual video configuration. Uses pages swapping, but unlike center() the current page number is kept in a global variable. Also waits for a vertical retrace signal before displaying the centered text.

```
void center2(char * text, struct vconfig config);
```

Input: char *text; a pointer to text
struct vconfig config; a video configuration structure, returned by init2video()

Output: none.

Example call: center("Press a button to continue",config);

3.3 clearscreen() - clears one of the two video adapters in the dual video configuration, and leaves the adapter selected as currently active. Use one of the constants BOOTH or CONS to select which screen will be cleared. Passing BOOTH to clearscreen() causes the CGA adapter to be cleared, and passing CONS causes the EGA adapter to be cleared.

3.4 close_files() - closes CSpeech files which were opened by open_files(). There MUST be one call to close_files() for each call to open_files(), also if you want to call open_files() more than once, close_files() must be called *before* calling open_files() a second time.

void close_files(int numfiles);

Input:	int numfiles;	the number of files opened by open_files().

Example call: close_files(numfiles);

3.5 disable_timing() - turns off the response boxes. Actually, what this routine does is restore the external keypad interrupt service routine to normal. Thus, button presses will still put an ASCII code in the system's keyboard character buffer, but the time of the response will not be noted, and the global variable g_button_count, which counts the number of timed button presses, will not be incremented. Therefore, get_responses() will not be aware of any button responses which occur after the call to disable_timing().

void disable_timing();

Example call: disable_timing();

3.6 drawline() - draws a line on the booth screens (CGA card). The line starts at p1 and extends to p2 and is drawn on the active page. This routine automatically selects the booth screens (with a call to *selectscreen(BOOTH)*) and leaves the booth screen active. Thus, if you want to print something on the console screen after drawing a line on the booth screens you must call *selectscreen()*. Config.maxx-1 is the largest possible x value on the screen and config.maxy-1 is the largest possible y value.

void drawline(int p1x, int p1y, int p2x, int p2y, struct vconfig config);

Input:	int p1x;	the x location of the first point.
	int p1y;	the y location of the first point.
	int p2x;	the x location of the first point.
	int p2y;	the y location of the first point.
	struct vconfig config;	the structure returned by *init2video()*.

Example call: drawline(p1x,p1y,p2x,p2y,config);

3.7 drawtext() - puts text on the booth screens starting at a particular row and column. The rows are numbered from 0 to config.textrows-1 and the columns are numbered from 0 to config.textcols-1. As with *drawline()*, this routine selects the booth screen and leaves that as the currently active video display, so it is necessary to call *selectscreen(CONS)*, in order to print on the console screen.

void drawtext(int row, int col, char *text);

Input:	int row;	the row number of the first character in text.
	int col;	the column number of the first character in text.
	char *text;	character string to be printed on the booth screens.

Example call: drawtext(config.textrows/2,config.textcols/2,"+");

3.8 enable_timing() - sets the system up for millisecond timing of button presses from the response boxes. This routine (1) clears extraneous characters from the system's keyboard buffer, (2) sets two global variables (g_msec and g_button_count) to zero, and (3) replaces the external keypad (serial port) interrupt service routine with one of our own. The PHONLAB serial port interrupt service routine stores the value of g_msec at the time of a button press, increments g_button_count, and then calls the old serial port service routine. This means that the identity of the button pressed is stored in the system's keyboard character buffer, and can be retrieved by normal input/output routines like getch(). This is normally done in get_responses().

void enable_timing();

Example call: enable_timing();

3.9 fast_timing() - changes the tick rate of the system clock from one tick every 55 ms to one tick every ms, and sets up a new interrupt service routine for clock interrupts. The interrupt service routine increments the global millisecond counter (g_msec) once per ms. Fast_timing() and slow_timing() keep track of the system time and restore the system clock to the correct time after an interval of millisecond timing. Also note that millisecond timing puts an extra burden on system operation; the computer's CPU must stop and service the clock interrupt every ms rather than every 55 ms, therefore it is advisable to use millisecond timing only when necessary. Also, in any program that calls fast_timing(), the timer must be set back to normal by calling slow_timing() before program termination. Otherwise, the system clock will keep sailing along at the fast rate and the correct system time will never be restored.

void fast_timing();

Example call: fast_timing();

3.10 free_memory() - turns off the DT2801A and frees the buffers allocated by load_files() used to store a set of CSpeech files in memory. It is necessary to call free_memory() before terminating any program in which files have been loaded into memory with load_files(). The routine also frees the large DMA buffer allocated by load_files() and turns off the DT2801A.

void free_memory(int numfiles);

Input: int numfiles; is the number of files read by load_files().

Example Call: free_memory(numfiles);

3.11 free_play_buf() - turns off the DT2801A and frees the buffer allocated by set_play_buf(). It is necessary to call free_play_buf() before terminating any program in which set_play_buf() has been called.

void free_play_buf(void);

Example call: free_play_buf();

3.12 get_listener_ids() - prints (using printf()) a request for a listener id number for each listening station. "Enter an id number for the listener at station # (0 if none):" is printed once for each listening station (the UCLA setup has two stations). The routine reads the number typed at the keyboard, stores the id numbers in a buffer used by get_responses(), and counts the number of listeners present for this run. The return value is the number of listeners.

int get_listener_ids();

Example call: numlisteners=get_listener_ids();

3.13 get_responses() - collects button press responses from the button boxes (via the extended keypad) during the response interval and stores the response value (button number), response time, and listener id number in a data array. Buttons are numbered from the left (1 through the number of buttons on the response box). The calling program must pass the address of a data buffer to get_responses() which assumes that the array is two dimensional, the first dimension being data type (id number, button, or time) and the second dimension being response instance. Following a call to get_responses() the data array will be filled in the following manner:

	listener id	button number	response time
first response	[0][0]	[0][1]	[0][2]
second response	[1][0]	[1][1]	[1][2]
nth response	[n][0]	[n][1]	[n][2]

Get_responses() is written to accommodate a variety of hardware setups. The number of response alternatives (buttons on the button boxes) is passed to the routine and it uses this information in decoding responses. NOTE: one assumption in get_responses() is that the response box buttons are wired to number keys on the external keypad with the buttons for the first listener starting at "1" and the buttons for the second listener starting at "n+1", where "n" is the number of buttons on each box. So, in a two button hardware setup the left buttons are wired to odd number keys on the keypad and the right buttons are wired to even numbers. Also get_responses() keeps track of whether a response has already been logged from a listener and keeps only one response (the first). If a listener fails to respond in the allotted response interval, values of 0 are entered for button number and response time for that listener.

```
void get_responses(int data[][3], int nsubs, int bps, int maxtime);
```

Input: int data[][3]; the name (or address) of an array to store the response data.
int nsubs; the number of listeners from whom to expect responses.
int bps; the number of buttons per listener;
int maxtime; the duration of the response interval in ms.

Output: int data[][3] is filled with data for this trial.

Example call: get_responses(data,2,2,2000);

3.14 init2video() - initializes the EGA and CGA cards for dual video mode. The EGA card (experimenter's screen) is set for 80X25 monochrome text mode, and the CGA card (booth screens) is set for high resolution black and white graphics. Before exiting any program in which a call to init2video() has been made it is necessary to call restore2video(), otherwise the computer will be in an abnormal video state on program exit. Also, init2video() should be called before calling any other dual video routines (see section 2.4). The address of a video configuration structure must be passed to init2video. The type of the structure is:

```
struct vconfig {        /* information for BOOTH video display */
        int maxx;       /* number of graphics x pixels */
        int maxy;       /* number of graphics y pixels */
        int textcols;   /* number of text columns */
        int textrows;   /* number of text rows */
}
```

This structure is filled by init2video() and is passed to the various dual video routines. NOTE when using dual video routines it is necessary to link your program to the library ctp_m3l.lib because the dual video routines use low level routines from the C TOOLS PLUS (ctp) library supplied by Blaise Computing

```
link /CO vidtest.obj,,,ctp_m3l+phonlab;
```

void init2video(struct vconfig far *config);

Input: struct vconfig far *config; a pointer to a video configuration structure.

Output: The booth screen is cleared.

Example call: init2video(&config);

3.15 l_play() - plays out a CSpeech file which has been previously opened with open_files(). The files are numbered from 0 to numfiles-1.

void l_play(int i);

Input: int i; a file number from 0 to numfiles-1.

Example call: l_play(i);

3.16 l_playwave() - plays a CSpeech file which has been opened by open_file(). Set_play_buf() must be called before calling open_file()/l_playwave().

void l_playwave(FILE *fd, long nsamples, long samrate);

Input: FILE *fd; a file pointer returned by open_file().
long nsamples; the number of samples in the file, returned by open_file().
long samrate; the sampling rate in Hz, returned by open_file().

Example call: l_playwave(fd,nsamples,samrate);

3.17 labels() - prints two text labels on the bottom corners of the listeners' video monitors. For identification or discrimination experiments this routine can be used to present the response alternatives. To avoid handedness effects in reaction time experiments the association between buttons and response alternatives can be counterbalanced across blocks within an experiment session.

short labels(char *left, char *right, short active);

Input: char *left; character string for the left label.
char *right; character string for the right label.
short active; number of the current active page in video memory.

Output: short active page after this call to labels.

Example call: active=labels("split","slit",active);

3.18 labels2() - is the dual video version of *labels()*. It puts labels on the booth screens (CGA) and does nothing to the console screen (EGA).

void labels2(char *left, char *right, struct vconfig config);

Input: char *left; character string for the left label.
char *right; character string for the right label.

struct vconfig config; structure returned by *init2videb()*.

3.19 load_files() - loads a list of CSpeech files into memory for future playback using play(). The routine opens a list file which contains the file names of CSpeech files to be presented. Here's an example of a list file (which might be named split.lis) for a hypothetical five step slit-split continuum:

```
split1.snd
split2.snd
split3.snd
split4.snd
split5.snd
```

Use a text editor (such as M, the MicroSoft text editor) to create the list file. In addition to load_files() there are a couple of utility programs (level.exe and set_amp.exe) which use list files. Load_files() establishes a DMA buffer for the playback routine, reads each file into memory, stores the length and sampling rate of each file, and returns the number of files read. Later, when the program calls play() this information is accessed and the CSpeech file is output through the DT2801A. Thus, files with different sampling rates can be intermixed in an experiment (not that one would actually have a need to do this). Note that the user's program must not call load_files() more than once unless free_memory() is called between each successive call to load_files(). Note also that l_play() will not play files opened by load_files().

int load_files(char *name);

Input: char *name; the name of list file.

Output: the number of files loaded into memory.

Example call: numfiles=load_files("split.lis");

3.20 open_file() - opens a single CSpeech file for audio output using l_playwave(). We use this routine in programs which require random access to a large number of files, because of memory limitations with the load_files()/play() routines and system limitations on the number of files which can be open at one time with the open_files()/l_play() routines. One other note: the size of the input buffer associated with the CSpeech file is increased from 256 bytes to 8192 bytes. This increase in buffer size improves the speed of disk transfer during digital to analog conversion.

FILE *open_file(char *name, long *nsamples, long *samrate);

Input: char *name; character string containing the name of the CSpeech file.
long *nsamples; address of a variable to store the number of samples.
long *samrate; address of a variable to store the sampling rate.

Output: a pointer to the CSpeech file.

Example call: fd=open_file("split1.snd",&nsamples,&samrate);

3.21 open_files() - opens a list of CSpeech files for later presentation using l_play(). See load_files() above for a description of the expected format of the list file. This routine is normally used for files which are too large to fit into memory using load_files(). Note that open_files() can be called more than once in a program, but that close_files() must be called between successive calls to open_files(). The routine allocates a DMA buffer for the playback routine, opens the list of files, reads their lengths and sampling rates, and returns the number of files successfully opened. File pointers, file lengths, and sampling rates are stored for later reference in l_play(). As with open_file(), the size of the input buffer associated with each file is increased, thus improving the speed of disk access during digital to analog conversion.

int open_files(char *name);

Input: char *name; a character string containing the name of the list file.

Output: the number of files successfully opened.

Example call: numfiles=open_files("split.lis");

3.22 play() - plays a CSpeech file from memory. The files used by play() must have been loaded into memory by load_files() prior to any call to play(). Load_files() numbers the CSpeech files from 0 to n-1 where n equals the number of files loaded into memory.

void play(int i);

Input: int i; the number of the CSpeech file, from 0 to n-1.

Example call: play(i);

3.23 randomize() - randomizes without replacement a set of numbers from 0 to n-1. The routine uses the value of the system clock as the seed value for the random number generator, therefore, each successive call will produce a different random order. The algorithm used is called a Monte Carlo shuffle. After calling randomize() the buffer passed to the routine will contain integers from 0 to
n-1 in a random order.

void randomize(int n,int *order);

Input: int n; the number of items to be randomized.
int *order; the address of a buffer to store the random order.

Output: order[] contains a sequence of random numbers.

Example call: randomize(numtokens,order);

3.24 restore_graphics() - sets the CGA video mode to the system default mode. Call this routine before exiting from any program which has called set_graphics().

void restore_graphics();

Example call: restore_graphics();

3.25 restore2video() - sets the graphics displays in a dual video setup to the system default modes. Call this routine before exiting from any program which has called *init2video().*

void restore2video();

Example call: restore2video();

3.26 rt_feedback() - displays a single reaction time value in the center of the listeners' video displays for a fixed number of milliseconds. After displaying the number the routine clears the screen.

short rt_feedback(long rt, int delay, short active);

Input: long rt; a long integer value, usually a reaction time value.
int delay; the number of milliseconds to keep the number on the screen.
short active; the current active page in video memory.

Output: the current active page in video memory.

Example call: active=rt_feedback(ave_time,500,active);

3.27 rt_feedback2() - is the dual video version of *rt_feedback()*.

void rt_feedback2(long rt, int delay, struct vconfig config);

Input: long rt; a long integer value, usually a reaction time value.
int delay; the number of milliseconds to keep the number on the screen.
struct vconfig config; the structure returned by *init2video()*.

Output: none.

Example call: rt_feedback2(ave_time,500,config);

3.28 selectscreen() - selects one of the two video adapters (using the global constants BOOTH and CONS) in the dual video configuration for future display commands. The most common use for this routine is to select the console screen prior to printing messages for the experimenter, because the routines for displaying things on the booth screen all automatically select the booth video adapter. Also, note that your program should call selectscreen(BOOTH) before initiallizing the mouse.

void selectscreen(int adapt);

Input: int adapt; one of the constants BOOTH or CONS.

Example call: selectscreen(CONS);

3.29 set_play_buf() - allocates a large buffer for DMA transfers during digital to analog conversion and initializes the DT2801A. Use this routine with l_playwave() to play out a single file. Do not call set_play_buf() if you are using either load_files() or open_files() to prepare files for D/A conversion. These two routines have internal calls to set_play_buf(). Also, do not call set_play_buf() more than once in a program, unless you call free_play_buf() prior to the second call to set_play_buf(). The DMA buffer must be freed by free_play_buf() before program termination.

void set_play_buf();

Example call: set_play_buf();

3.30 set_graphics() - puts the CGA card into high resolution graphics mode and turns the text cursor off. Call once at the beginning of any program which uses video display routines.

void set_graphics();

Example call: set_graphics();

3.31 slow_timing() - resets the clock tick rate of the system clock to one tick every 55 ms, restores the old clock interrupt service routine, and updates the system time of day. The time of day update converts the elapsed milliseconds since calling fast_timing() into hundredths of a second, adds this time to the old system time of day which was saved by fast_timing(), and resets it with the new value.

void slow_timing();

Example call: slow_timing();

3.32 t_scope() - is an implementation of a video tachistoscope. The routine clears the subjects' screen, displays text in the center of the screen for a specified number of milliseconds (setting g_msec to 0 just after showing the text), then clears the screen again. The routine waits for a vertical retrace of the screen before showing the text (and resetting the millisecond timer), and before clearing the screen. Note that this routine requires that fast_timing() has been called in order for the presentation time to be accurate.

short t_scope(char *text, int time, short active);

Input:	char *text;	character string to be displayed.
	int time;	number of milliseconds to show the text.
	short active;	current active page of video memory.

Output: current active page of video memory after the routine.

Example call: active=t_scope("go",20,active);

3.33 t-scope2() - implements a video tachistoscope in the dual video environment. Its mode of operation is identical to the single video routine *t_scope()*.

void tscope2(char *text, int time, struct vconfig config);

Input:	char *text;	character string to be displayed.
	int time;	number of milleseconds to show the text.
	struct vconfig config;	structure returned by *init2video()*.

Output: none.

Example call: t_scope2("go",20,config);

3.34 wait() - pauses program operation for a specified number of milliseconds. This routine assumes that fast_timing() has been called. G_msec is set to 0 and the routine loops until g_msec is equal to or larger than the input number of milliseconds.

void wait(int msec);

Input: int msec; number of milliseconds to delay.

Example call: wait(500);

4 Utility Programs

This section describes three utility programs which can be used to prepare speech stimuli for an experiment, calibrate the listening level at the headphones, and listen to individual CSpeech files. Two of the most important tools which are used in preparing stimuli are (1) the CSpeech waveform editing and speech analysis software (Milenkovic & Read, 1990), and (2) the Klatt software speech synthesizer (Klatt & Klatt, 1990). The reader is referred to separate documentation concerning these tools. We use an implementation of the Klatt synthesizer which produces sound files which have CSpeech headers on them, so these files can be read into CSpeech for verification of the synthesis parameters and read and played by the relevant PHONLAB routines.

4.1 Level

The utility program level.exe takes a list file as input and (1) checks that the sampled data format is appropriate for the playback routines in PHONLAB, (2) converts the file to the appropriate format if necessary, (3) finds the peak RMS amplitude in each CSpeech file, and (4) if requested scales the speech samples so that each file has the same (target) peak RMS amplitude level.

Here's an example of a list file (which might be named split.lis) for a hypothetical seven step slit-split continuum. Use a text editor (such as M, the MicroSoft text editor) to create the list file.

```
split1.snd
split2.snd
split3.snd
split4.snd
split5.snd
split6.snd
split7.snd
```

As shown below, level takes two optional command line arguments. The first (“-l”) requests that the samples be scaled to the target peak RMS amplitude, and the second is the name of the list file. If these arguments aren't present in the command line the program prints requests for the information.

```
C:>level (-l) FILENAME
```

Regardless of whether you request that the samples be scaled, the program asks for a target RMS amplitude level (45dB is a reasonable value for this parameter), and shows the existing peak RMS level for each file and the peak RMS level after scaling. The program also prints a message if any waveform peaks got clipped when scaling to a particular target RMS amplitude level. If -l is not entered on the command line, or if you say that you don't want the amplitudes leveled, the program won't replace the old waveforms with the new ones. It is a good idea to try a test run without leveling the amplitudes to be sure that the target value won't produce any peak clipping.

4.2 Set_amp

Set_amp.exe plays a sound file repeatedly so the experimenter can adjust the listening level at the headphones. When you type set_amp at the DOS cursor the program asks two questions: (1) what is the name of the list file? and (2) which CSpeech file in the list file should be played? The expected answer to this last question is the number of a CSpeech file where the first file is number 1 and the last file in the list is N.

4.3 Play

Play.exe is a stand alone playback routine. It expects the name of a CSpeech file to be entered on the command line and will complain if this is not done. This program uses open_file() and l_playwave() which were described above. To play a CSpeech file type:

```
C:>play FILENAME
```

5 Conclusion

We have chosen to write a description of the speech perception facilities at UCLA at this time because the system has been functioning for some months and most of the bugs seem to be worked out. However, this is not a static system. We anticipate in the coming months to implement a dual video system in which the experimenter's monitor shows running tabulations, trial by trial information, and other system messages, while the listeners' monitors show button labels, reaction time feedback, or other relevant messages. We also plan to revise the playback routines to handle stereo output for dichotic listening experiments. And, of course, we expect to continue to improve the reliability of the code.

References

Brysbaert, M., Bovens, N., d'Ydewalle, G, & van Calster, J. (1989) Turbo Pascal timing routines for the IBM microcomputer family. *Behavior Research Methods, Instruments, & Computers*, **21**, 73-83.

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

Milenkovic, P.H. & Read, C. (1990) CSpeech Core User's Manual, CSpeech Version 3.1. Paul Milenkovic, 409 N. Eau Claire Ave. 102, Madison, Wisconsin 53705.

Appendix: An example program

This appendix is a program listing of an example program which makes use of the hardware and software described above. The program is a simple perceptual identification paradigm in which identification data and reaction time data for each stimulus in a continuum is collected.

```
/****************************************************************/
/*       ident.c  - an example program showing the use of the     */
/* phonlab library of routines for speech perception experiments.  */
/* Keith Johnson, Jan, 1992                                         */
/****************************************************************/
#include <stdio.h>                /* standard C functions */
#include <conio.h>
#include <graph.h>
#include <phonlab.h>              /* the library of specialized routines */

#define MAX_RESP_TIME 5000        /* time allowed for a response */
#define N_BUTTONS 2               /* number of buttons on each box */
#define N_TOKENS 9                /* maximum number of stimuli */

main ()
{
/************************************************/
/* variable declarations                        */
/************************************************/
        FILE *out;                /* output file pointer */
        int token,numtokens;      /* loop control variables */
        int block,numblocks;      /* loop control variables */
        int order[N_TOKENS];      /* random order buffer */
        char inname[12],outname[12]; /* filename strings */
        char resp1[8],resp2[8];   /* button labels */
char *result;               /* for gets() */
short active;               /* for use in screen handling */
int isi;                    /* interstimulus interval */
int numsubs;                /* number of listeners */
int resp[2][3];             /* store button press data here */
int i;
```

```
            disable_timing();                      /* finished collecting responses */
            slow_timing();                         /* clock speed back to normal */
                                                   /* write data to disk */
            for (i=0;i<numsubs;i++) fprintf(out,"%d\t%d\t%d\t%d\n",
               resp[i][0],order[token]+1,resp[i][1],resp[i][2]);
        }
                                                   /* Display message between blocks */
        if (block==(numblocks-1)) active=center("That's all",active);
        else active=center("Please wait",active);
        i=getch();                                 /* wait for keypress */
      }
/***********************************************************/
/*  restore system to original state                                         */
/***********************************************************/
      free_memory(numtokens);
      restore_graphics();
      fclose(out);
}
```

An example output file from the program ident.c

```
IDENT.C perceptual identification
List file name: vow.lis
Response alternatives: [i]  [u]
ISI = 1000  Max response time = 5000
id #  token   button   time
14      6       1       735
14      5       1       814
14      9       2       753
14      8       2       713
14      4       1       536
14      1       1       555
14      3       1       556
14      7       2       636
14      2       1       560
14      6       2       706
14      7       2       555
14      4       1       518
14      5       1       537
14      9       2       564
14      8       2       633
14      3       1       569
14      1       1       572
14      2       1       563
```

The Effects of Implosives on Pitch in SiSwati

Richard Wright

Aaron Shryock

0. Introduction

It is widely known that the fundamental frequency (F_0) at the onset of a vowel is significantly higher after a voiceless consonant than it is after a voiced consonant (ex. House and Fairbanks 1953, Mohr 1971, Hombert *et al* 1979). Recently, Traill *et al* (1987) conducted a study to determine the effect of breathy voiced, or depressor, consonants on F_0 of the following vowel. Depressors were shown to have a dramatic lowering effect on F_0. The perturbation of F_0 caused by various consonants is of interest to theories of speech perception because it represents a possible cue to consonant manner. Likewise, it is fundamental to theories of tonogenisis in which the raising or lowering of F_0, once a cue to stop manner, persists after a consonant is lost, thus contributing to the development of tones (Hombert *et al*, 1979). Despite the interest and despite thorough studies on the effects of other types of consonants, there has been no conclusive study of the effect of implosive consonants on F_0 of the following vowel[1]. It has been predicted that implosives should have a raising effect similar to voiceless stops, e.g. Hombert 1978. This prediction is based in part on the phonological patterning of implosives with respect to tone (Hyman and Schuh 1974, Schuh 1978). To investigate the Phonetic effect of implosives on the F_0 of the onset of the following vowel, we conducted a study examining the consonantal effects on F_0 in SiSwati.

SiSwati is an Nguni language spoken in Swaziland and in parts of eastern South Africa. It was chosen for this study because its consonant inventory contains voiceless aspirated, voiced implosives and breathy voiced consonants, and because it has two contrasting tones: high and low. Tone is of interest because a study by Hombert (1978) suggests that in languages with contrastive tone, the duration of the F_0 perturbation may be shorter than it is in languages lacking contrastive tone.

In implosives, the larynx is lowered during oral closure resulting in rarefaction of air in the supraglottal cavity. When the oral closure is released, the negative pressure causes momentary ingressive air flow immediately followed by egressive air flow. Nearly all implosives are voiced. Lowering of the larynx is normally associated with a lower F_0, while stiffening of the vocal folds is associated with higher F_0. In order for the rarefaction of air in the supraglottal cavity to occur during the lowering of the larynx, the vocal folds are stiffened to maintain closure. The rarefaction of air in the supraglottal cavity causes a drop in pressure across the glottis which results in a greater rate of air flow at the onset of voicing. Increased airflow is associated with a higher F_0. Any one of these components of implosives might lead to a perturbation of F_0. It is important to keep in mind, however, that the duration of perturbation seen for other consonant types is longer than can be explained simply using a mechanical explanation.

[1] The one study that was carried out by Painter (1978) provides no conclusive evidence because there are several factors which detract from the findings. First, there was only one subject. Second, nonsense tokens were used creating an unnatural environment for the consonant. Third, implosives were confounded with labiovelars because both have negative oral pressure. The negative oral pressure in labiovelars is located between the velar closure and the labial closure while the negative oral pressure in implosives is found between the place of closure and the glottis. The differences in pressure is crucial in considering the possible causes of F_0 perturbation.

1. Previous Studies

A good example of a study on the effects of consonants on F_0 is Hombert (1978). In this study, two experiments were performed. In the first experiment, the effect of voiceless aspirated stops, voiced oral stops and nasal stops in English was investigated. F_0 was measured from the onset of voicing of the vowel. The results indicated that voiceless aspirated stops had a raising effect and the voiced stops and nasals had a lowering effect on the onset F_0 of the following vowel. The duration of perturbation was between 50 and 80 milliseconds (ms). Interestingly, there was considerable between-speaker variation in the degree and duration of the perturbation.

In the second experiment, the effect of voiced and voiceless stops in Yoruba, a language with three contrastive tones (high, mid and low) was investigated. The results from this experiment agreed with those of the first; voiceless stops had a raising effect and voiced stops had a lowering effect. The degree of perturbation was graded with raising effects being greatest in low tone vowels and least in high tone vowels, and with lowering effects being greatest in high tone vowels and least in low tone vowels. The duration of the perturbation was shorter in Yoruba than it was in English, indicating that the presence of contrastive tone may reduce the length of time that the consonantal effect persists into the vowel.

A particularly thorough study of the effect of depressor (breathy voiced) consonants on the F_0 of the following vowel was conducted by Traill et al (1987). The language studies was Zulu which has both depressor consonants and two contrastive tones. The results indicated that after depressor consonants, high tone was realized as a rising tone that started below the lowest point of the low tone range and rose throughout the duration of the vowel, reaching the high tone range only in the last 20 ms of the vowel.

2. A Pilot Study

With a study on implosives in mind, we conducted a pilot study on SiSwati consonants. The two purposes of this study were to establish that implosives in SiSwati were truly implosive, and to collect data on implosive, breathy voiced and voiceless consonants that would direct later studies.

One speaker, whose native language is SiSwati, participated in the pilot study. Bilabials were chosen as the set of consonants to study. This decision was made for two reasons. First, a nasal catheter is needed to obtain oral pressure measurements for alveolar stops whereas oral pressure during bilabial stops can be measured by simply inserting a small tube between the lips. The catheter is a fairly intrusive method of measurement and seemed unnecessary for the purposes of our study. Second, judgments we and others made by ear and by measurements from wave form displays indicated that the alveolar stop which has been described as implosive in SiSwati is not reliably imploded.

Using pressure and flow measurements taken with a mask worn over the mouth and nose and with a small tube inserted between the lips of the speaker, we determined that SiSwati implosives are in fact implosive. The negative oral pressure reached between 20 and 40 cm^3 H_2O, *and the larynx was lowered during the oral closure.*

Secondly, a list of tokens was read by the speaker and recorded in a sound treated room at the UCLA Phonetics Laboratory. The list consisted of ten real **CVX** words where C=[p, p^h, ɓ, b^ɦ, m], and V=[à] or [á]. Each consonant was found before both a high tone and low tone[2]. The recording was digitized at 16 kHz using C-speech on an IBM PC. F_0 values were measured using

[2] The word list was compiled from A Concise Siswati Dictionary, D. K. Rycroft, and by consultation with native SiSwati speakers.

a pitch tracking program in C-speech. Measurements were taken starting at the onset of the voicing of the vowel.

We found that implosives had no reliable raising effect in the F0 of vowels with high tones and that there was a slight effect on the F0 of vowels with low tones. The breathy voiced consonants showed the lowering effect described in Traill et al (1987). However, the duration of the lowering effect was longer, lasting throughout the vowel with the F_0 <u>never</u> rising to the high tone range. Voiceless aspirated consonants had the effect seen in earlier studies; the F_0 of the following vowel was raised with the effect being greater on low tone vowels than it was on high tone vowels.

3. A Second Study

In order to test the findings of the pilot study a follow-up study was undertaken. Four subjects, whose native language is SiSwati, participated in the second study. Three of them were male and one was female. Three of the subjects were recorded at a radio station in Swaziland and one was recorded in a sound treated room at the UCLA Phonetics Laboratory. For three of the subjects, the wordlist consisted of 8 real word **CVX** tokens, where **C**=/ɓ, $\text{m}^{\text{ɦ}}$, m, p^{h}/[3] and where **V**= /á/ or /à/. The tokens were spoken in the frame "Tsani ____ futsi." (Say ___ again). Speaker 1 recorded 6 repetitions of the token sets, speakers 2 and 3 recorded 8 repetitions each. The same tokens were used for speaker 4 and three more tokens were added, these included /$\text{b}^{\text{ɦ}}$/, and /m/. Speaker 4 recorded 10 repetitions of the token set. The recordings were digitized at 19.5 kHz using CECIL on an IBM PC clone. F0 was measured using the pitch tracking program in CECIL . Measurements at vowel onset and at 5 ms intervals thereafter. The results were submitted to a two tailed unpaired T-tests, where the dependent variable was the F_0 values and the independent variables were the consonantal categories and where significance was set at .01.

4. Results

The results of the second study, illustrated in figures 1-4 below, agree with those of the pilot study. The voiceless aspirated stops show the typical raising effect seen in previous studies. There is gradation in the raising of F_0 similar to that found by Hombert (1978); low tone vowels are subject to greater raising of F_0 than are high vowels.

The voiced implosive consonants showed no reliable raising nor lowering effect. Although F_0 appears to be raised slightly at the onset of low tone vowels, this a reflection of the normal low tone contour as can be seen when the implosives are compared with the depressor consonants in figure 5. High tone is plotted using solid lines low tone using broken lines and filled symbols.

[3] Because of unavoidable problems in collecting data in the field, /$\text{b}^{\text{ɦ}}$/ was omitted from the data set.

Figure 1

Implosive vs Voiceless Aspirated Stops: Speaker 1

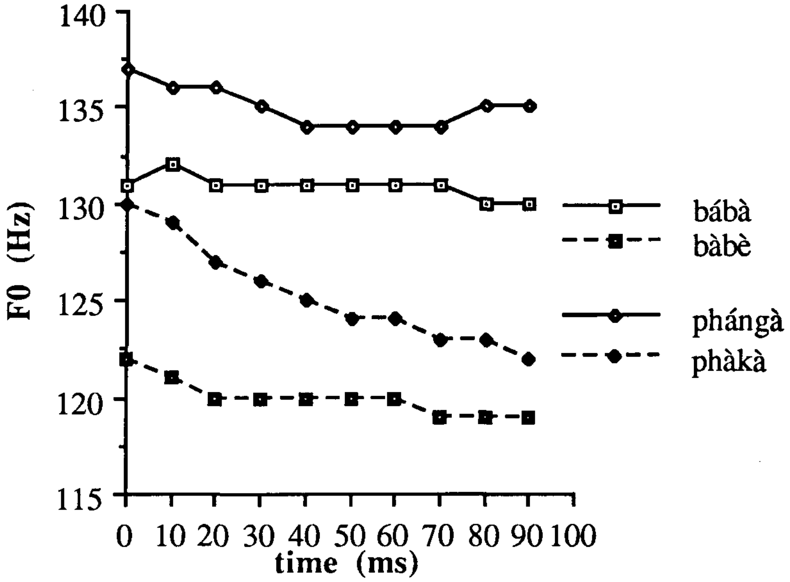


Figure 2

Implosive vs Voiceless Aspirated Stops: Speaker 2

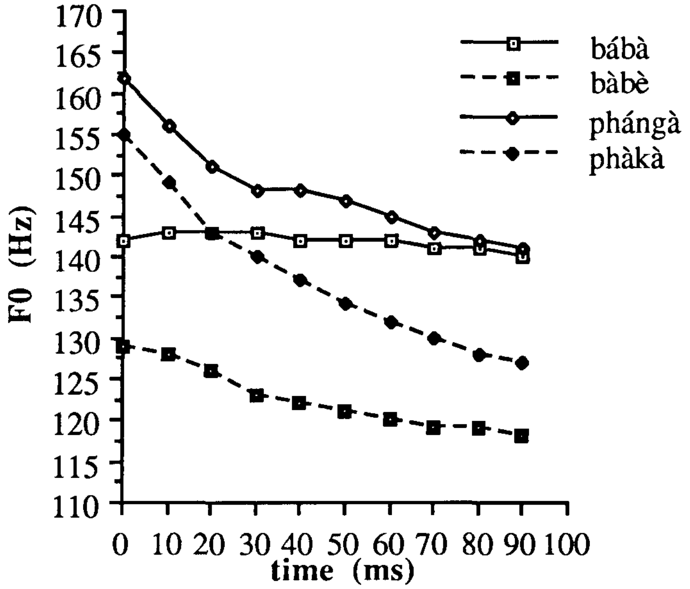


Figure 3

Implosive vs Voiceless Aspirated Stops: Speaker 3

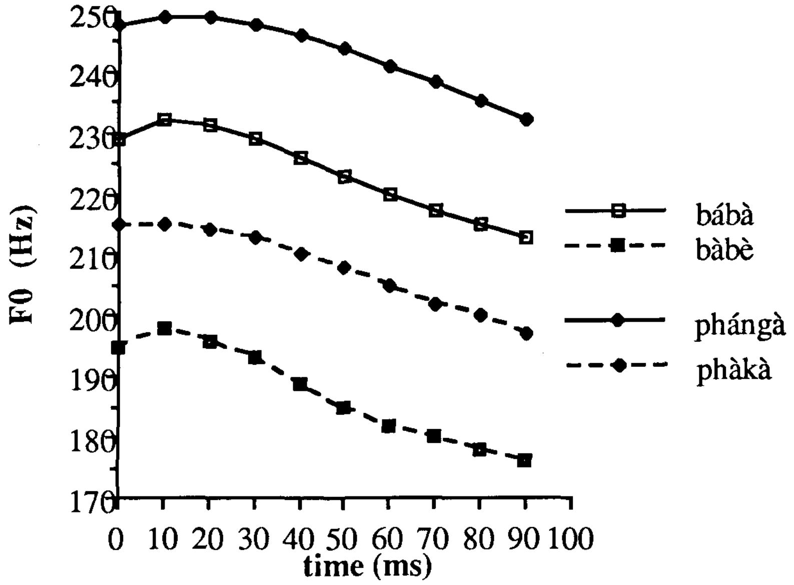


Figure 4

Implosive vs Voiceless Aspirated Stops: Speaker 4

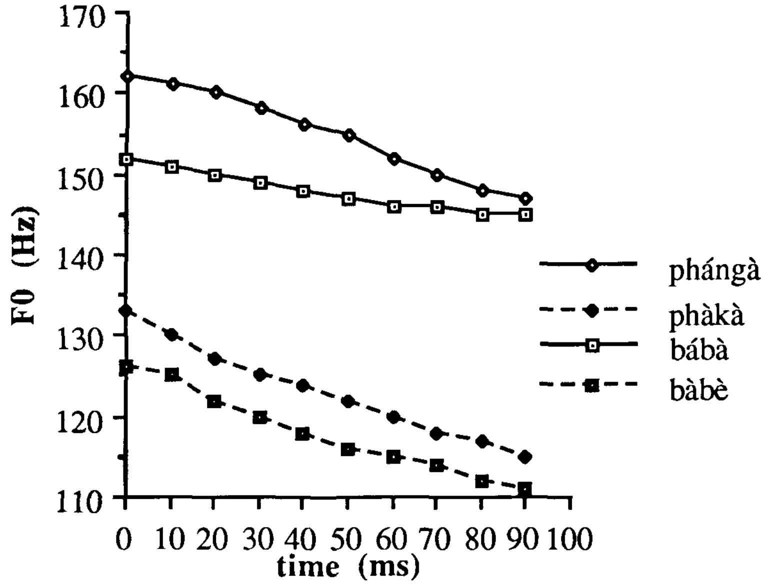
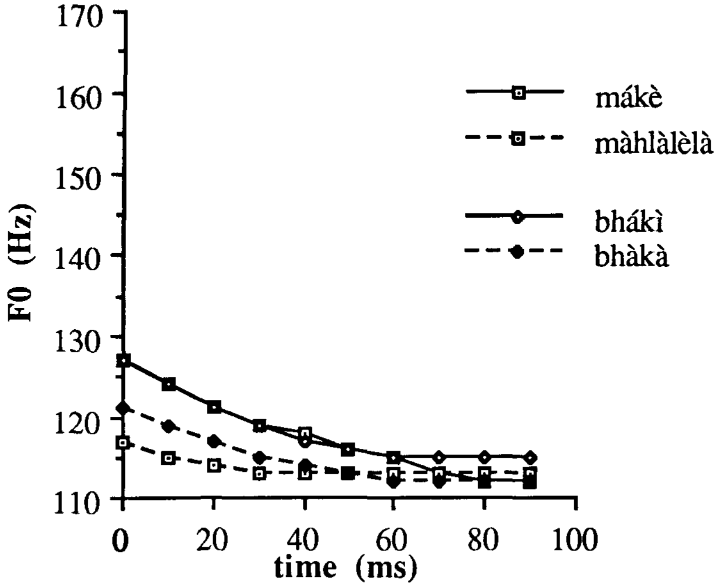


Figure 5

"Depressors": Speaker 4



For all speakers, the voiced implosive stop had no reliable raising effect. There was a reliable raising effect for all speakers for all instances of the voiceless aspirated stop. The duration of the perturbation is considerably longer than was found in Yoruba (Hombert, 1978). The data of speaker 3 presented some difficulty since the F_0 after the voiceless aspirated stop never reached its high tone target. However, the raising effect is reliable when compared across consonants. As in Hombert's (1978) data, there was considerable between speaker variation in the degree and duration of the perturbation.

The depressor consonants showed a dramatic lowering of the F_0 such as that described by Traill et al (1987). However, rather than rising throughout the duration of the vowel in the case of high tone, the F_0 remained depressed for the duration of the vowel. This phenomenon was observed consistently for all speakers. Figure 5 is speaker 4's data presented as an illustration. It is interesting to note that while high tone vowels were realized as having an F_0 within the low tone range, there is a reliable difference between the F_0 of high tone vowels and the F_0 of low tone vowels after depressor consonants. Speaker 5's data was chosen because it contains the larger set of consonants. The effect of the two types of breathy consonants was identical. Note that even with the depressor consonants there is a sloping low tone contour. This type of depression has been observed independently in SiSwati and in some of the Mijikenda languages.[4]

5. Conclusion

Voiceless unaspirated stops in SiSwati have a raising effect on F_0 in both high and low tone vowels. The effect persists at least 40 ms into the vowel often persisting up to 80 ms. The degree of perturbation is greater in low tone vowels than it is in high tone vowels. While this study agrees with Hombert's (1978) findings on the degree and gradation of raising of F_0 after voiceless consonants, it does not agree with his finding that the duration of the effect in tone languages is shorter. The difference in duration may be due to the fact that SiSwati has two tones

[4] Charles Kisseberth, personal communication.

while Yoruba has three, and SiSwati tone functions more like an accent system with mobile tone assignment while Yoruba has rigidly fixed lexical tone.

The breathy consonants show a strong depressor effect on F_0, agreeing with Traill et al (1987). It appears that SiSwati has phonologized the depression to such a degree that a high tone fails to appear on the vowel following the depressor consonant. Further studies are needed in order to determine the exact nature of this phenomenon.

The implosive stop showed no raising effect, contrary to some hypotheses in the literature. Neither did it show the lowering effect on F_0 typical of voiced stops. This finding has implications for theories of tonogenesis in which the perturbations caused by stops contribute to the development of tones. It may also have implications for reconstructive historical linguistics because implosives pattern separately from voiceless stops. This may be important in the light of theories, such as that of Stewart, which propose both lenis and fortis classes of stops in Proto-Bantu, in which the voiced lenis class are hypothesized to have been implosives. It is interesting that implosives, which have qualities that are correlated with both raising and lowering, seem to be neutral to F_0. This neutrality may be due to the various qualities canceling each other out. Finally, it is important to note that what has been defined as "implosive" varies greatly across and even within languages (Lindau, 1984). Therefore, it may be possible to test the neutralization hypothesis by comparing implosives in different languages with varying degrees of ingressiveness of airflow (ex. Sindhi vs Degema), or varying degrees of voicing amplitude and closure time (ex. Degema and Hausa), to see if there is any F_0 perturbation that can be correlated with a variation in quality of the implosive.

Acknowledgments

We are very grateful to Kim Thomas without whom our data could never have been collected. We would also like to thank all the speakers who participated in the study.

References

Hombert, J-M. 1977. Consonant Types, Vowel Height and Tone in Yoruba. *Studies in African Linguistics* **88** (2), 173-190.

Hombert, J-M. 1978. Consonant Type, Vowel Quality, and Tone. In Fromkin V. A. *Tone: A Linguistic Survey*. pp. 77-112. New York: Academic Press.

Hombert, J-M., J.J. Ohala and W.E. Ewan 1979. Phonetic explanations for the development of tones. *Language* **55**, 37-58.

Hyman, L. and R.G. Schuh 1974. Universals of tone rules: Evidence from West Africa. *Linguistic Inquiry* **5**, 81-115.

Lindau, M. 1984. Phonetic differences in glottalic consonants. *Journal of Phonetics* **12**, 147-155.

Painter, C. 1978. Implosives, Inherent Pitch, Tonogenesis, and Laryngeal Mechanisms. *Journal of Phonetics* **6**, 249-274.

Rycroft, D.K. 1981. *A Concise Siswati Dictionary*. Johannesburg: Witwatersrand University Press.

Schuh, R.G. 1978. Tone Rules. In Fromkin V. A. *Tone: A Linguistic Survey*. pp. 77-112. New York: Academic Press.

Traill, A., J.S.M Khumalo and P. Fridjhon 1987. Depressing Facts About Zulu. *African Studies Journal* **46** (2), pp.255-274.