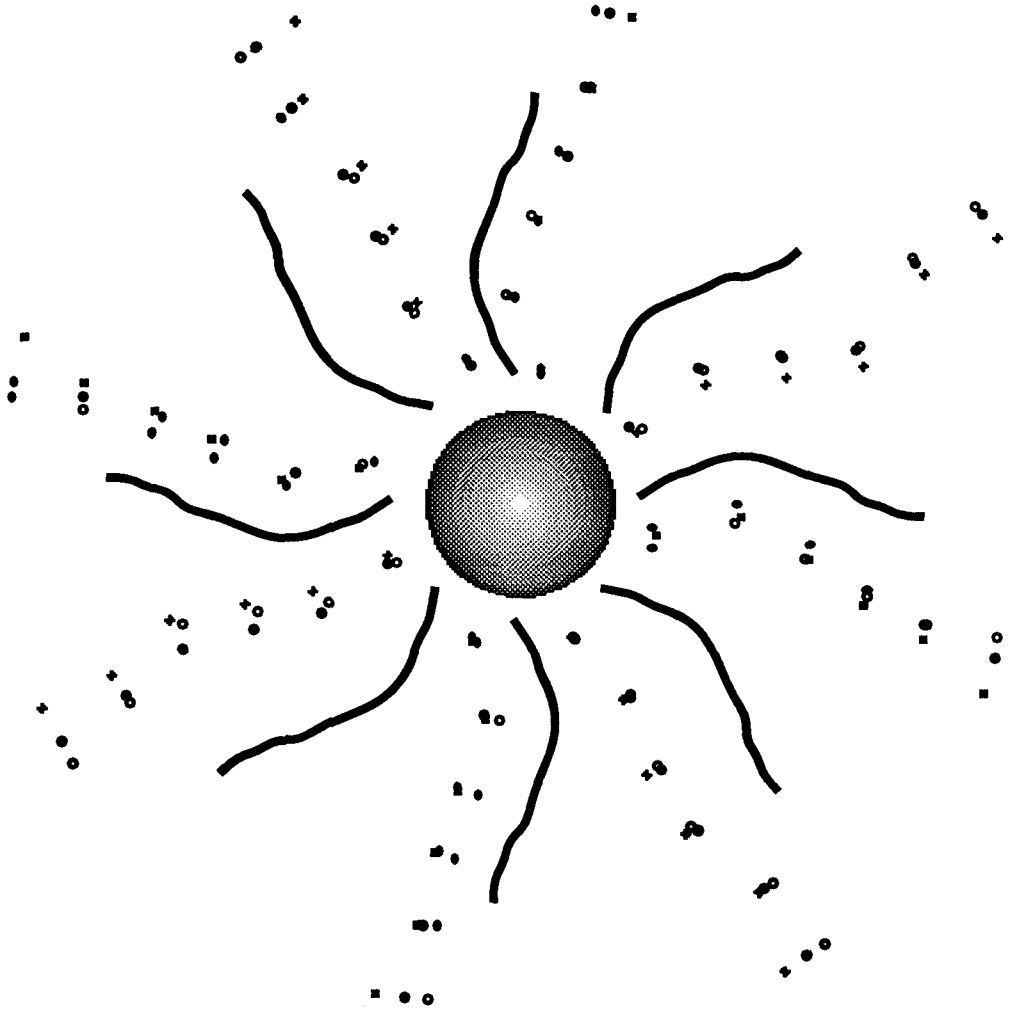
UCLA Working Papers in Phonetics



Number 80 December, 1991

The UCLA Phonetics Laboratory Phonetics

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
Paul Kirk
Jenny Ladefoged
Peter Ladefoged
Mona Lindau
Ian Maddieson
Joyce McDonough
Pam Maelzer
Paroo Nihalani
Bonny Sands
Stephan Schuetze-Coburn
Michael Shalev
Aaron Shryock
Siniša Spacik
Donca Steriade
Henry Teheranizadeh
Kimberly Thomas
Anthony Traill
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 NS-22726
USHHS grant 1 RO1 DC00642
USHHS grant 1 T32 DC00029
NSF grant BNS 9107004
and the UCLA Department of Linguistics.

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

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

UCLA *Working Papers in Phonetics* is edited by Ian Maddieson

UCLA Working Papers in Phonetics 80

December 1991

Table of Contents

A Re-evaluation of Implosives in Sindhi

Paroo Nihalani
National University of Singapore

1. Introduction

Sounds of one language may differ from those of another because of the phonetic value of the segments along the same continuum. To take an example, the linguistic specification that distinguishes between [p] and [b] in English is that they are [- voice] and [+voice] respectively. The articulatory instruction that accompanies the feature [+voice] is "vibrate the vocal folds". In order to implement this instruction a number of articulatory instructions have to be performed, such as keeping the vocal folds sufficiently lax, reducing the distance between the vocal folds, keeping the airflow through the glottis powerful enough to cause vibration, and maintaining the difference between the subglottal and supraglottal air pressure by lowering the larynx, allowing air to escape through a small velic opening, and/or expanding the walls of the pharynx. "Vibrate the vocal folds", however, is the primary instruction that is associated with the linguistic feature [+ voice] and the rest of the articulatory gestures are ways of implementing this instruction. Speakers of different language backgrounds choose different combinations of parameters for the implementaion of voicing of stops. The phonetic implementation of these diffferences is as important as those in the sound patterns. In order to illustrate the point, I will discuss some phonetic differences between implosives in Sindhi and a few other languages.

Implosives have been traditionally characterized as glottalic ingressive sounds produced by lowering the vibrating glottis (Catford, 1939; Pike, 1943). Lindau (1984, P. 152) notes that Hausa implosives are produced with aperiodic, inefficiently closing vocal cord vibrations and that there is considerable speaker to speaker variation between implosives in languages, and that languages may differ in the way that they maintain distinction between implosives and the corresponding plosives. Ladefoged (1964, p. 6) noted that his Igbo implosives only produced negative pressures 8% of the time. Ladefoged therefore observes:

> "... the action of the vocal cords in the production of these implosive sounds has been one of a leaky piston... Often the piston is so leaky that the airstream is not actually ingressive nor the sounds really implosive. In many of the languages I have observed (cf. Ladefoged 1964) the pressure of the air in the mouth during an ingressive glottalic stop is approximately the same as outside the mouth, since the rarefying action of the downward movement of the glottis is almost exactly counterbalanced by the leakage of lung air up through the vocal cords. Although these sounds may be called implosives, in ordinary conversational utterances air seldom flows into the mouth when the stop closure is released." (Ladefoged 1971: 25-26)

In this connection, Painter (1978: 254) observes: "Despite Ladefoged's caveat (1964: 6) that his Igbo implosives only produced negative pressures 8% of the time...my physiological data for Gã, Sindhi and Yoruba show negative pressures most of the time". More recently, Nihalani (1986) has shown that there exist natural languages like Sindhi (spoken in India and Pakistan) and Kalabari (spoken in Nigeria) in which implosives do involve an ingressive air flow in addition to the downward displacement of the vibrating glottis. The quantitative measurements of the air flow dynamics run counter to Ladefoged's assumption that there are no real implosives.

Ladefoged (personal communication) has commented that Nihalani's findings (1986) are based on his own speech (one single speaker), and that the aerodynamic data are collected from citation forms. Ladefoged has valid criticisms in that we should always use a large enough sample on which to base our generalizations. It is obviously crucial to any study of this sort to have as many speakers as practicable, in order to increase the possibility of making more meaningful language-specific generalizations.

The purpose of this study was to expand the data on pressure-flow dynamic from a much large number of informants in order to explore the aerodynamic characteristics of implosives in Sindhi and also to determine whether these articulatory strategies are consistent within a language or vary according to speaker-specific idiosyncracies.

2. Test materials

Data on intraoral pressure and oral air flow were collected from 3 speakers (1 male and 2 females, based in Los Angeles). A minimal pair contrasting the bilabial implosive and the voiced bilabial plosive in syllable initial position was selected, namely /ɓaru/ "child" and /baru/ "load". The language informants were requested to utter words in the carrier phrase "hi ____."

3. Instrumentation

The language informant speaks into a specially constructed mouthpiece pressed against the face, which takes the oral air flow through a calibrated resistance so that a pressure transducer provides a signal that is directly proportional to the rate of air flow (see Ladefoged 1991 for details). If one can find a language informant who is willing to tolerate a nasal catheter, then it is possible to record the pressure build-up behind stop closures anywhere in the vocal tract. As an alternative, a simple way of obtaining supraglottal air pressure and air flow data on just bilabial sounds was used by inserting a small tube between the lips.

These parameters were digitized along with the audio signal from a microphone at the rate of 11000 samples/sec. Figure 1 is an example of the aerodynamic data recorded in the Phonetics Lab, UCLA. The top channel records the audio signal, the middle channel represents oral air flow and the bottom channel represents intraoral air pressure.

4. Results

Figure 1 gives the aerodynamic record of the word [ɓaru] "child" for one of the female speakers. The closure phase in the articulation of the implosive sound is characterized by a straight line Q-C (channel 2) indicating the absence of air flow in either direction through the mouth. The large periodic fluctuations in the delimited segment R-S on the pressure tracing (channel 3) reflect vibrations of the vocal cords. A mid-line was drawn through these ripples by hand. The maximum supraglottal pressure (Psupra) was measured on this mid-line. The measurements of Psupra were made at the point of release of closure. Table 1 presents the Peak Psupra values of the syllable-initial implosives and plosives.

Table 1. Peak measurements of Supraglottal air pressure, in cm H20.

	b	ɓ	Difference
HW	7.5	-2	5.0
SS	6.5	-5	1.5

In the production of the implosives [ɓ], the vocal folds are brought together before the larynx is lowered. The vocal folds remain fairly tightly together throughout the articulation so that air will not pass through the glottis in such large volume as to destroy the negative

pressure necessary for an implosive. Lowering of the larynx obviously enlarges the supraglottal cavity behind the oral closure, which results in generating negative pressure inside the mouth. Since the larynx lowers only after the vocal folds are constricted, the lips brought together and velopharyngeal port closed, the rarefaction process in the expanding supraglottal cavities is not affected so much so that the air is sucked in when the outer closure is released. These results are typical of the other female as well.

Another interesting feature was noted consistently in the speech of both these two speakers. Implosives are produced with a relatively short closure duration. Table 2 presents the duration of voicing in both implosives and plosives. Note that the voicing of implosives ranges between 70% to 72% of the corresponding plosives.

Table 2. Duration of voicing, in milliseconds

	b	ɓ	Difference
HW	14	10	4.0
SS	12.5	9	3.5

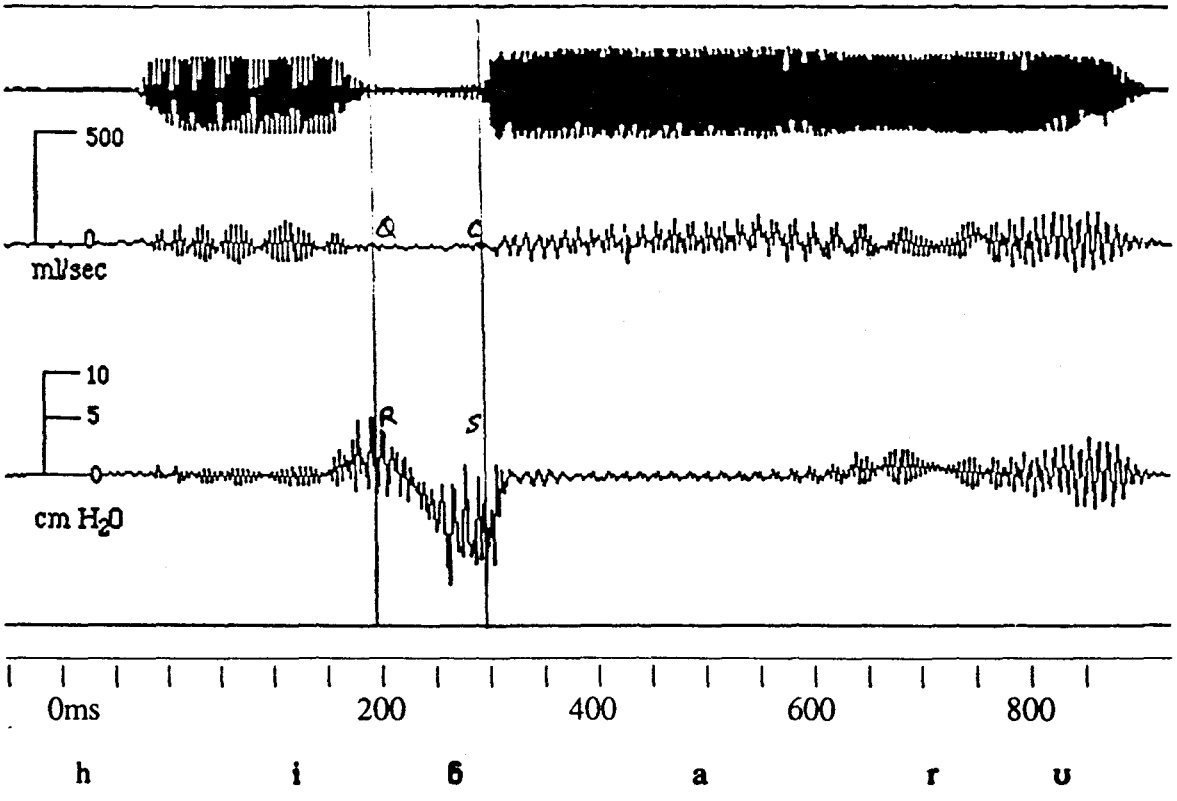


Figure 1. Acoustic waveform, air flow and intraoral pressure records of the word [ɓaru] "child".

The third speaker, however, produced implosives with a voiceless beginning of the closure. The closure displays highly aperiodic vibrations, whereas the voiced plosive [ɓ] in the speech of the third speaker has periodic voicing vibrations during the closure phase. So the voicelessness or aperiodicity in the case of the third speaker may serve to keep the implosives apart from the voiced plosives. However, the spectrograms (see Fig. 2) made from an independent recording of the same speaker clearly indicate the presence of vocal

fold activity throughout the period of closure in the articulation of implosives. Thus considerable variation was noted within the speech of the same speaker. I don't know how to solve this anomaly.

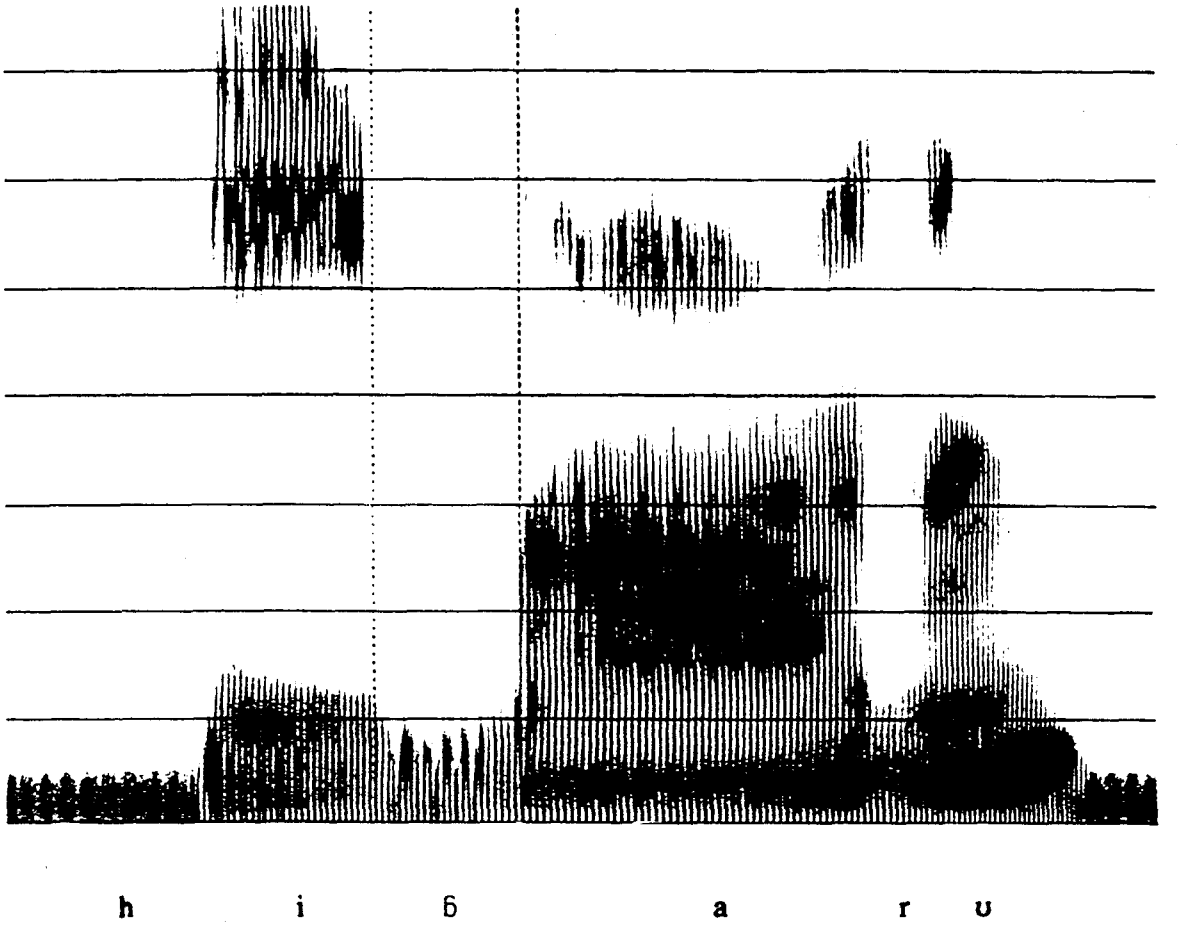


Figure 2. Spectrogram of the phrase [hi: ɓaru]

5. Discussion

The aerodynamic records show that in the case of 2 out 3 speakers the movement of the larynx occurs while the vocal cords are vibrating. This downward movement of the vibrating glottis enlarges the supraglottal cavity behind the closure. These vibrations are maintained by a small amount of lung air which is not of sufficient volume to destroy the partial vaccuum caused by the downward laryngeal movement and thus prevent the occurrence of suction pressure. The negative pressure ranging between -2 cm H_20 to -5cm H_20 was generated in the mouth. On separation of the articulators, the airflow was found to be ingressive. Thus the quantitative measurements, on the whole, confirm the results reported earlier by Nihalani (1986).

6. Conclusion

The preceding discussion makes it clear that Sindhi implosives show negative pressure most of the time in contrast to the implosives observed by Ladefoged in which negative pressure was produced only 8% of the time. Given that the implosives that Ladefoged has investigated are produced without any ingressive, unlike the Sindhi implosives, it is necessary to encode this information in the phonetic representation, whatever may be the phonological characterization of implosives.

In conclusion, I would like to propose the implosives are best characterized linguistically in terms of [+/- suction] feature rather that the greater degree of downward displacement of the larynx, which is just a physiological mechanism adopted in order to maintain the pressure difference for suction. The absence of suction in other types of implosives (such as in Hausa and other Nigerian languages) could be explained in terms of implementation of the linguistic features in terms of physiological mechanism. While giving a precise account of what makes a particular language sound the way it does, it is necessary to describe the phonetic properties of individual segments. Differences of each language will have to be described in terms of language-specific low level rules of "phonetic implementation", and these must form part of the phonological description of natural language. An understanding of the mapping processes from discrete and timeless phonological units to continuous articulatory and acoustic quantitative physical manisfestations is the important goal of linguistic phonetics.

Acknowledgement

I am grateful to Professor Peter Ladefoged for inviting me to join as a member of the Phonetics Lab group at UCLA to work on the phonetic structure of Indian languages when I was on sabbatical leave from the National University of Singapore.

References

Catford, J.C. (1939). On the classification of stop consonants. *Le Maitre Phonétique*, 3rd Series, 65, 2-5.

Ladefoged, P. (1964). *A phonetic study of West African languages*. Cambridge: Cambridge University Press.

Ladefoged, P. (1971). *Preliminaries to linguistic phonetics*. Chicago: University of Chicago Press.

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

Ladefoged, P. (1991). Computerized phonetic fieldwork. *UCLA Working Papers in Phonetics*, 78: 1- 6.

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

Nihalani, P. (1975). Velopharyngeal opening in the formation of voiced stops in Sindhi. *Phonetica*, 32: 98-102.

Nihalani, P. (1986). Phonetic implementation of implosives. *Language and Speech*, 29: 253-262.

Painter, C. (1978). Implosives, inherent pitch, tonogenesis and laryngeal mechanisms. *Journal of Phonetics*, 6: 249-274.

Pike, K. L. (1943). Phonetics, Ann Arbor: The University of Michigan Press.

Evidence for click features: Acoustic characteristics of Xhosa clicks

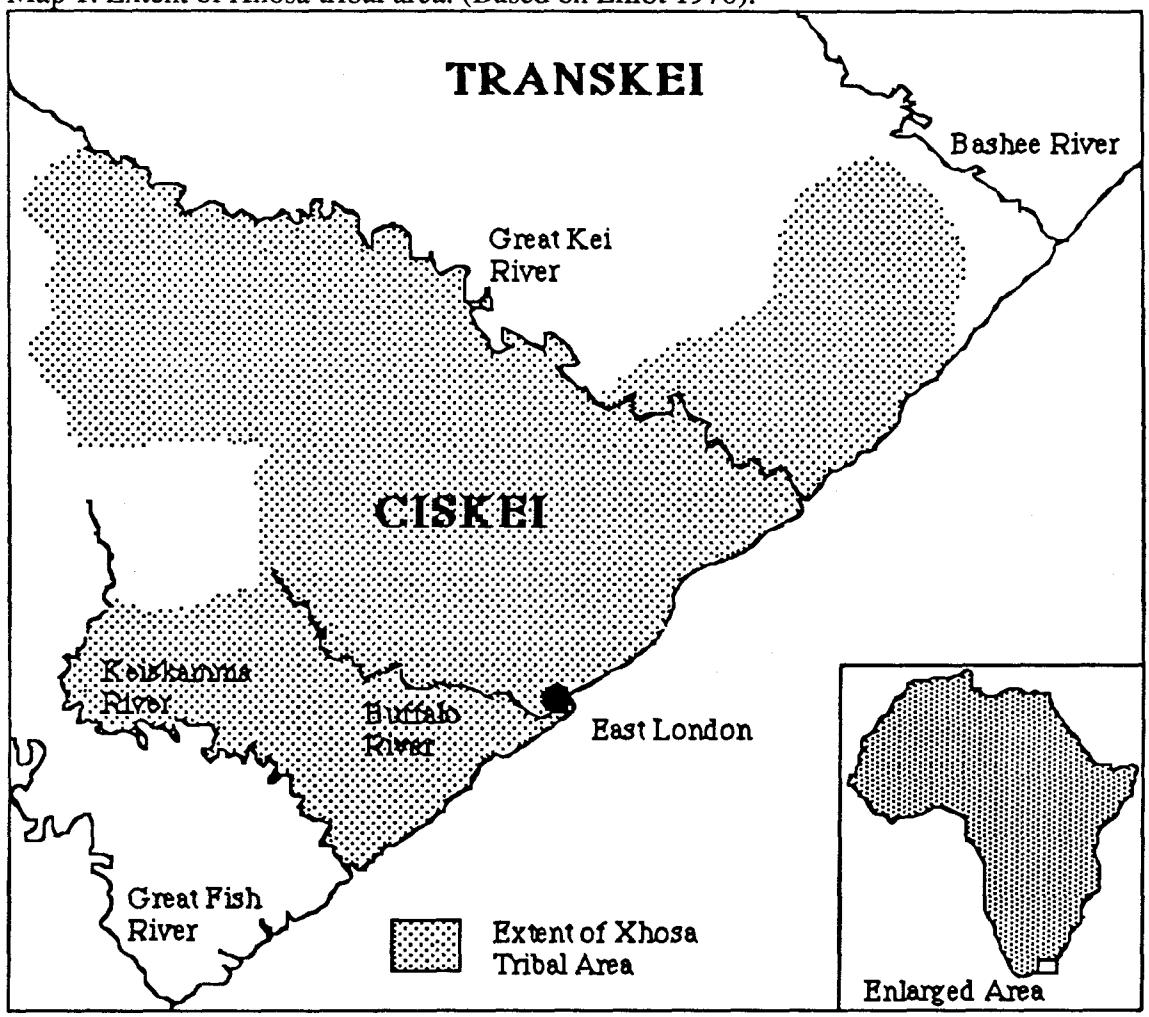
Bonny Sands

Introduction

Clicks are sounds which are typologically limited in distribution and complex in articulation. These two facts contribute to the perception of clicks as atypical consonants. It is precisely because of their rarity that clicks are of considerable linguistic interest. Their uniqueness causes them to be difficult to incorporate into classificatory systems. It also raises the question of whether the definitions of features as applied to other consonants apply equally well to clicks. This study will provide appropriate phonetic evidence to form a basis for the discussion of this topic.

Clicks are confined geographically to southern and eastern Africa. They occur in the southern Bantu languages Xhosa, IsiZulu, SiNdebele, SiSwati, and Gciriku, in all the Khoisan languages, including Sandawe and Hadza of Tanzania; and in Dahalo, a Cushitic language of Kenya. The non-Khoisan languages with clicks are assumed to have borrowed them from Khoisan languages.

Map 1: Extent of Xhosa tribal area. (Based on Elliot 1970).



The click sounds to be examined in this study are those in Xhosa, a Bantu language in the Nguni cluster whose other members include IsiZulu, SiNdebele and SiSwati. Xhosa is spoken by approximately six million people living primarily in the Ciskei and the Transkei in South Africa (Grimes 1988). (See Map 1). There are also Xhosa speakers in all of the major cities in South Africa.

The Nguni languages seem to have acquired clicks during the proto-Nguni stage or earlier, when there was a period of interaction with Khoisan groups in Natal (Finlayson 1987, Bill 1974). Contact with Transvaal Khoi or Cape Khoi speakers would have occured before 1000 A.D., (the date after which proto-Nguni is presumed to have begun to diverge; Ownby, 1981), and also after the divergence of proto-Nguni. Historical records show that the Xhosa had extensive contact with speakers of Eastern Cape Khoi dialects (Ownby 1985), and intermarriage and bilingualism are known to have been common among Nguni and Khoisan people in the time when the first Europeans arrived in South Africa (Lanham 1964). The integration of clicks in the Xhosa sound system is believed to have been accelerated by lexical borrowing and sound substitution due to the custom of hlonipha, the avoidance of certain words in order to show respect (Finlayson 1982). Hlonipha requires substitution of different words for the tabooed items.

The consonant inventory of Xhosa includes fifteen phonemic clicks. Clicks are produced with a velaric ingressive airstream, which involves a velar closure, and another closure made further forward in the mouth. Air in the pocket between these two closures is rarefied by lowering the central portion of the tongue so that when the anterior closure is released there is an ingressive flow of air. The velar closure is maintained until after the release of the anterior closure. Various names have been used for the types of anterior closures; the terms that will be used here are: dental, alveolopalatal, and alveolar lateral. The anterior closure can be referred to as the influx. The efflux, or click accompaniment, refers to modifications on the basic click such as nasalization, voicing and aspiration. Each of the click types in Xhosa may occur with one of five distinctive accompaniments. The general terms that can be used for distinguishing them are: voiceless unaspirated, voiceless aspirated, nasalized, breathy voiced, and nasalized with breathy voice. The set of clicks and the symbols that will be used for their transcription are seen in Table 1. The velar closure is represented in the click transcription for each accompaniment with a consonant symbol preceding the click symbol. Tone will not be marked.

Table 1. Symbols for Xhosa clicks

	Voiceless Unasp.	Voiceless Aspirated	Breathy Voiced	Nasalized	Nasalized Breathy
Dental	kǀ	kǀh	gǀɦ	ŋǀ	ŋǀɦ
Alveolopalatal	k!	k!h	g!ɦ	ŋ!	ŋ!ɦ
Lateral	kǁ	kǁh	gǁɦ	ŋǁ	ŋǁɦ

Clicks are not marginal sounds in Xhosa. They account for fifteen of the contrastive consonants in the language and occur in an estimated 38% of entries in the lexicon (Louw 1977b citing Finlayson). Often only Khoisan languages are considered to be "click languages", but this is clearly not the case.

1. Articulatory characteristics of Xhosa clicks

Before an acoustic study of Xhosa clicks can be carried out, it is necessary to provide a basis on which to analyze these sounds. That is, the articulatory characteristics of these sounds must be known. Although it is generally accepted that Xhosa has 15 distinct clicks, there is a great deal of variation in the descriptions of the types and accompaniments. While some of this may be attributable to speaker or dialect variation,

much of it is due to the fact that many of the descriptions were not done by specialists in phonetics. The terminology may not always be applied in a way which conforms to standard usage. The most reliable contemporary source of phonetic descriptions of clicks is Traill (1978, 1981, 1985). Although his work is primarily on !Xóõ, his works serve as a reference, as many of the sounds which he analyzes are analogous to those in Xhosa. The sounds described in this work are from Xhosa, unless stated otherwise.

A. Click types.

The click type in Xhosa typically refered to as a dental click is represented by the phonetic symbol [ǀ] ([ǀ] by the IPA prior to 1989), and orthographically as 'c'. It has been called a dental click in pedagogical materials and by various researchers (Jordan 1966, Louw 1977a, and Bill 1974) and also a dental affricate click (Louw et al. 1980, Beach 1938), as the release of the click is accompanied by fricative noise. It is described as being produced with the tongue tip pressed against the front teeth (Jordan 1966); or with the tip of the tongue against the upper front teeth and the sides of the tongue against the teeth and gums, as in Zulu, Bushman and Nama (Beach 1938). Like all clicks, this click must have the tongue raised on all sides to form a complete closure and create a cavity between the front and back closures. X-ray tracings of the [ǀ] click as produced by a !Xóõ speaker are shown in Figure 1. The tongue tip touches the teeth and the alveolar region, but before the release of the click, the area of contact is decreased and primarily dental.

The Xhosa dental click has also been described as an alveolar click, involving the blade of the tongue against the alveolae (Wilkes 1987). It may be that there are two distinct articulations for the [ǀ] click, with a laminal articulation being more commonly used by Xhosa in the Ciskei and the western parts of the Transkei, and an apical articulation used by most Xhosa speakers from the eastern districts of the Transkei (Louw 1977a). The laminal articulation is made with tongue tip pressed against the lower front teeth and the front part of the tongue blade being sucked away from the upper front teeth (Louw 1977a). The apical articulation, which is also used by Zulu speakers, is made with tongue tip being raised and sucked away from the upper front teeth. (Louw 1977a). Bearing these differences in mind, the [ǀ] click type is referred to as dental in this work.

The click type which is represented as 'q' in the orthography is often transcribed [ǃ] ([ǃ]by the IPA prior to 1989). The terminology used to describe the place of articulation of this click includes: mid-palatal (Jordan 1966), palatal alveolar (Wilkes 1987, Bill 1974), cerebral (Wilkes 1987, Beach 1938), palatal (Lanham cited in Davey 1975, Louw 1977a), alveolopalatal (Finlayson and Louw pers. comm. 1989), post-alveolar (Ladefoged 1982), retroflex (Beach 1938) and alveolar (Beach 1938). Descriptions of the role of the anterior part of the tongue in the production of this click type vary greatly. Jordan states that the front of the tongue is against the front palate (1966), while Beach states: "the tip of the tongue is placed against the alveoli, usually quite far back, though the exact position varies in the pronunciation of a single individual" (Beach 1938). Beach describes one Xhosa speaker who makes use of an alveolar or post-alveolar articulation, as well as what he terms a cerebral or retroflex articulation. Palatograms of the post-alveolar and cerebral articulations of this Xhosa speaker are shown in Figure 2a and b (after Beach 1938). The palatograms show very different articulations, the former showing contact on the back part of the alveolar ridge, and the latter showing contact which is much further back. Beach says that the palatogram of the cerebral articulation shows contact between the tongue tip and the palate. The same variation in place of articulation seems to be seen in Zulu, Xhosa, and Khoisan languages (Louw 1977a, Beach 1938). Beach states that various writers on Nama have described the [ǃ] type of click with the cerebral articulation as being made by "curling up the tip of the tongue against the roof of the palate" but that the usual place of release for this click in Nama is alveolar (generally "post-alveolar" or "palato-

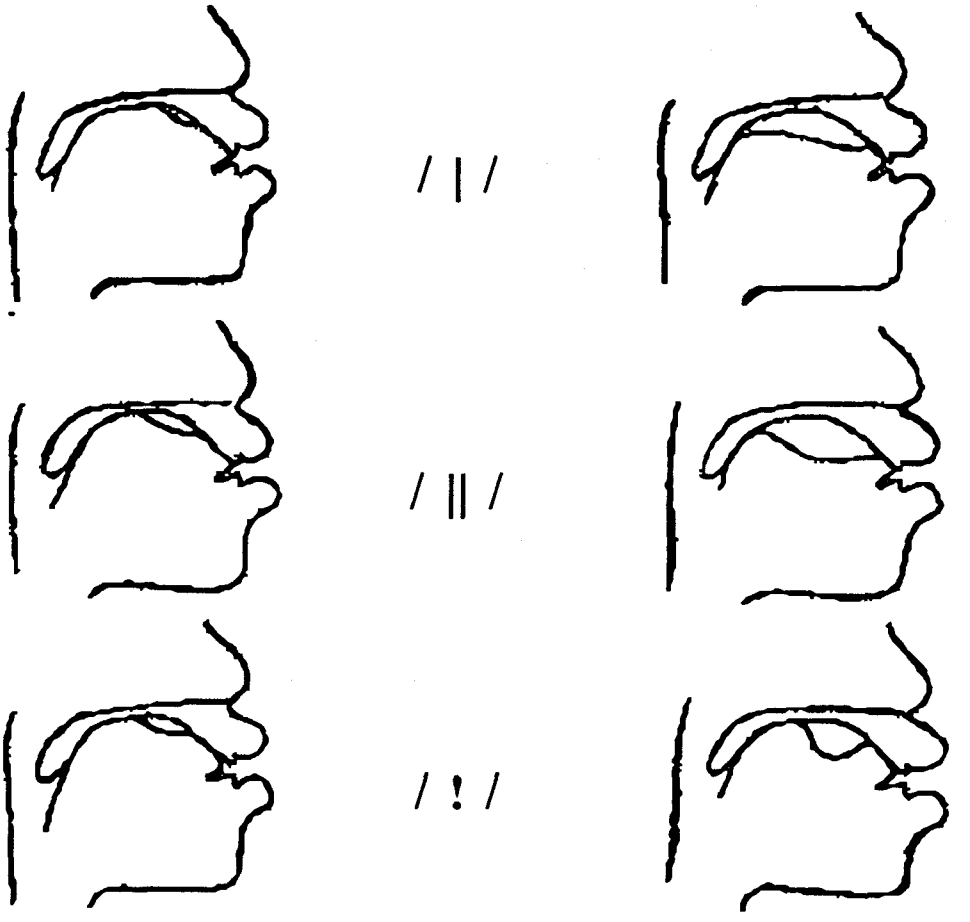


Figure 1: The dental, lateral and alveolopalatal nasal clicks in !Xóõ, based on x-ray tracings in Ladefoged and Traill (1984). The diagrams on the left are based on tracings from the frame showing the smallest cavity enclosed by the tongue during each click closure: those on the right are based on tracings from the frame immediately before the release of the closure.

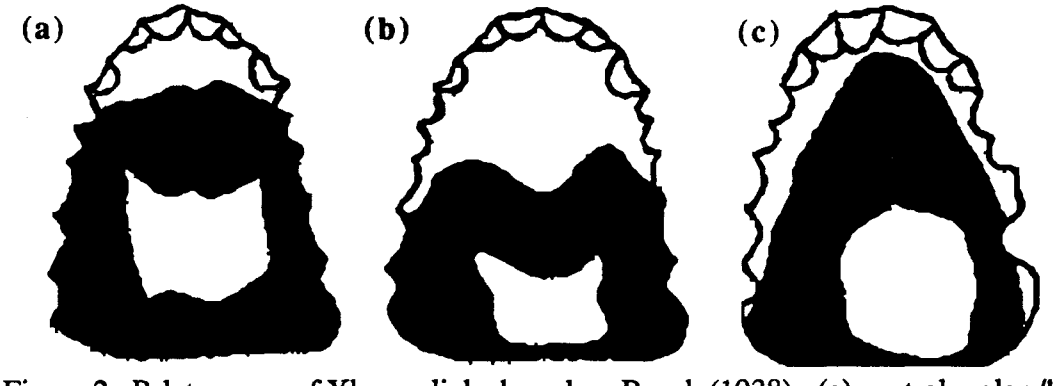


Figure 2: Palatograms of Xhosa clicks based on Beach (1938): (a) post-alveolar /!/, (b) "cerebral" /!/, (c) bilaterally released /ǁ/. Areas of contact between the tongue and the palate are indicated by dark shading.

alveolar") (Beach 1938). It is possible that the cerebral click that Beach describes is made with a sublaminal articulation. His palatogram is typical of those produced by sublaminal retroflex articulations in Indian languages (Dart 1991). One factor in the difficulty in describing the place of articulation for the [!] click is that the San people often have a flatter palate and lack an alveolar ridge. All these descriptions indicate an articulation made towards the back of an alveolar position or the front of the palate. The apical constriction is made with more pressure than for an apical pulmonic stop, and thus has a wider area of contact. In order to distinguish the place of articulation of this click from that of sounds made further forward on the alveolar ridge, or with a smaller area of contact, the Xhosa click which is orthographically 'q' is referred to as alveolopalatal in this work.

There are several possible explanations for the great deal of variation in the description of the [!] click. Some descriptions indicate that the [!] click is made with an alveolar closure initially, but with a palatal one immediately before release. A teaching guide to descriptive phonetics of Xhosa (Louw et al. 1980) states that the click is produced with the tip of the tongue against the palato-alveolar region. Another account of the production of the Xhosa [!] states that the tongue tip is against the alveolar position initially, and shifts to a palatal position as the tongue is pulled downwards (Wilkes 1987). Yet, contrary to this account, the X-ray tracings in Figure 1 of the [!] click in !Xóõ both show an alveolar place of contact.

The third click type found in Xhosa is termed lateral. It is represented orthographically as 'x' and is often transcribed [ɮ] ([ɬ] by the IPA prior to 1989). It has been described as being made with the tip of the tongue on the alveolar ridge and the sides of the tongue up against the side teeth, forming a completed "ring-like" closure, with the closure broken by pulling the tongue away at one side from the side teeth (Louw et al. 1980, Wilkes 1987). Beach states that it is made with the tongue tip placed against the alveoli and the sides of the tongue touching the teeth and gums (Beach 1938). While it has not been noticed in Xhosa that the point of contact of the tongue blade varies during the articulation of the lateral click, this can been seen in !Xóõ, in Figure 1. The tongue blade makes contact behind the alveolar ridge and the tip contacts the base of the teeth during the click, but the blade no longer touches the back part of the alveoli immediately before release of the closure.

It has been observed that in the Xhosa lateral click, the part of the side of the tongue which is pulled away is near the back of the oral cavity (Louw 1977a). In Khoi languages such as Korana and Griqua, it is the part of the side of the tongue which is more forward which is pulled away, resulting in a somewhat different sound (Louw 1977a). The Xhosa articulation results in a sound which is similar to that of the dental click in Khoi (Louw 1977a). Beach (1938), however, remarks that the lateral click is made the same way in Zulu, Xhosa, Bushman and Nama: "with the influx generally made over a considerable length of the side-edge (or edges) of the tongue and not at a single point" (Beach 1938). Beach notes that in Xhosa, the release of the click may be at one or both sides of the tongue, but more usually at only one side. A palatogram of a click in Xhosa which Beach reports to be bilaterally released is shown in Figure 2c. The click has an alveolar closure which does not touch the front teeth and covers the alveolar ridge. The sides of the tongue do touch the side teeth, but not as extensively as for the [!] clicks in Figure 2a. The bilateral release cannot be seen in a static palatogram, however.

The lateral click in Xhosa has been referred to as a lateral affricate (Louw et al. 1980, Beach 1938). The release of the click is usually accompanied by friction.

There is some degree of individual variation in the pronunciation of the lateral click. Beach (1938) claims that the click may be released on either side of the mouth or on both sides simultaneously, and that while the tip of the tongue is usually lowered slightly later than the side or sides of the tongue, some speakers were convinced that air entered over the tip of the tongue as well as at the side in their pronunciation of the lateral click.

It is apparent that the articulation of the Xhosa lateral click sound has not been completely described. It is not clear whether the click release is simultaneous for most points along a side of the tongue or not, and whether this varies across speakers. If we take into account the reports of Beach's speakers, it is also not clear whether the frication noise of the click is due to the release of the side or sides of the tongue from near the upper gums and teeth or to the release of the tip of the tongue from the alveolar ridge. Some of the frication may be due to the influx of air from between the upper and lower teeth.

There is a great deal of variation in the descriptions of the click articulations in the literature. Differences in terminology can account for some of the lack of agreement in the literature about how clicks are produced. Undoubtedly, speaker and dialect variation can also account for some of the differences in the descriptions. Careful descriptions of click articulations which take into account variation between speakers and dialects, and provide instrumental data illustrating the click articulations are called for. Unfortunately, an articulatory study of this sort is beyond the scope of this paper.

B. Click accompaniments.

The dental, alveolopalatal and lateral click types occur contrastively with an oral, voiceless, unaspirated accompaniment (Louw 1977a) which is also referred to as the radical form of the click (Bill 1974, Wilkes 1987). Clicks with this accompaniment are usually transcribed as [ʇ, !, ȸ], and written as 'c', 'q' and 'x'. They are transcribed here as [kʇ, k!, kȸ] to provide a transcription which is parallel with those of other accompaniments. Although the back closure must be released after the front closure, the symbol for the velar closure precedes the click symbol in order to avoid any implication that there is a noisy velar release. There seems to be some dialectal variation in the pronunciation of the voiceless unaspirated accompaniment. Jordan states that all unaspirated voiceless stops, including clicks, are ejective (Jordan 1966) and Louw notes that it is pronounced with a glottal stop in the speech of many Rharhabe Xhosa (Louw p. c. 1989). Beach makes no mention of these clicks being produced with a glottal stop, but states that these clicks in Xhosa and Zulu are pronounced with a practically silent velar release, unlike those in Nama (Beach 1938). The silent velar release would be consistent with a closed glottis.

Clicks also occur contrastively with a voiceless aspirated accompaniment, and are often transcribed as [ǀh, !h, ǁh], and written as 'ch', 'qh' and 'xh'. They are transcribed here as [kʇh, k!h, kȸh]. There is a period of aspiration following the release of both closures. This period is characterized by burst noise or stricture friction immediately at the release of the click and also by glottal friction. The noise of the glottal friction continues beyond the cessation of the burst noise and continues up to the onset of voicing. The puff of air in the aspirated stops of Xhosa may be more forceful than in English (Jordan 1966).

Only one contrastive oral click accompaniment involves some kind of voicing. Clicks with this accompaniment are often transcribed as [ʞg, !g, ȸg] (Traill 1985), written as 'gc', 'gq' and 'gx', and are transcribed here as [gʇfi, g!fi, gȸfi]. The quality of voice in this accompaniment has been described as oral and voiced (Jordan 1966), partially devoiced or devocalized (Bill 1974, Wilkes 1987), murmured or voiced (Louw et al. 1980), and as delayed breathy voice (Louw 1977a). Louw states that all murmured stops are initially voiceless to some degree and that often the stop is voiceless up to the point of release with the murmur being heard only on the following vowel (Louw 1977a). This accompaniment will be referred to as breathy voice.

Clicks occur contrastively with a plain nasal accompaniment in Xhosa. They are written 'nc', 'nq' and 'nx' and are often transcribed [ŋʞ, !ŋ, ȸŋ]. This accompaniment will be referred to as nasal and transcribed [ŋʇ, ŋ!, ŋȸ].

In addition to the plain nasal clicks, there is a click accompaniment represented in the orthography as 'ngc', 'ngq' and 'ngx', which is transcribed here [ŋʇfi, ŋ!fi, ŋȸfi]. Clicks with this accompaniment type have been referred to as being voiced nasal

compounds (Jordan 1966), nasalized murmured (Louw et al. 1980), and voiced aspirated nasalized (Bill 1974). Some accounts suggest that what is represented in the orthography as one click derives from more than one source and represents two distinct clicks for the Rharhabe or western speakers of Xhosa. The distinction between these two sounds in the speech of some Rharhabe Xhosa speakers is between what is called a prenasalized breathy voice click in which there is not strong evidence of breathy voice on the nasal component and a breathy voiced nasalized click in which the whole compound has breathy voice (Louw pers. comm. 1989). Both of these clicks are in contrast with the plain nasal click. The prenasalized breathy voiced clicks are derived from a synchronic rule of prenasalization. The nasalized clicks which are breathy voiced are the result of borrowings in which the nasal element was already present in a vowel following a click (Louw 1977a). When necessary, these separate pronunciations will be referred to as prenasalized with breathy voice and breathy voiced nasal. However, as most speakers do not make a distinction between clicks which are prenasalized with breathy voice and those which are breathy voiced nasal, clicks represented in the orthography as 'ngc', ngq' and 'ngx' will be usually referred to as nasalized breathy voiced clicks.

Another nasal compound which occurs in Xhosa is a prenasalized voiceless click, written 'nkc', 'nkq' and 'nkx' and transcribed here as [ŋkʇ, ŋk!, ŋkʖ]. This compound does not have the status of an accompaniment as it is a derived compound, a sequence of two segments, and not a single, contrastive segment. It is derived by the prenasalization of an aspirated click. The aspirated clicks pattern with other aspirated stops by becoming unaspirated when prenasalized (Louw 1977a).

2. The phonological system

A. Incorporation of clicks into the phonology.

We will start by looking at how the clicks were first incorporated into the sound system of Xhosa. The borrowing of clicks into proto-Nguni has been called a "startling invasion of the phonemic system of one language by another" (Haugen cited in Lanham 1964). This is an apt description of what happened to the phonemic system of Xhosa as fifteen clicks as well as six other consonants were added (Lanham 1964) to produce the current inventory, shown in Table 2.

Table 2: Xhosa consonant system based on Lanham (1964) and Davey (1975).

ɓ									
p	t	ts	tʃ	c	k	kx	kʇ	kʖ	k!
b	d		dʒ	ɟ	g		gʇɦ	gʖɦ	g!ɦ
ph	th	tsh	tʃh	ch	kh		kʇh	kʖh	k!h
f	s	ɬ	ʃ			x			
v	z	ɮ				ɣ			
	l			j	w				
m	n		ɲ		ŋ		ŋʇ	ŋʖ	ŋ!
mɦ			ɲɦ				ŋʇɦ	ŋʖɦ	ŋ!ɦ
h									
ɦ									

While many clicks were borrowed into Xhosa, they were not borrowed unsystematically. The Khoisan languages have much more extended click inventories than do Xhosa or any of the other Nguni languages. In some San languages such as !Xóõ, there are five click types; labial, dental, alveolar, alveolopalatal and lateral. The Khoi languages Naron and Nama share all these types except the labial. The bilabial click may not have been part of the segment inventories of the languages which influenced Nguni in that bilabial clicks were not borrowed. None of the Nguni languages has the click type

transcribed /ʗ/, which is variously described as alveolar, palatoalveolar and palatal. In Xhosa, this click type seems to have merged primarily with the dental click type: Nama /kʗuib-s/ 'willow' (cf. Xhosa /umgɪfiunube/ 'Cape willow') (Louw 1979). It may be that dental and /ʗ/ click were merged because Xhosa does not contrast two anterior coronal stops.

The number of accompaniment contrasts in Xhosa is reduced as compared with Khoisan systems. The voiceless unaspirated clicks in Xhosa often correspond to the clicks with glottal closure in Nama, as evidenced by related words such as Nama /kɪ'awe/ 'plan, scheme' (cf. /kɪeba/ 'plan, scheme') (Louw 1977a). They also correspond to clicks with aspirated and unaspirated voiceless accompaniments: Khoi /kǂhiri/ 'rush past goal' (cf. /kǂila/ 'hasten away') and Khoi /kʖo:/ 'fall off, die' (cf. /kʖwela/ 'grow thin') (Davey 1975). Clicks in Xhosa with voiced, aspirated and nasalized accompaniments typically correspond to clicks with like accompaniments in Khoi. Nasal clicks in Xhosa often correspond to glottalized clicks followed by a nasalized vowel: Khoi /ǃ'ũ/ 'delay in doing' (cf. /amaŋɪuŋɪuŋɪu/ 'unwise, imprudent') (Louw 1977a). As Xhosa does not have phonemically nasalized vowels, the nasalization of the vowel was re-interpreted as belonging to the click.

B. Synchronic status of the clicks in the phonology.

The synchronic status of the clicks in the phonological system can be seen by observing their behavior in phonological rules.

1. Prenasalization. Prenasalization occurs with adjective stems, deverbatives, and noun class 9 and 10 prefixes (Louw et al. 1980). The introduced nasal assimilates to the place of articulation of the following consonant. When the consonant is a click, the nasal is realized as a velar, indicating that clicks must be specified for the feature [+back], and possibly also [+high]. But in Xóõ, nasals assimilate to the anterior closure of a following click (Traill 1977).

The lateral click does not pattern with the lateral approximant when prenasalized. The introduced nasal deletes before /l/ but not before any of the lateral clicks (Davey 1975). The voiced and voiceless lateral fricatives become affricates after these nasals, which makes them more similar to the affricated lateral clicks. Although not all [+lateral] segments behave alike with respect to prenasalization, the [-sonorant] laterals behave similarly, in that nasals before these laterals do not delete. Clicks with nasal accompaniments behave like pulmonic nasals in that they cause the nasal prefix to delete: underlying /i+N+ŋɪaNgathi/ becomes surface [iŋɪaŋga:thi] (Davey 1975). There is no change in accompaniment for the nasal consonants. Nasal clicks, nasal pulmonic consonants and the lateral approximant all cause a preceding nasal to delete, but as the clicks are generally considered to be obstruents and therefore not [+sonorant], there is no straightforward way of grouping these together.

The anterior and velar closures of the clicks do not change when the click is prenasalized. With the velaric suction mechanism intact, the clicks do not become egressive. On the other hand, bilabial implosives become egressives when prenasalized: [ɓalisa] 'tell' (cf. [imbali] 'story') (Louw et al. 1980). The implosives can be characterized as [+constricted glottis], a feature which is lost when prenasalized, but there is no reason for clicks to be specified for this laryngeal property.

The nasal prefix affects the accompaniment of the following consonant, but not in precisely the same way for clicks as for pulmonic consonants. Voiceless unaspirated clicks become breathy voiced when they are given the class 9 prefix : [kɪɛŋɪfiɛlɛza] 'recite' (cf. [iŋ|fiɛŋɪfiɛlɛza]), [k!wɛŋga] 'tear', (cf. [iŋ!fiwɛŋga]) and [kʖak/a] 'flabbergast' (cf. [iŋʖfiak/i]) (Louw et al. 1980). This is the origin of the so-called prenasalized breathy voiced clicks in the Rharhabe dialect of Xhosa, and one of the origins of the nasalized breathy voice accompaniment. Voiceless unaspirated stops do not become breathy voiced when preceded by a nasal, but optionally become ejectives. The ejective variant of the

voiceless unaspirated stops can also be found in word-initial position, but is more common after a homorganic nasal (Davey 1975). Voiceless fricatives which are prenasalized do not become breathy voiced like the voiceless unaspirated clicks, but become ejective affricates (Davey 1975).

The voiceless unaspirated clicks and the voiceless unaspirated non-click consonants can be considered to have different specifications for laryngeal features. Khumalo (1981) considers the voiceless unaspirated non-click consonants to be ejectives, specified with the features [-spread glottis] and [+constricted glottis]. These features need not change when the consonants are prenasalized. If the voiceless unaspirated clicks are not [+constricted glottis], then when prenasalized they assimilate to the voicing of the preceding nasal. Being [+spread glottis], they will be breathy voiced. If all the voiceless unaspirated consonants are assumed to share the same laryngeal specifications, then the different effects of prenasalization could be accounted for by ascribing a certain feature to clicks such as [+click] or [+velaric]. Only [+click] consonants would become voiced and [+spread glottis] when following a nasal.

Clicks with the aspirated accompaniment pattern with pulmonic consonants with this accompaniment. Aspirated consonants become unaspirated when following a nasal: [kɪhwela] 'peel' (cf. [iŋkɪwɛla] 'shaving'), [k!huβa] 'go forward' (cf. [iŋk!uβɔ] 'progress'), [kʖhɛnts/a] 'dance' (cf. [iŋkʖɛnts/i] 'good dancer') (Louw et al. 1980). These derived click compounds are the so-called prenasalized voiceless clicks. Although sometimes mistakenly considered a separate accompaniment, these are derived compounds.

The voiced consonants which are usually described as having breathy or delayed voicing, and the breathy voiced clicks behave somewhat similarly when prenasalized. Prenasalization of a breathy voiced pulmonic consonant causes the consonant to no longer have any delay before the voicing starts (Louw, p.c.).

*2. **Tone**.* Clicks pattern with other consonants regarding tone. In Xhosa, all voiced obstruents which are usually described as having breathy voice or delayed voicing, and all breathy voiced nasals are depressors, or consonants that have a lowering effect on certain tones. The oral and nasalized breathy voiced clicks are included in the group of depressors (Davey 1975, Claughton 1983, Traill et al. 1987).

*3. **Phonological features**.* All of the front click closures can be specified as [+coronal] and the back closures as [+back]. The features [anterior], [lateral], and [distributed] distinguish the click types. Clicks may be separated into distributed (/ǁ/, /ǀ/) versus non-distributed (/!/); apical (/!/, /ǁ/) versus laminal (/ǀ/); anterior (/ǀ/, /ǁ/) versus non-anterior (/!/); and lateral (/ǁ/) versus non-lateral (/ǀ/, /!/). It is not clear, however, which features would be optimal for describing the clicks as there is little phonological evidence concerning the features indicating the place and aperture of the anterior constrictions of the clicks. It should be remembered that the lateral click is phonetically lateral at the release phase, but not during the closure phase. The dental and lateral clicks have affricated releases. These facts give rise to the possibility of different phonetic feature specifications for the release and closure phases.

The clicks can be said to share the same set of laryngeal features as the non-click consonants: [spread glottis], [constricted glottis], [voiced], with depressors being those consonants specified for the features [+spread glottis] and [+voiced]. The feature [nasal] is needed in addition to the laryngeal features to distinguish between accompaniments.

There is phonological evidence that the velar click closure must be specified with a feature, ie. [+back]. A preceding nasal will assimilate to the velar place of articulation of a following click. This is an argument against considering the velar closure as a secondary feature of the clicks in Xhosa. There are no restrictions on which vowel qualities may

follow clicks in Xhosa. In !Xóõ and other Khoisan languages, however, there is a constraint whereby only back vowels may follow back consonants, including clicks.

There is no phonological evidence that the dental and lateral clicks pattern with the affricates. Given that the dental and the lateral clicks have affricated releases, one might expect that they behave like contour segments, behaving like fricatives during the release, and like stops before the release. There is no phonological evidence that the dental and lateral clicks must be specified as [-continuant, +continuant].

Evidence from !Xóõ (Traill 1985) shows that the alveolar, and possibly the dental click pattern with the other dental consonants in conditioning a rule of /a/-raising. There is no clear phonological evidence in Xhosa that clicks which are [+coronal] pattern with pulmonic coronals.

There is no evidence for the lateral click patterning with other laterals, though laterality is clearly a salient phonetic detail of the release portion of these consonants. The lateral click can be distinguished from the other clicks without the use of the feature [lateral], and there is no evidence indicating that the closure phase of the click must be [-lateral] and the release phase [+lateral].

There is much that is unknown about the feature values of clicks. The phonological evidence does not show the status of features such as [continuant], [anterior], [coronal], [lateral] and [distributed] with respect to clicks. An invariant acoustic property which is argued to exist for some feature or place of articulation should also exist for clicks sharing that feature or place of articulation. A feature such as [coronal] should have the same definition for pulmonic stops and fricatives and clicks. Unfortunately, the work on invariance has largely ignored clicks in the determination of acoustic properties of features. It is not clear whether the phonetic properties of the features are the same for clicks as they are for pulmonic consonants. This study of acoustic characteristics of clicks provides a basis for the discussion of this issue.

3. Previous instrumental phonetic studies of clicks

While numerous descriptions have been made of the articulatory characteristics of Xhosa clicks, little quantitative descriptive work has been done. With the exception of the palatograms in Beach's study, there exists virtually no published instrumental data on the Xhosa clicks. An acoustic study can provide insights into the articulation of click types and accompaniments. This is particularly important given the dearth of physiological data available on the Xhosa clicks, and the confusion surrounding their precise articulations. However, we can draw on the excellent and extensive data that are available for certain Khoisan languages. Acoustic studies of clicks have been done on !Xóõ (Traill, 1985), Naron (Kagaya 1978), and Nama (Ladefoged and Traill 1984).

Traill's (1985) study on !Xóõ includes palatograms and x-ray tracings of click articulations as well as wide and narrow band spectrograms. He also reported aerodynamic data including oral and nasal flow, and oral and pharyngeal pressure, along with waveforms of the clicks, and conducted a fiberoptic examination of the larynx.

Traill provides a fine articulatory description of the click accompaniments in !Xóõ, and also has a few remarks about the acoustic characteristics of the click types. He distinguishes the unaffricated alveolar and alveolopalatal clicks from the dental, lateral and bilabial clicks which are released with friction and are longer. He notes that the lateral click has energy at lower frequencies than the dental click, that the bilabial click has more low frequency energy than the dental click and also has a more random distribution of noise, and that the alveolar click has more high frequency energy than the alveolopalatal click.

Traill provides acoustic data on vowel quality in !Xóõ and gives one example of the interaction between clicks and vowels. He remarks that the vowel /a/ is pronounced as "either a lowered-high and slightly centralized vowel [ɐ], or as a raised-mid central [ɜ]" when preceded by a "dental" consonant and followed by /i/ or /n/. The alveolar click /ʗ/ is

mentioned as counting as a dental consonant for the purposes of this rule. No details are provided as to whether dental and alveolar clicks of all accompaniments follow this rule.

Kagaya (1978) looks at the frequency characteristics of the dental, alveolar, alveolopalatal and alveolar lateral clicks for one male speaker of Naron. Using wide-band spectrograms, he notes the frequency ranges of click noise for each of these clicks, dividing the clicks into a strong and weak intensity part. He also notes the frequency bands within the regions of greatest intensity. He characterizes all clicks as having a wide frequency range and strong intensity compared with pulmonic fricatives. He finds that the alveolar and lateral clicks have a wide frequency range compared with the other two clicks. The dental and alveolar clicks have the dominant intensity band of the click noise in a relatively high frequency region while the alveolopalatal and lateral clicks have the dominant intensity band in the low frequency region, with the affricated clicks showing a downward shift in the frequency of the strong intensity band with time. He finds no indication that the frequency distribution of the strong intensity band of the click burst is influenced by either the tone of the syllable or the click accompaniment.

Kagaya also examines the duration of burst noise of the clicks, concluding that the dental and lateral clicks have a long burst duration compared to the alveolar and alveolopalatal clicks, although he claims that the difference in duration is generally smaller than that between affricates and stops. He also states that "a bigger back cavity capacity may be required for the long click noise type than for the short click noise type."

In order to compare his acoustic measures with frequencies predicted by acoustic modeling, Kagaya estimates the vocal tract shape for each of the four click types. He estimates the lengths of the front cavity, which is the distance between the lips and the front closure, the back cavity, and the constriction or anterior closure. His findings indicate that the shorter the length of the front cavity, or the longer the length of the constriction, the higher the resonant frequencies. The volume of the chamber between the two closures is estimated as is the minimum pressure in this chamber. The dental and lateral clicks are estimated to have the greatest negative pressure. Kagaya argues that central frequency is in proportion to the pressure difference between the front and back cavity, so that the dental and lateral clicks have higher central frequencies than the alveolar or alveolopalatal clicks. However, this would predict that nasal clicks would have different spectra from oral ones.

Ladefoged and Traill's (1984) study on clicks in Nama and !Xóõ has measurements of pharyngeal pressure, oral and nasal flow, waveforms, wide-band spectrograms and x-ray tracings. They note that the dental and lateral clicks are more affricated than either the alveolar or alveolopalatal click types. They note that the aspirated clicks tend to be affricated after the release of the velar closure, and that "the back of the tongue moves away from the velum more slowly in the aspirated than in the unaspirated clicks." They note that the lateral click has slightly lower mean frequency than the dental, and the /ʗ/ click has a higher mean frequency than the /!/ click. They also address the nature of the /ʗ/ click, noting that the click is palatal at the time of release.

This study expands on previous studies by considering spectral characteristics of clicks before different vowels before a number of both male and female speakers. The question of what coarticulatory relations exist between clicks and their neighboring vowels is addressed. Temporal characteristics of clicks such as total duration, VOT, degree of frication, and duration of closure are also examined.

4. Acoustic analysis

The data analyzed in this study were taken from a recording, kindly supplied by Professors Louw and Finlayson, of four male and four female Xhosa speakers saying words containing each of the 15 phonemic clicks before each of the vowels /i,e,a,o,u/. There was one repetition for each click and vowel combination. For four of the speakers, there were no tokens of a breathy voiced lateral click before /i/. For one male speaker, the

set of tokens was incomplete. Recordings were done on a reel to reel recorder and then transferred to cassette tape. The frequency response characteristics of the original recording are unknown.

Wideband spectrograms were made of two male and two female speakers of the fifteen contrastive clicks before each of the five vowels /i,e,a,o,u/. Spectrograms were made using the Kay Sonagraph with speech sampled at 40,960 Hz. Frequencies range up to 16,000 Hz, which is a wider range than typically used for consonants. This range was used because clicks often have energy present in the very high frequencies. An analysis bandwidth of 600 Hz was used.

In order to examine temporal characteristics of clicks, speech was low pass filtered at 10,000 Hz and digitized at a Sample Rate of 11,128 Hz, with an 8 bit resolution. Waveforms of the nine oral clicks were made for all eight speakers before the five vowels /i,e,a,o,u/. Waveforms were made using the Macintosh SoundEdit program. LPC spectral analysis was carried out using the Macintosh UCLA/Uppsala Soundwave program.

A. Temporal characteristics

The temporal characteristics of Xhosa clicks examined are: duration of frication noise, click closure duration, voice onset time, and total click duration.

1. Affrication. Qualitative judgements concerning the duration of friction noise of each of the three click types were made using spectrograms and waveforms.

As seen in Figure 3, dental clicks are made with an affricated release, shown by the high degree of friction in the period before the onset of the vowel. The amplitude of the burst and friction is typically low relative to the amplitude of the following vowel. The amplitude is usually weakest at the onset and then remains relatively constant till the onset of voicing. The burst typically has a gradual onset, low amplitude relative to that of the vowel, and noise continues throughout the period of aspiration.

As seen in Figure 4, the alveolopalatal clicks are not affricated or made with a prolonged release. This is in sharp contrast with the dental clicks. The noise at the release of the click typically has a very sharp onset and high amplitude relative to that of the following vowel. The amplitude decreases very quickly after this and there is very little noise before the beginning of the vowel that would indicate affrication. The concentration of energy in the first 10 to 20 ms of the release phase indicates that the release is made quickly.

As seen in Figure 5, the waveforms of the lateral clicks are similar in many ways to those of the dental clicks. There is a great deal of noise throughout the duration of the release up to the onset of voicing which indicates affrication. It is of higher amplitude in the initial part of the release, but is persistent throughout. The lateral click, like the dental clicks is produced with a long constriction area and with a constriction release of a long duration. The lateral clicks tend to be noisier and have sharper onsets than dental clicks.

The lateral and the dental clicks in Xhosa have affricated releases while the alveolopalatal click does not. This concurs with data on the duration of click noise for clicks in Naron. Kagaya (1978) shows the dental and lateral clicks as having a much larger average click noise duration than either the alveolar or alveolopalatal clicks. In his data, the dental and lateral click noise durations average 30 ms, with a range of 20-40 ms, while the unaffricated clicks average only 10 ms, with a range of 5-20 ms.

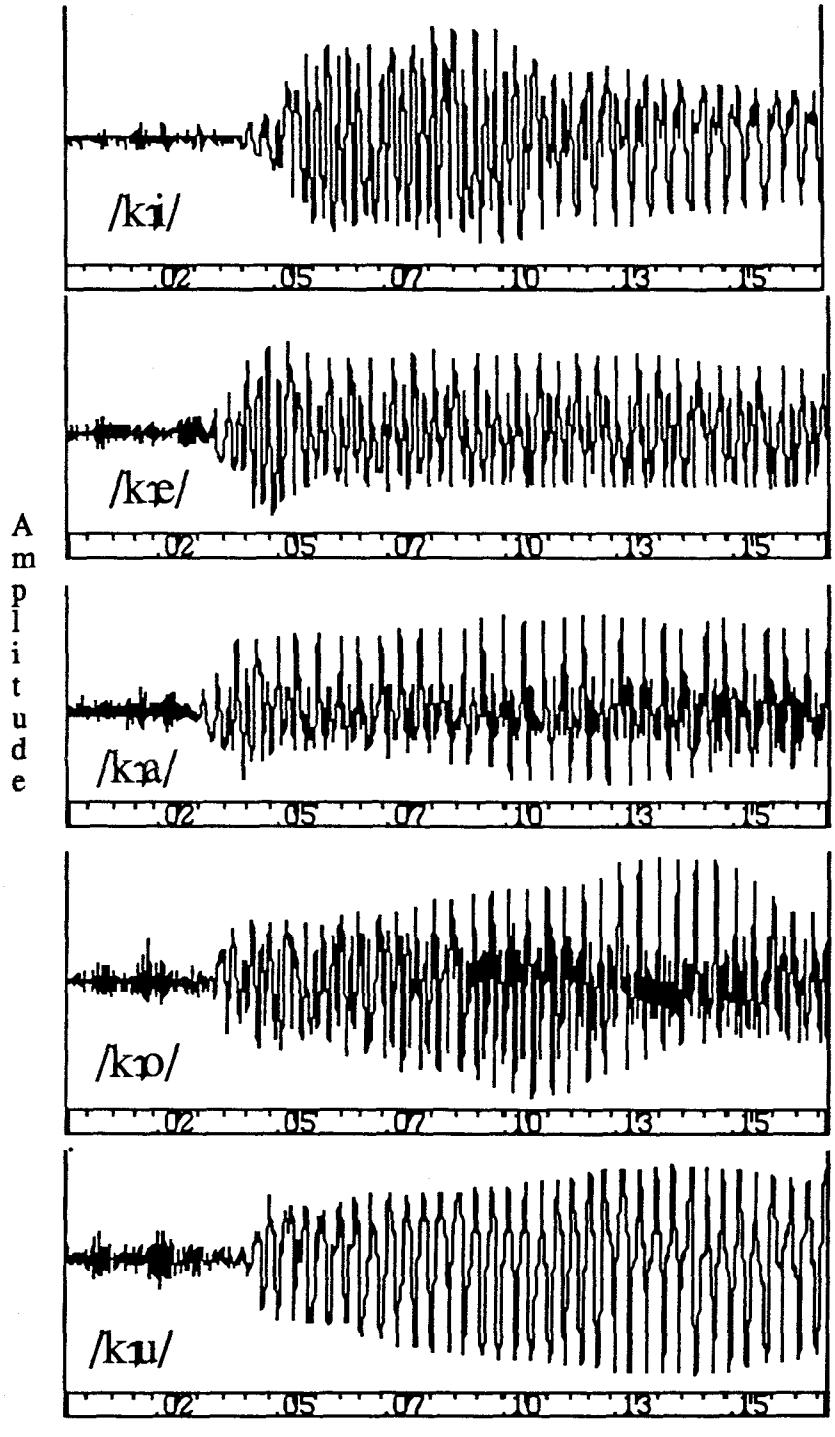


Figure 3: Waveforms of voiceless unaspirated dental clicks of one female Xhosa speaker, before each of the five vowels.

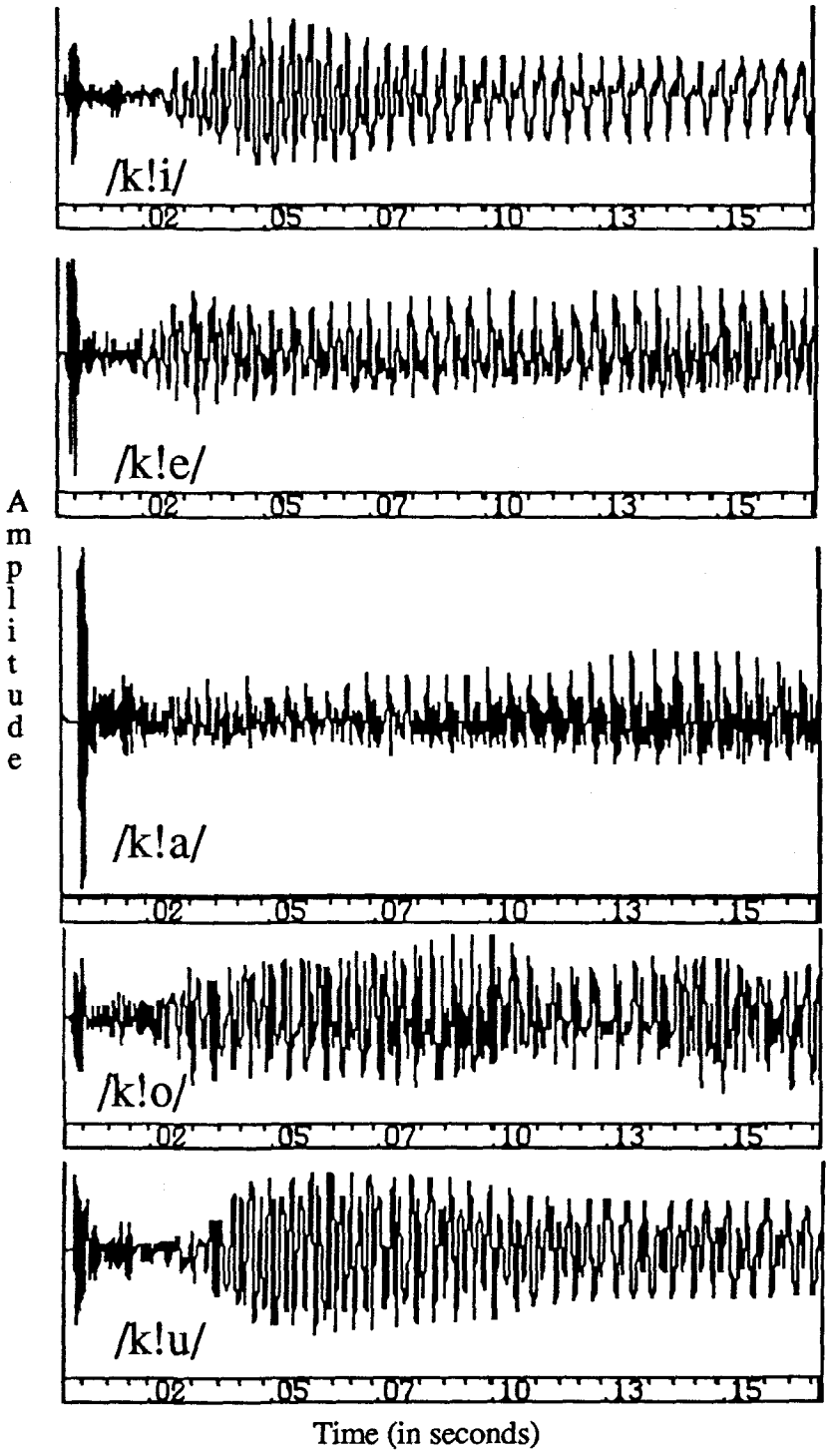


Figure 4: Waveforms of voiceless unaspirated alveolopalatal clicks of one female Xhosa speaker, before each of the five vowels. The waveform of the click before /i/ includes part of a lateral approximant following the vowel.

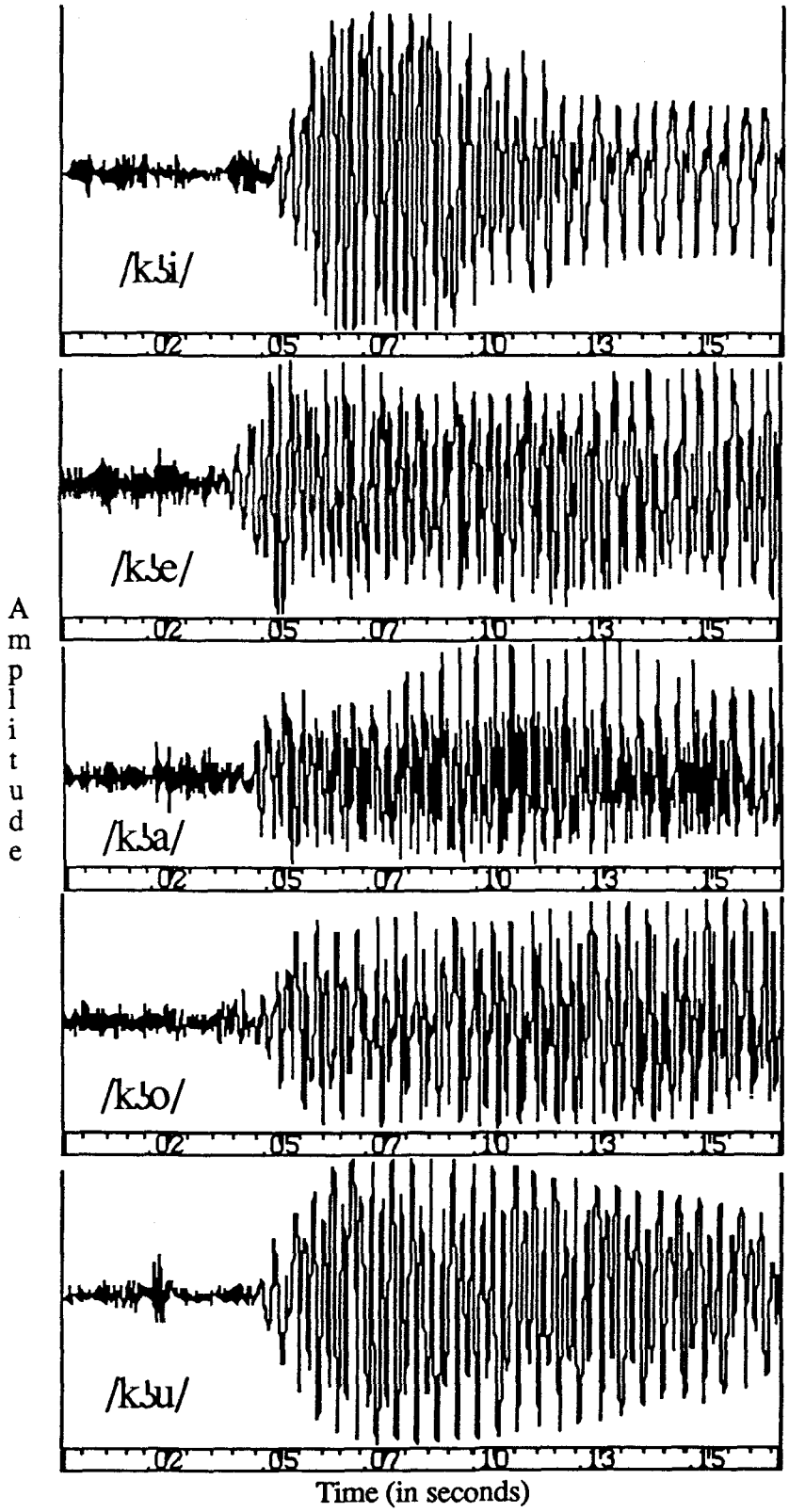


Figure 5: Waveforms of voiceless unaspirated lateral clicks of one female Xhosa speaker, before each of the five vowels.

2. Closure duration. Using waveforms, measurements were taken from the end of regular voicing of the vowel preceding the click to the beginning of the burst associated with the release of the front articulation of the click. These measurements are referred to as click closure durations. They essentially measure the duration from the back click closure to the release of the front click closure. Closure durations of the dental, alveolopalatal and lateral voiceless unaspirated clicks were measured for seven speakers before each of the five vowels /i,e,a,o,u/.

As can be seen in Table 3, the click closure duration is longest for the alveolopalatal clicks, followed by the lateral clicks. The dental clicks have the shortest closure duration. Using the Fisher PLSD and Scheffe F-test, the difference in closure duration between only the dental and alveolopalatal clicks was found to be significant at 99%.

Table 3: Mean closure durations with standard deviations for the three click types with the voiceless unaspirated accompaniment, for one token before each of the vowels /i, e, a, o, u/ for 7 speakers.

Dental		Alveolopalatal		Lateral	
mean	s.d.	mean	s.d.	mean	s.d.
127.7	31.4	159.4	45.5	143.5	44.9

3. VOT. The measure of the duration from the onset of the click burst to the onset of the following vowel will be referred to here as voice onset time, or VOT. Measurements were made using waveforms, as in Figure 6. VOT measurements were made for words containing clicks made with each of the three oral accompaniments before each of the vowels /i, e, a, o, u/ for 7 speakers. If there was a period of irregular voicing before regular voicing began, measurements were made up to the point of the irregular voicing.

The lateral and dental click types might be expected to have longer voice onset times than the alveolopalatal clicks, on the grounds that they are made with long constrictions in the vocal tract and are usually described as having affricated releases. Mean VOTs and standard deviations can be seen in Table 4.

Table 4: Mean VOT and standard deviations for the three click types with the oral accompaniments of one token before each of the vowels /i, e, a, o, u/ for 7 speakers. (The mean for the lateral breathy voiced clicks were before /e, a, o, u/.)

	Voiceless unaspirated		Voiceless aspirated		Breathy voiced	
	mean	s.d.	mean	s.d.	mean	s.d.
dental	50	20	133	28	47	19
alveolopalatal	45	22	131	30	26	9
lateral	63	32	137	39	42	10

The aspirated clicks obviously have much longer VOT than either the voiceless unaspirated or the breathy voiced clicks. There is a tendency for the voiceless unaspirated clicks to have a longer VOT than the breathy voiced clicks, although this effect cannot be seen for all the click types. The breathy voiced clicks are typically not voiced during the closure phase. There was found to be a significant difference between the voiceless unaspirated and breathy voiced accompaniments for the alveolopalatal and lateral click types ($p<.01$), using a two-tail, paired t-test. There was no significant difference in VOT between these accompaniments for the dental click type.

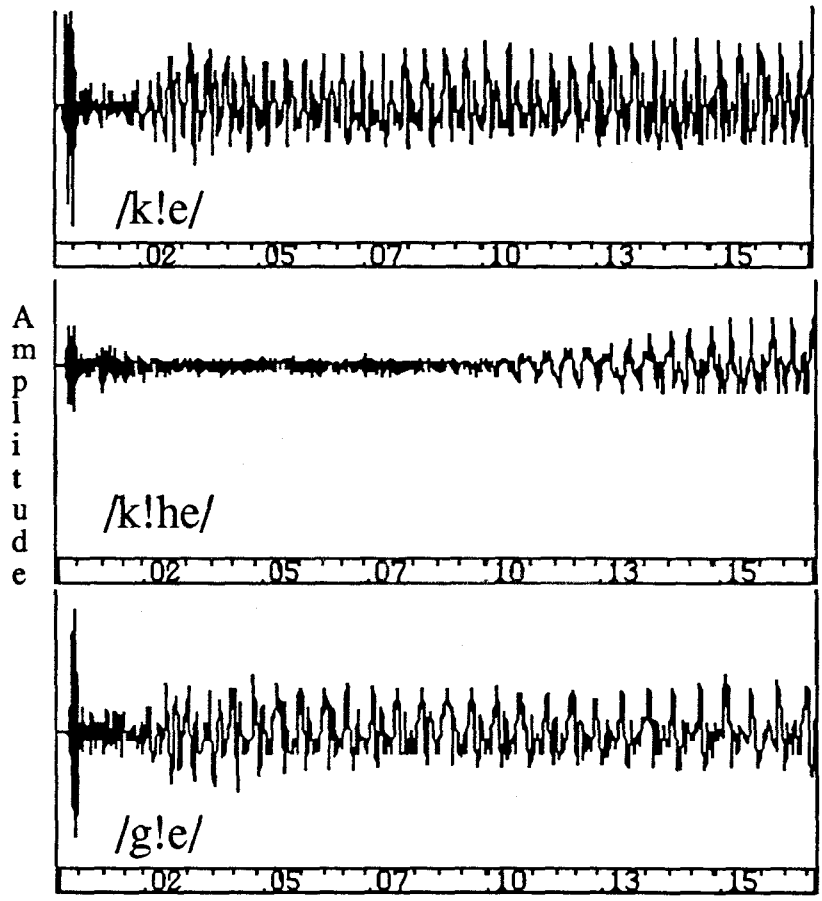


Figure 6: Waveforms of the voiceless unaspirated, voiceless aspirated, and breathy voiced alveolopalatal clicks of one female speaker of Xhosa before the vowel /e/. The waveform of the voiceless unaspirated click also includes some of a lateral approximant following the vowel.

All of the click types with the voiceless aspirated accompaniment have similar VOTs. This would seem to be because the period of aspiration extends beyond the period of frication, obliterating any effect of duration of frication on VOT. Any effect of click type on VOT would be seen more strongly by examining clicks with the voiceless unaspirated and breathy voiced accompaniments.

There is a distinct tendency for the more highly fricated dental and lateral clicks to have longer VOTs than the alveolopalatal clicks, although this difference is not significant across all accompaniments. All comparisons were made using paired, two-tail t-tests. There were no significant differences between click types found for the aspirated clicks. The dental and lateral click types were both found to have significantly different VOTs from the alveolopalatal clicks ($p<.01$). The voiceless unaspirated lateral clicks were found to have significantly different VOT from the alveolopalatal clicks with the same accompaniment ($p<.01$). The voiceless unaspirated dental clicks had significantly different VOT from the voiceless unaspirated alveolopalatal clicks, but only at a level of probability $<.05$. For the voiceless unaspirated clicks only, there was found to be a significant

difference in VOT between the dental and lateral click types ($p<.05$). So for this accompaniment, dental clicks tend to have shorter VOT than lateral clicks.

4. Click duration. Total click durations were obtained by adding the click closure durations to the VOTs. As click closure durations were only measured for the voiceless unaspirated clicks, total click durations are calculated only for these. However, Kagaya (1978) found no evidence of click duration being influenced by the accompaniment of the click sound in Naron, or the tone, for that matter.

As seen in Table 5, the mean total duration of the voiceless unaspirated alveolopalatal and lateral clicks were very similar, while the dental clicks have the shortest mean duration. A preliminary study of duration in !Xóõ based on 60 tokens of 4 male speakers, also found that voiceless unaspirated lateral clicks were the longest of the three click types. They averaged 181 ms, while the voiceless unaspirated alveolopalatal and lateral clicks averaged 168 and 171 ms, respectively (Traill 1991).

Table 5: Mean total durations with standard deviations for the three click types with the voiceless unaspirated accompaniment, for one token before each of the vowels /i, e, a, o, u/ for 7 Xhosa speakers.

Dental		Alveolopalatal		Lateral	
mean	s.d.	mean	s.d.	mean	s.d.
179	29	203	43	207	43

B. Spectral characteristics

1. Click burst spectra. Spectra were made on the Kay DSP Sonagraph using power spectra of speech sampled at 40,960 Hz, and filtered and analyzed up to 16,000 Hz, which is a wider range than typically used for analyses of consonants. This range was used because clicks often have energy present in the very high frequencies, although a range extending up to 8000 Hz would have been sufficient. As the frequency response characteristics of the original recording are unknown, and as the recordings were played back into the Sonagraph using a tape recording with an upper limit of about 14,000 Hz, frequencies above this are not reliably analyzed. The power spectra of the click bursts of eight speakers for the voiceless aspirated, voiceless unaspirated and breathy voiced clicks before the vowels /i/, /e/ and /a/ were analyzed, giving 72 tokens of each click type, with the exception of the lateral clicks. For four speakers, there were no tokens of breathy voiced lateral clicks before /i/, and for one male speaker, there were no tokens of the voiceless unaspirated lateral clicks before /i, e/. The spectra in this study were made using a 25 ms window starting at the release of the consonant. Since the back click closure (cf. IV.C.1) is released shortly after the release of the front closure, some noise from the back release may be included in the 25ms window used.

As seen in Figures 7 and 8, the dental clicks have a diffuse spectrum, and a great deal of high frequency energy. Dental clicks typically have energy present from 0 to 9000 Hz, and energy of lesser amplitude present up to 16,000 Hz. The amplitude level of the dental clicks is lower than that of the lateral or the alveolopalatal clicks. All of the dental clicks before /i, e, a/ have a diffuse spectrum, and further, half have a falling shape.

Although the dental clicks all have energy present in a wide frequency range, there are certain frequencies which contain higher amplitude energy. Of the clicks before /i, e, a/, approximately 53% have the greatest amplitude energy present in a band approximately 1000 Hz wide which is centered between 1000 and 1200 Hz. A further 10% have prominent energy centered within this range. All of the dental clicks have some energy within this range, but for some tokens it is not prominent compared with energy in neighboring frequencies. There were no consistent differences due to the vowels /i, e, a/.

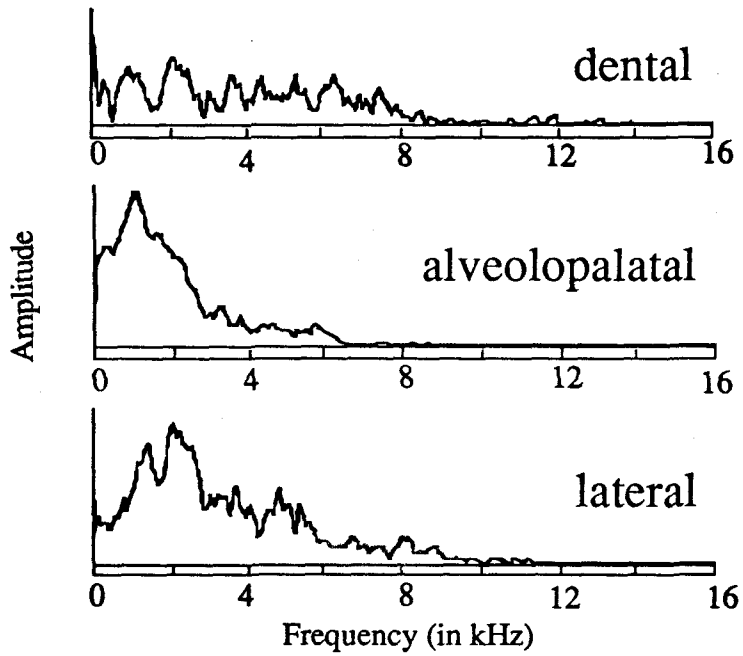


Figure 7: Mean spectra of the dental, alveolopalatal and lateral voiceless unaspirated clicks before the vowels /i,e,a/ for two male Xhosa speakers. Each curve is the mean of six spectra. Spectra were averaged over only two speakers so dissimilar tokens were not averaged together.

For the alveolopalatal clicks, as seen in Figures 7 and 9, there is typically one main band of energy in the low frequency range. Above this band of energy, the amplitude drops off rather quickly. Energy generally continues up to 5000 Hz, but it is typically of

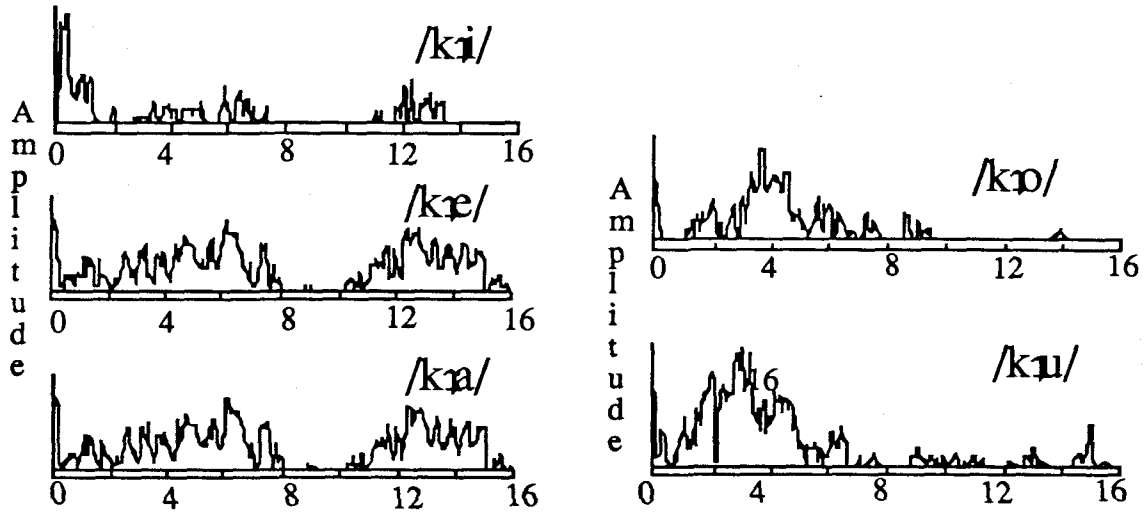


Figure 8: Spectra of voiceless unaspirated dental clicks of one female speaker of Xhosa, before each of the five vowels.

very low amplitude relative to that in the lower frequencies. Energy may also continue up to 9000 Hz or beyond, but is of much lower amplitude. For clicks before /i, e, a/, the

center of the prominent energy band occurred within the range of 1300 to 1400 Hz for 39% of the tokens, in the range of 1400 to 1700 Hz for 29% and in the range of 1000 to 1300 Hz for 19%. There is some tendency for the frequency range of this band to be higher for the female speakers than for the males. The remainder of tokens either had the most prominent energy above 1700 Hz, or had prominent energy in a band much wider than 1000 Hz. In many of the tokens, energy above the prominent low frequency band also occurs in distinct peaks. For alveolopalatal clicks before /i, e, a/, there is a peak centered between 2400 and 2700 Hz for 26% of the tokens. An additional 20% have a prominence centered between 2000 and 2300 Hz. Some 68% of tokens have fairly prominent energy between 3800 and 4800 Hz. It may be that all alveolopalatal clicks have audible energy in this range which does not appear in spectra designed to show the prominent peaks, as it is of such low amplitude relative to the low frequency band of energy. There were no consistent differences due to the vowels /i, e, a/.

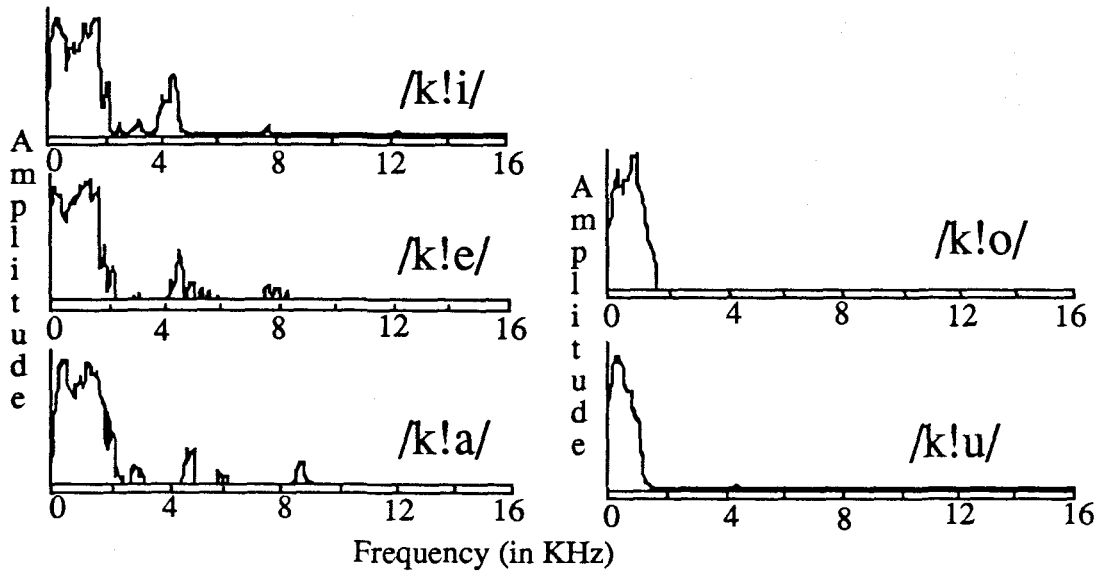


Figure 9: Spectra of voiceless unaspirated alveolopalatal clicks of one female speaker of Xhosa, before each of the five vowels.

The lateral click bursts, as seen in Figures 7 and 10, have a diffuse spectrum in the frequency range of 0 to 5000 Hz. They often have energy up to 8000 Hz or beyond, but it is typically of lower amplitude relative to energy below 5000 Hz. The energy in the spectrum is greatest in three broad frequency ranges, which are lower for male speakers than for female speakers. The spectrum can be delineated into prominence regions presumably because of zeros caused by side cavities to the lateral channel of airflow. No consistent effects of the vowels /i, e, a/ were seen.

The lateral click spectra of the four female speakers have high amplitudes in the low frequency regions. All tokens had energy present between 1000 and 2000 Hz. In 88% of the tokens, energy with either the highest or the second highest amplitude in the spectrum occured in this range. In 35% of the tokens, energy in this region was as high or higher in amplitude than energy elsewhere in the spectrum. Energy in the low range was lower for the male speakers, ranging from 900 to 1900 Hz. In only 25% of the cases did the highest amplitude energy in the spectrum occur in this range.

The highest amplitude energy in the spectra of 68% of the female speakers occurred within a second wide region of energy, which ranged from 2100 to 4000 Hz. For 12% of speakers there was energy present in this region which was equal in amplitude to that in other regions. The second region of the male speakers was lower in frequency. The

majority of tokens, 72%, had the highest amplitude peaks between 2000 and 2900 Hz, while another 9% had peaks in this range equal in amplitude to peaks in other regions.

The third main region of energy in the lateral click spectrum is between 4000 and 4800 Hz for the female speakers and between 3000-4500 Hz for the male speakers. Peaks in these regions for 56% of female and 47% of male speakers had the second greatest amplitude in the spectrum.

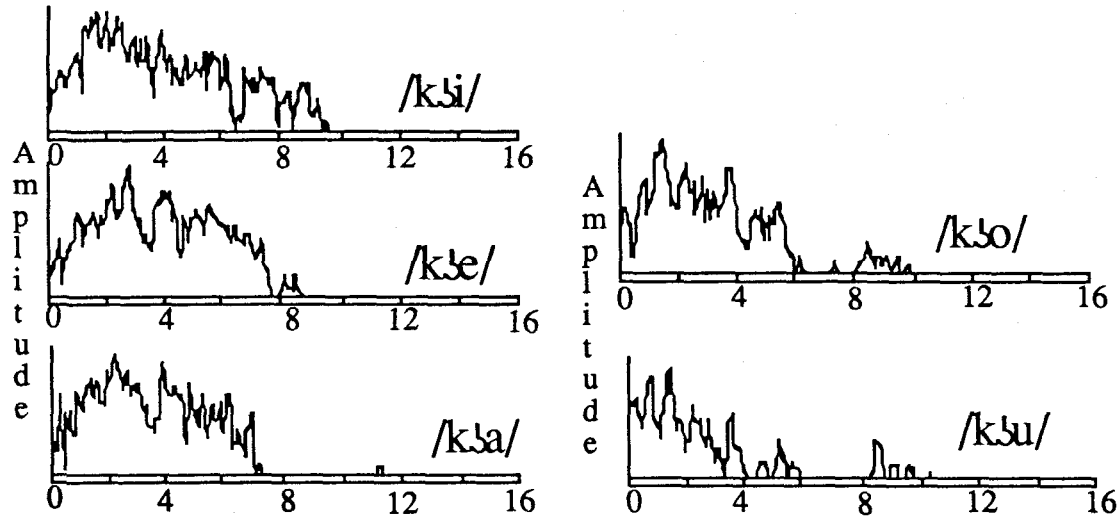


Figure 10: Spectra of voiceless unaspirated lateral clicks of one female speaker of Xhosa, before each of the five vowels.

The relevance of the spectral characteristics of the click bursts to phonetic feature theory is discussed in IV.D.1.

*2. **Coarticulation.*** The degree of coarticulation between a stop consonant and a following vowel can be examined by comparing the spectral pattern of the consonant burst before different vowels. If vowel position is anticipated in the consonant, the burst will show modifications that echo some characteristics of the vowels.

There were no consistent differences between the power spectra of the clicks before the vowels /i, e, a/. In particular, no consistent effect of the high front vowel /i/ is seen. This is the vowel which commonly causes extensive coarticulation effects with other consonants. It is possible that certain vowel effects were not seen as the wide frequency range used may have compressed information in the lower frequency region. However, notable differences between the power spectra of the clicks preceding /i, e, a/ and those preceding the rounded vowels /o/ and /u/ were clearly seen.

This effect is an expected result of anticipation of the rounding of these vowels. This effect can be seen more clearly for the dental clicks if we compare spectra of clicks before the unrounded vowels /i,e,a/ with those before the rounded vowels /o/ and /u/, as in Figure 8. Those preceding the rounded vowels show a concentration of energy in the lower spectral region resulting from attenuation of amplitudes in the higher frequency range. The energy in the lower frequency band is greater in amplitude relative to the energy above 10,000 Hz for the clicks before rounded vowels. In particular, they show a peak in energy around 3000-4000 Hz. All of the dental clicks before /o, u/ have a diffuse spectrum with the exception of the aspirated dental clicks of one male speaker, which have a prominence around 2800 Hz.

The effect of a rounded vowel on a preceding click can also be seen for the alveolopalatal clicks, as in Figure 9. As for the dental clicks, those preceding the rounded vowels show a concentration of energy in the lower spectral region, that is, below 2000 Hz. Energy occurs in a narrower band for the clicks preceding rounded vowels.

As seen in Figure 10, for lateral clicks preceding rounded vowels, there is less energy in the higher end of the spectrum. The peak of energy which occurs below 2000 Hz tends to be at a lower frequency for clicks preceding a rounded vowel.

C. Characteristics of the back click closure

1. Duration. The back click closure must be maintained from the beginning of the click, which is approximately concurrent with the cessation of normal voicing for a preceding vowel, until after the release of the front closure. This timing of the back closure is necessary for the production of a velaric airstream mechanism. Where the release of this closure occurs is more difficult to determine. There are no published X-ray tracings showing the release of the front and the back closures in separate frames. A velar release burst is not easily seen in waveforms and spectrograms of the clicks. This is apparently because the release of the back closure occurs quite soon after the release of the front closure. The noise associated with the release of the back closure is obscured by the noise of the front release. In the alveolopalatal clicks which have a less fricated burst, it is sometimes possible to see a small burst occuring approximately 5 ms after the initial burst. This is not visible in all tokens, however. Of the three alveolopalatal clicks in Figure 6, a second burst can be seen in the aspirated click.

Aerodynamic records of Nama clicks made by Ladefoged and Traill (1984) provide articulatory evidence that the back closure is released shortly after the front release. A drop in pharyngeal pressure is associated with the release of the back closure. For voiceless unaspirated clicks, pharyngeal pressure is approximately 7-8 cm H_20 just prior to release of the front click closure, with pharyngeal pressure dropping off sharply after the release. Pharyngeal pressure also decreases at this time for the voiceless aspirated clicks, but the decrease in pressure is more gradual.

2. Coarticulation

a). Effect of vowel on back closure. Other than knowing that the back closure of clicks in Xhosa is typically described as velar, not much about its precise place of articulation is known. It may be that the back closure of the click is retracted during the production of the click so that the actual release is post-velar, and it may be that the back closure varies due to the type of front closure. Additionally, it is not clear whether back click closures are influenced by vowel context. Specifically, we may ask whether clicks before front vowels are made with fronted velar closures. That is, does the back click closure behave like a pulmonic velar stop, which is known to vary according to vowel context across languages?

In order to investigate the nature of the back click closure, a perception study was carried out. Words containing dental, lateral and alveolopalatal voiceless unaspirated clicks before the vowels /i, e, a, o, u/ from 7 Xhosa speakers were digitized at 20 kHz on the Kay Sonagraph. The noise of the release of the front click closure was removed by using the Kay Sonagraph editing capabilities in an attempt to isolate the release of the back closure. The resulting edited tokens were re-recorded onto tape. The stimuli contained the period of aspiration following the burst of the front click closure and the following vowel, with an inter-stimulus interval of approximately four seconds. A listening tape was prepared in which these stimuli were randomized and grouped by speaker, in a forced choice task. Each token appeared only once in the presentation.

The listeners, six phoneticians, were asked to classify the place of articulation of the prevocalic segment into one of three categories; coronal, velar, or uvular. The reason for including the coronal response was to reflect the assumption that for the affricated clicks,

information about the front articulation is included in the noise after the initial click release burst.

As listeners were given a forced choice, a confidence rating scale was also used. Listeners were asked to rate the confidence of each categorization on a scale of one to three. When analyzing the results, the response category selected by the listener was coded with the confidence rating assigned (an integer between one and three). The categories which were not selected were assigned a zero response, as were those which were assigned a one rating. In this way, a token which was rated velar with a confidence rating of two would be reported as a 66% adjusted velar response, and a token with a confidence rating of three would receive a 100% adjusted response.

Listeners tended to avoid the coronal label for the alveolopalatal clicks, but this can be considered to reflect characteristics of the front closure rather than the back closure. The affricated clicks were labeled coronal more often than the alveolopalatal clicks.

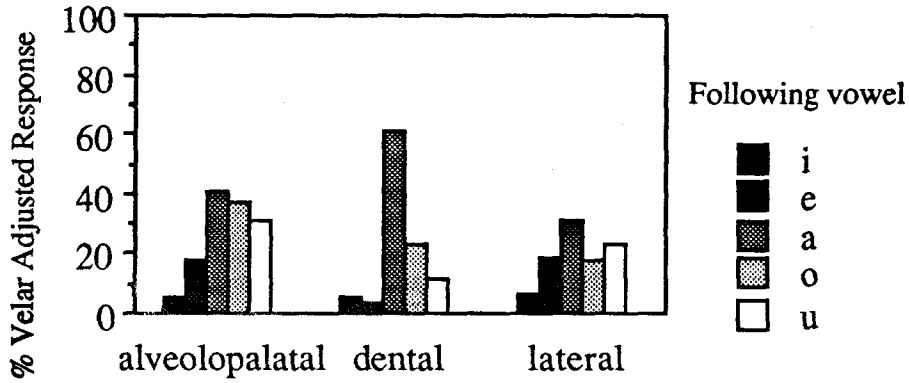


Figure 11: Percent velar adjusted responses for 6 subjects.

In Figure 11, which represents the percent adjusted velar responses grouped by vowel, it can be seen that the stimuli containing /a/ were most confidently labeled velar. A velar interpretation is also strongly favored for stimuli taken from the alveolopalatal clicks. The strongest velar response was for the group of dental clicks with the vowel /a/.

Stimuli containing the front vowels /i/ and /e/ tended to favor a uvular interpretation, as shown in Figure 12. This is particularly the case for the dental clicks. Obviously, this percept is not attributable to coarticulation, since front vowels would be expected to produce a more front back closure. It could instead reflect an articulatory dissimilation, by which the back release of clicks is actually more retracted with front vowels than with back vowels. Or it may be a case of perceptual compensation. In the context of front vowels, listeners expect to hear fronted velars, so they interpret non-fronted velars to be uvular.

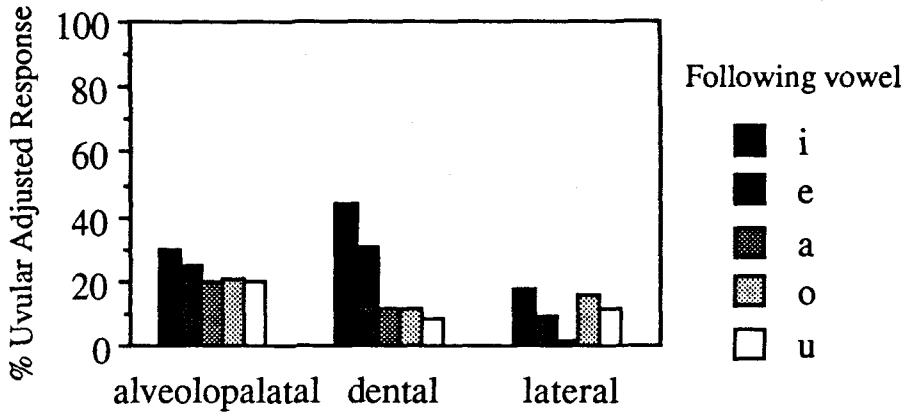


Figure 12: Percent uvular adjusted response scores for 6 subjects.

As tokens with /i/ produced the highest percentage of uvular responses and those with /a/ the highest percentage of velar responses, formant transitions following these vowels were compared. We found the highest uvular response rate for dentals with /i/, and a low rate for laterals. But as seen in Figure 13, the difference between the onset of vowels following dental and lateral clicks was marginal.

The dental clicks show marginally lower F2 and F3 than the lateral clicks, but as seen in Figure 14, dental clicks also have the lowest F2 for /a/, where the response was overwhelmingly velar. Given that the formant transitions are not significantly different, as

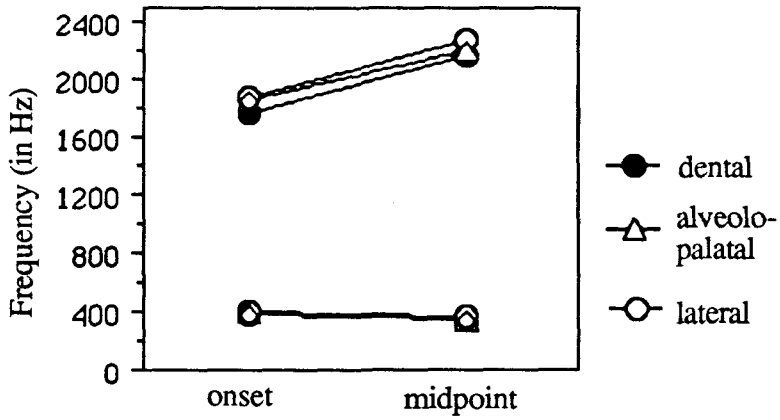


Figure 13: F1 and F2 of /i/ averaged over 7 speakers for the three contrastive oral clicks. Formant measurements were taken at the vowel onset, and once during the steady state portion of the vowel.

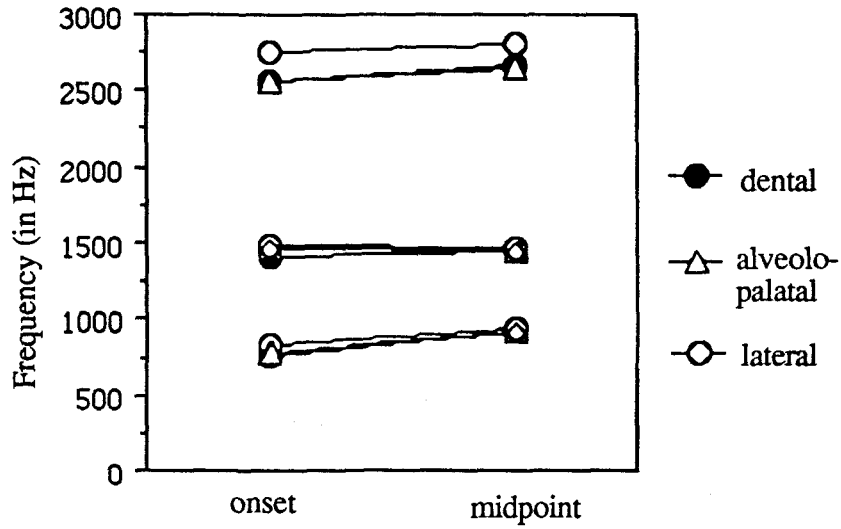


Figure 14: F1, F2, F3 of /a/ averaged over 7 speakers.before the 3 contrastive oral clicks. Formants were measured at the vowel onset and once during the steady state portion of the vowel.

will be presented in section IV.C.2.b below, it seems more likely that these uvular responses arise not from disimilation, but from failure to coarticulate with the following front vowel. Listeners interpret transitions from non-fronted velars to be uvular. So there is no perceptual evidence indicating that the type of back click closure varies due to the type of front click closure or to that of a following vowel.

b). Effect of click type on vowel. There is no discussion in the literature of the effect of a click on the articulation of a neighboring vowel in Xhosa. We might expect some information about click type to be contained in the vowel onset transitions, as this is often considered to be the primary cue for place of articulation of pulmonic stops. Alternatively, vowels following clicks might be expected to have onset transitions which are indicative of a dorsal consonant since the release of the back click closure follows the release of the front one, and because F2 depends primarily on tongue body position.

In order to examine the effect of click type on vowel quality, measurements were made of formants at the onset of the vowel, and during the steady state portion of the vowel for the first three formants of the vowels /i, e, a, o, u/ after dental, lateral and alveolopalatal voiceless unaspirated clicks, for 7 Xhosa speakers were analyzed. Formants were measured using LPC analysis on a Macintosh computer using UCLA/Uppsala Soundwave; an analysis window of approximately 23 ms (256 points) was used. Formants were measured in the middle of the vowel and at the onset of voicing, and averaged. To measure the formants at the beginning of the vowel, a window was positioned so that it was completely in the voiced portion of the vowel. No significant differences in the vowel formant onsets were found for vowel by front click closure, using a repeated measures ANOVA with the factors vowel and click type. There is no significant acoustic evidence indicating that the vowel formant onset transitions vary due to type of front click closure. As seen in Figure 14, the formant transitions do not show the convergence of F2 and F3 as is typical of velar consonants. These transitions indicate that a remnant of both dorsal and coronal gestures is present at the onset of the vowel following a click.

D. Comparison between clicks and non-clicks. Clicks form a class of sounds which have distinct acoustic and articulatory properties. Nonetheless, it is reasonable to assume that there are certain characteristics shared between clicks and consonants made with other airstream mechanisms, particularly given the fact that clicks were readily integrated into the phonemic systems of Bantu languages.

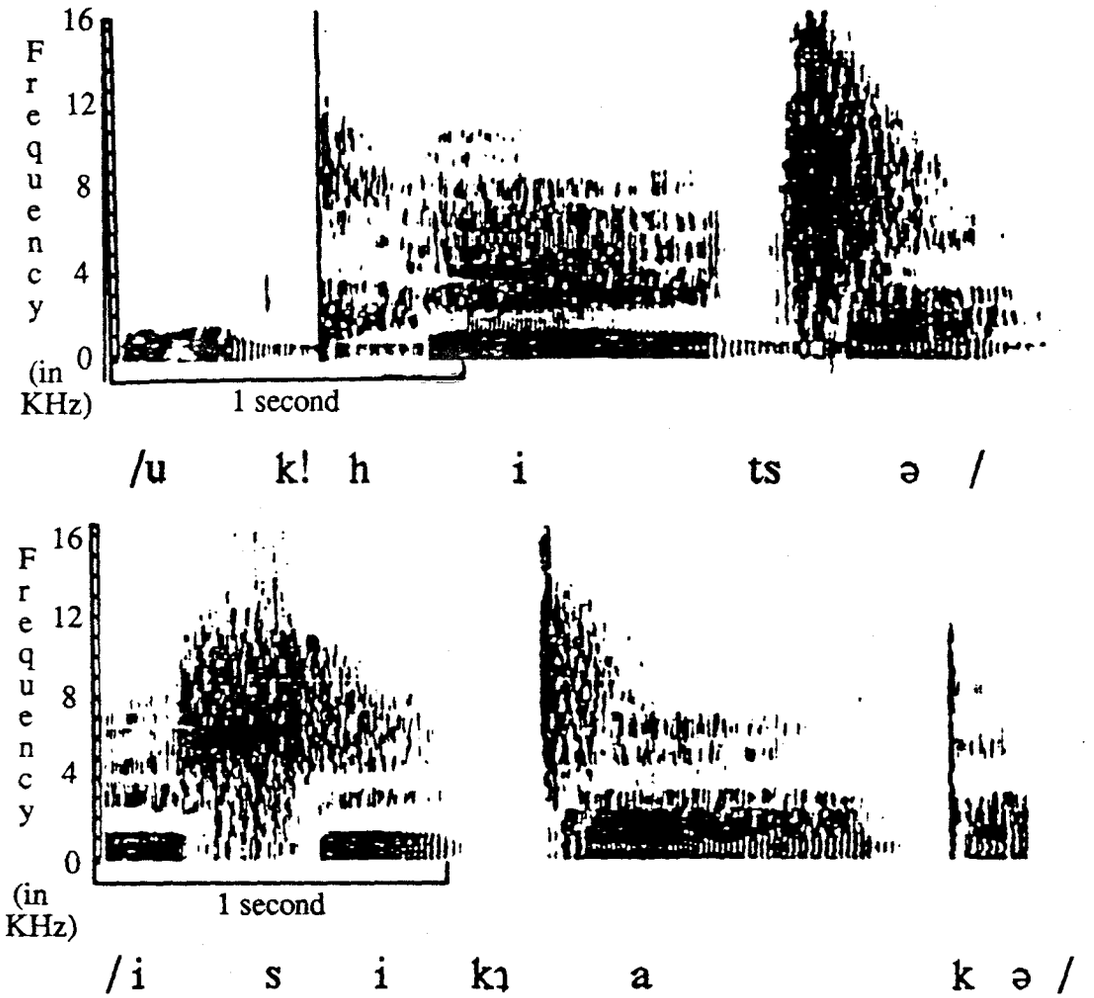


Figure 15: Wideband spectrograms of words containing voiceless aspirated alveolopalatal and voiceless unaspirated dental clicks uttered by a male Xhosa speaker.

1. Frequency. Although there are many spectral similarities between clicks and pulmonic consonants, clicks in particular are characterized by a wide frequency range. In Figures 15 and 16, it can be seen that clicks have energy in a wide frequency range as compared with pulmonic stops [k] and [t^h]. Coronal fricatives and affricates have little high amplitude energy in the low frequency regions compared with clicks. It has been noted by Kagaya (1978) that clicks are characterized by a wide frequency range (almost 0 to 8 kHz) and strong intensity compared with other fricatives caused by the eggresive airstream.

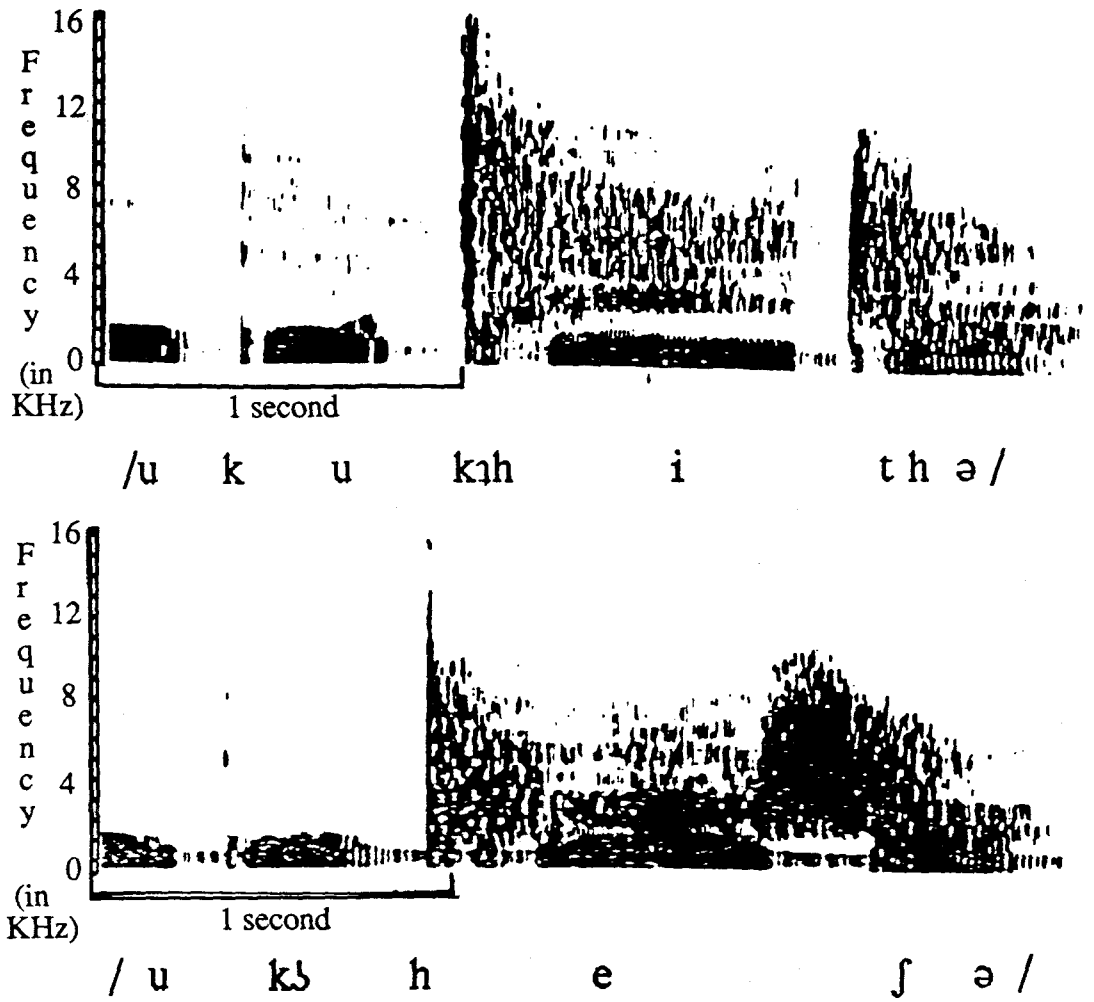


Figure 16: Wideband spectrograms of words containing voiceless aspirated dental and lateral clicks uttered by a male Xhosa speaker.

The bursts of dentals and other anterior stops are often considered to have a spectral shape which is diffuse and are also characterized by a predominance of high frequency energy (Stevens 1989, Blumstein 1986, Blumstein & Stevens 1979). These properties are also found with dental clicks. Dart (1991) found that in French, apical dental burst transient spectra have more energy above 3500 Hz than laminal dentals, and that the laminal dentals, like the dental clicks, have an essentially flat spectrum.

Coronal consonants which are not anterior are usually characterized as having a compact spectral shape (Blumstein 1986). The alveolopalatal clicks also have a compact spectral shape. It can be seen that the alveolopalatal clicks do pattern with other consonants in having a compact spectrum with a predominance of energy in the lower frequencies. Pulmonic retroflex stops have the effect of bringing together F3 and F4 (Stevens 1989), because of their large sublingual cavity. The presence of a similar cavity can be assumed for apical alveolopalatal clicks.

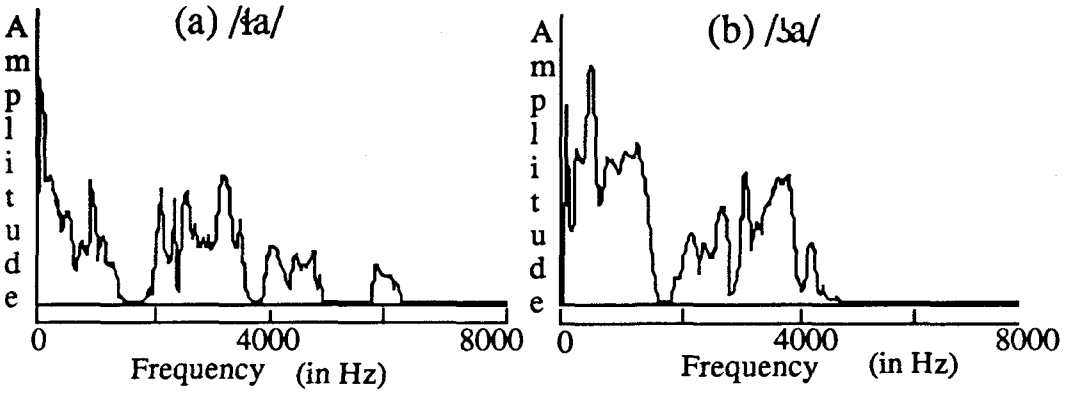


Figure 17: Power spectra of Zulu laterals, (a) voiceless lateral fricative, (b) voiceless lateral click

The lateral click bursts share certain acoustic characteristics with other laterals. Lateral clicks and lateral approximants typically have a raised F3, with energy at 3000 Hz and above. As was seen in Figure 14, lateral clicks had a slightly raised F3 compared with the other clicks. While lateral approximants typically have energy around 1200 Hz, the lateral clicks typically have a prominence between 900-2000 Hz. The similarities between lateral clicks and other laterals can be seen clearly by comparing the spectra of lateral fricatives with those of the clicks. Unfortunately recordings of lateral fricatives in Xhosa were not available, but those of the closely related language Zulu were, from other UCLA projects. The spectra of a lateral click and a lateral fricative before the vowel /a/ can be seen in Figure 17. Both consonants have a similar spectral range and shape. The spectra both show energy corresponding to the second formant roughly around 1200 Hz, and significant energy above 3000 Hz. Both spectra have a zero around 1800 Hz.

*2. **Amplitude**.* Clicks are characteristically strong in intensity compared with pulmonic consonants. This can be seen in Figure 18 which shows that clicks in final, unstressed syllables are of greater amplitude than pulmonic stops in the same environment. Although there can be a great deal of variation in the amplitude of clicks, they are typically much greater in intensity than pulmonic stops.

The difference in intensity between clicks and pulmonic consonants is related to the difference in degree of change in intraoral pressure upon the release of the stop closure. Pulmonic stops typically have an intraoral pressure of 7-8 cm H_2O. The back click closures are similar to velar consonants in that pharyngeal pressure before release is approximately 7-8 cm H_20, at least for the clicks in Nama (Ladefoged & Traill 1984).

Clicks are made with an ingressive airflow and have a negative intraoral pressure. Based on data in !Xóõ, Kagaya (1978) estimates the minimum pressure in the cavity between the front and the back closure for the three click types. The dental and lateral clicks are estimated to have the lowest minimum pressures, -75 and -70 cm H_20, respectively. The alveolopalatal click is estimated to have a minimum cavity pressure of -50 or -60 cm H_20. All of the clicks are estimated to have very low minimum pressures.

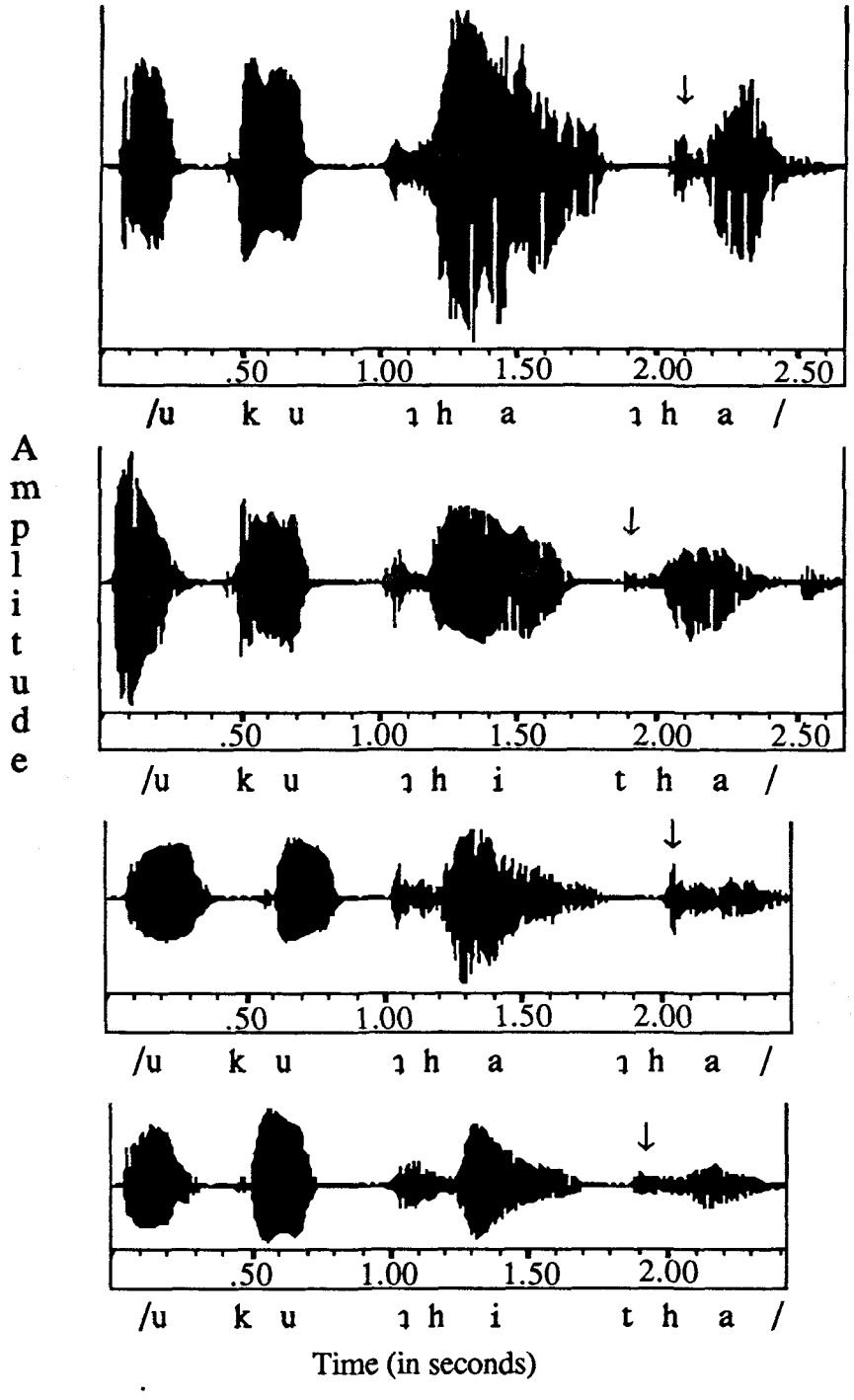


Figure 18: Waveforms of words spoken by two female Xhosa speakers (the top two panels are of speaker KP and the bottom two of speaker VS) illustrating voiceless aspirated dental clicks and alveolar stops in a word-final, unstressed syllable.

The change in pressure experienced during the production of a click is greater than the change in pressure associated with pulmonic consonants. However, it is not clear if the relationship between negative pressure and intensity is the same as for positive pressure.

Excluding bilabial clicks, for which no data are available, the further front the front click closure is, the lower the estimated minimum pressure. It is also the case that the clicks with the lowest minimum pressures are the affricated clicks, however, the alveolopalatal click is typically higher in amplitude than the affricated clicks. Presumably because they are affricated and thus have a more gradual change in pressure upon release, the dental and alveolar lateral clicks are lower in amplitude. It has been claimed that the amplitude of the burst of pulmonic stops increases the further back the place of articulation (Dixit and Brown 1978). This is also the case for the clicks.

*3. **Temporal characteristics**.* It has been shown that the dental and lateral clicks are more affricated than the alveolopalatal clicks. We will now consider whether these affricated clicks are similar in duration and amplitude of friction to pulmonic affricates, and whether the unaffricated clicks are similar to pulmonic stops.

Although the period of frication of clicks is much shorter than that seen in pulmonic fricatives and affricates, clicks have longer mean durations than either pulmonic stops or affricates. For instance, in one study, affricates averaged 123 ms (s.d. 42) and stops 89 (s.d. 34) (Crystal and House 1988b) while the shortest clicks (dentals) averaged 179 ms (s.d. 29). It must be remembered, however, that these measurements were made from tokens in running speech.

The lateral and alveolopalatal clicks have longer total durations than the dental clicks. For pulmonic consonants, mean durations of stops tend to be much shorter than the mean durations of affricates (Crystal & House 1988b). We might expect the alveolopalatal click to have a much shorter total duration than either of the affricated clicks, but even though the alveolopalatal clicks have shorter VOTs, they have longer closure durations. Kagaya (1978) says that the difference in duration between the affricated and unaffricated clicks is not as great as the difference usually seen between pulmonic affricates and stops. With respect to total duration, clicks fail to behave like pulmonic stops and affricates.

Although the closure durations are not a direct measure of the duration of the front click closure, but of the time between the back closure and the front release, click type does influence the closure duration. The alveolopalatal clicks, which are made with the most back front closure, have the longest mean closure duration. The dental clicks, which have the most anterior front closure of all the clicks, have the shortest closure duration. Other studies have not shown that the closure duration will be longer the further back the place of articulation (Crystal and House 1988a). Unlike pulmonic affricates which typically have shorter closure durations than pulmonic stops (Umeda 1977, Maddieson 1980), unaffricated alveolopalatal clicks have a longer mean closure duration than affricated clicks. This is evidence for affrication being a physiological by-product rather than a linguistically distinctive property.

*4. **Coarticulation**.* Coarticulatory relations between clicks and vowels are less extensive than those between other consonants and their following vowels. However, this is not surprising, considering that the tongue body cannot so freely vary its position in clicks. Presumably both the front and the back of the tongue have to be in particular positions to produce the consonant. Coarticulation involving the tongue position of vowels must be limited. (This is similar to the constraints observed in vowel to vowel coarticulations across a consonant with a secondary palatal or velar articulation.) The only coarticulation effect seen is that due to the anticipation of vowel rounding, since this does not involve a gesture used in the click prodution.

Although there was no significant effect of click type on vowel quality in Xhosa, the place of front click closure can influence the quality of an adjacent vowel in !Xóõ (Traill 1985), where the alveolar click patterns with dental consonants in conditioning a rule of /a/ raising.

5. Conclusion

At the beginning of this thesis, it was noted that the uniqueness of clicks causes them to be difficult to incorporate into classificatory systems. It has been seen that all clicks share some acoustic properties which could be used as the acoustic definition of a feature. Because affricated and unaffricated clicks do not pattern with pulmonic stops and affricates with respect to closure duration and total duration, this is evidence for affrication being a physiological by-product rather than a linguistically distinctive property. The alveolopalatal clicks pattern with non-anterior coronal consonants in having very compact release spectra and a predominance of energy in the lower frequencies. Like pulmonic laterals, the lateral clicks have energy at 3000 Hz and above. Lateral clicks have a prominence between 900 and 2000 Hz, and lateral approximants typically have a prominence around 1200 Hz. Coarticulatory relations between clicks and vowels are less extensive than those between other consonants and their following vowels. Neither the front nor the back click closure varies much according to vowel context.

References

Beach, D.M. 1938. *The Phonetics of the Hottentot Language*. W. Heffer & Sons Ltd., Cambridge.

Bill, M. 1974. The influence of the Hottentot languages on the Bantu languages. *Limi* 2.2. 63-77.

Blumstein, S. 1986. On acoustic invariance in speech. in *Invariance and Variability in Speech Processes* ed. J. Perkell, D. Klatt. Laurence Erlbaum. New Jersey. 178-201.

Blumstein, S. & Stevens, K. 1979. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *JASA* 66(4). 1001-1017.

Claughton, John S. 1983. *The Tones of Xhosa Inflections*. Communication No. 13. Dept. of African Languages, Rhodes Univ., Grahamstown.

Crystal, T. & House, A. 1988a. The duration of American-English stop consonants: an overview. *J. Phon* 16. 285-294.

Crystal, T. & House, A. 1988b. Segmental durations in connected-speech signals: Current results. *JASA* 83.4. 1553-1573.

Dart, S. 1991. *Articulatory and Acoustic Properties of Apical and Laminal Articulations*. Dissertation. UCLA.

Davey, A. S. 1975. *Some Aspects of the Phonology of the Noun in Xhosa*. M.Litt. Univ. of Edinburgh.

Dixit R. P & Brown W. S. 1978. Peak magnitudes of supraglottal air pressure associated with affricated and nonaffricated stop consonant productions in Hindi. *J. Phon* 6.353-365.

Elliot, Aubrey. 1970. *The Magic World of the Xhosa*. Collins, London.

Finlayson, R. 1982. Hlonipha--the women's language of avoidance among the Xhosa. *South African J. of African Languages*. Supp. 1982 (1). 35-60.

_____1987. Southern-Bantu origins. *South African J. of African Languages*. 50-57.

Grimes, Barbara F. (ed.) 1988. *Ethnologue: Languages of the World*. 11th ed. Summer Institute of Linguistics, Inc., Dallas, Texas.

Jordan, A.C. 1966. *A Practical Course in Xhosa*. Longmans Southern Africa (pty) Ltd. Johannesburg.

Jongman, A., S. Blumstein & A. Lahiri 1985. Acoustic properties for dental and alveolar stop consonants: a cross-language study. *J. Phon* 13, 235-251.
Kagaya, Ryohei. 1978. Sound spectrographic analysis of Naron clicks: A preliminary report. *Research Institute of Logopedics and Phoniatrics, Faculty of Medicine, Univ. of Tokyo, Annual Bulletin* 12. 113-125.
Khumalo, J.S.M. 1981. *Zulu Tonology*. Univ. of the Witwatersrand, Johannesburg.
Ladefoged, Peter. 1982. *A Course in Phonetics*. 2nd ed. Harcourt Brace Jovanovich, New York.
_____ & A. Traill. 1984. Linguistic phonetic description of clicks. *Language* 60. 1-20.
Lanham, L. W. 1964. The proliferation and extension of Bantu phonemic systems influenced by Bushman and Hottentot. *Proceedings of the Ninth International Congress of Linguistics*. London, Mouton & Co. 382-391.
Louw, J. 1977a. Clicks as loans in Xhosa. *Bushman and Hottentot Linguistic Studies 1975*. 82-100.
_____. 1977b. The linguistic prehistory of the Xhosa. *Zur Sprachgeschichte und Ethnohistorie in Afrika: Neue Beiträge afrikanistischer Forschungen*. eds. W.J.G. Möhlig, F. Rottland & B. Heine. Dietrich Reimer, Berlin. 127-151.
_____1979. A preliminary survey of Khoi and San influence in Zulu. *Khoisan Linguistic Studies 5*. 8-21.
_____1986. Some linguistic influence of Khoi and San in the prehistory of the Nguni. *Contemporary Studies on Khoisan* 2. 141-168.
_____A.S. Davey, & P.C. Taljaard (compilers). 1976. rev. issue 1980. *Xhosa I. Guide 1*. UNISA, Pretoria.
Maddieson, I. 1980. Palato-alveolar affricates in several languages. *WPP* 51.120-126.
Ownby, Carolan Postma. 1981. Early Nguni history: linguistic suggestions. *South African J. for African Languages* . Supp. 1981. 60-81.
_____1985. *Early Nguni History: The Linguistic Evidence and Its Correlation with Archaeology and Oral Tradition*. Dissertation. UCLA.
Stevens, K. 1989. On the quantal nature of speech. *J. Phon* 17.3-45.
Traill, Anthony. 1977. The phonological status of !Xóõ clicks. *Khoisan Linguistic Studies* 3. 107-131.
_____1978. Another click accompaniment in !Xóõ. *Khoisan Linguistic Studies*. 5. 22-9.
_____1981. *Phonetic and phonological studies of !Xóõ Bushman*. Doctoral thesis, Univ. of the Witwatersrand. Johannesburg.
_____1985. *Phonetic and Phonological Studies of !Xóõ Bushman*. (Quellen zur Khoisan-Forschung, 5.) Helmut Buske, Hamburg.
_____1991. The feature geometry of clicks. manuscript. Univ. of Witswatersrand. Johannesburg.
_____J. S. M. Khumalo, & P. Fridjhon. 1987. Depressing facts about Zulu. *African Studies* 46.2.87. 255-274.
Umeda, N. 1977. Consonant duration in American English. *JASA* 61.3. 846-858.
Wilkes, A. (compiler, revised ed. by B.P. Putu, Z. Jama). 1987. *Course Xho100, Block A, Language Description: Phonetics*. Dept. of African Languages, Vista Univ.

VOWEL PERCEPTION IN A SECOND LANGUAGE

Barbara Blankenship

INTRODUCTION

The relationship between a listener's identification of a vowel and the vowel's acoustic properties is an important question in language perception research. Although it is apparent that the way a speech sound is produced depends on the language background of the speaker, it is less well known that a listener's interpretation of speech sounds is also affected by his language experience. The purpose of this paper is to explore speech perception among bilingual listeners.

It is well established that the location of the first three prominences (formants) of the frequency spectrum of a vowel are highly correlated with the perceived identity of the vowel. If one ignores rhotacized vowels such as that in the American English pronunciation of "bird", the frequencies of the first two formants provide sufficient information to differentiate all the remaining English vowels. (Whether this is the information actually used by listeners is a separate question.) The frequency of the first formant (F1) correlates roughly with what may be called the height of the tongue and that of the second formant (F2) correlates roughly with how far forward the bulge of the tongue is located when pronouncing the vowel. Thus it is possible to draw an approximation of the tongue position for any vowel by plotting the frequency of F1 on the vertical axis and the frequency of F2 on the horizontal axis of a graph. Fig. 1 shows the F1 and F2 frequencies, and thus the nominal tongue positions, for five American English vowels, based on the Peterson and Barney (1952) average values for 33 male speakers. To facilitate comparison with conventional vowel charts, the figure has its point of origin in the upper right corner, with F1 along the abscissa and F2 along the ordinate.

The single point for each vowel is an idealization. In actual speech, a vowel has slightly different values of F1 and F2 each time it is spoken, due to the influence of nearby segments in the utterance and to random variation. Thus a more accurate picture of the vowel's range of possibilities is a larger shape on the graph, as shown in Fig. 2, (based on Disner's (1983) representation of the Peterson and Barney (1952) data for 33 male speakers.) Such a shape is referred to here as a vowel area.

The graph provides not only a picture of the possible tongue positions in speaking, but also a convenient abstraction for the "map" a listener might use in distinguishing the vowels he hears, if one assumes that the listener identifies each incoming vowel sound by normalizing it for the particular speaker, then comparing it to a mental map of vowel formant frequencies. A listener's mental map may not look like this graph, but the graph enables us to compare perceptual differences between listeners. An experimental subject identifies sounds near the center of a vowel area very consistently. Sounds near the edge of an area, particularly where two vowel areas overlap, are identified less consistently. The regions of uncertainty, though similar, are not identical across speakers of the same language.

This paper will investigate the variation in such maps due to language experience, specifically whether the listener is monolingual or bilingual.

Vowel Perception Research

Much of the experimentation in vowel perception has been concerned with establishing the average response to stimuli by speakers of a given language. (Synthesized vowel stimuli: Ainsworth and Millar, 1971; Fox, 1984; Nearey, 1989; Johnson, 1989; Miller, 1989; Beddor and Hawkins, 1990. Natural vowel stimuli: Peterson and Barney, 1952; Fox, 1982, and others). A related area of investigation is the cross-language study of vowel perception, in which monolingual speakers of different languages are shown to contrast in their responses to the same stimuli. (Synthesized vowels: Stevens, Libermann, Studdert-Kennedy, and Ohman, 1969; Terbeek, 1977. Natural vowels: Flege and Hillenbrand, 1984; Munro, 1990; Bohn and Flege, 1990.) Willis (1971) has also described single-language dialect differences reflected in vowel perception.

Figure 1
Average F1 and F2 values
for 5 American English vowels

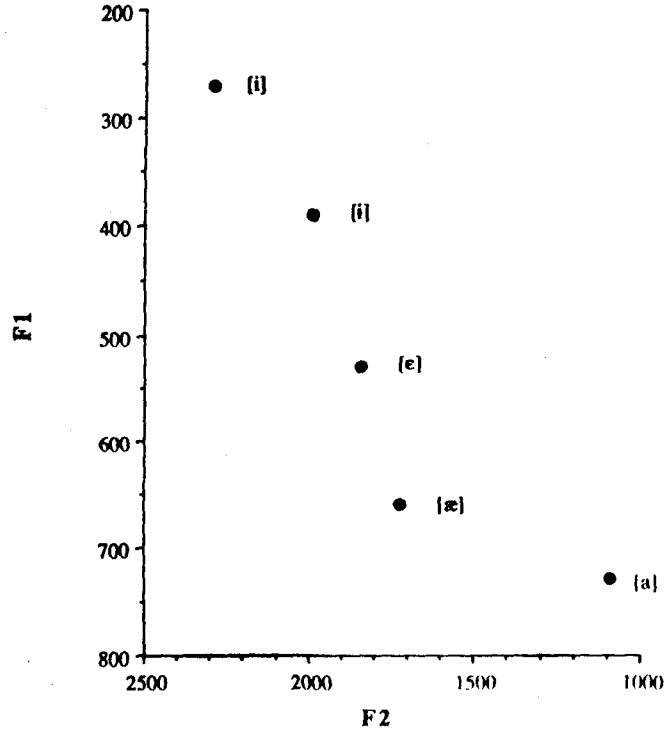
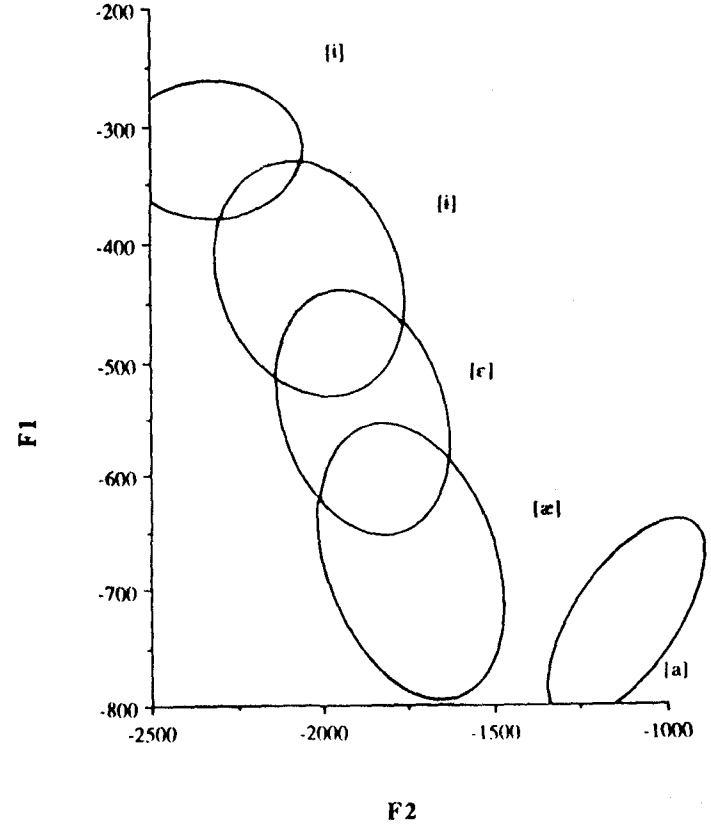


Figure 2
Average F1 and F2 areas
for 5 American English vowels



Typically the listeners in these experiments have been native speakers of the language or dialect under study. Second-language (L2) vowel perception, and the effect of second-language learning on native language (L1) vowel perception, are less widely investigated. Yet there are interesting questions to be answered in connection with second language acquisition. Consider what happens when a person learns a second language. Since it is unlikely that the two languages will have identical vowels, the learner cannot use his unaltered L1 map to perceive L2 sounds. Not only is L2 likely to have new vowels that don't correspond to any L1 area on the map, but there will also be similar vowels in L1 and L2 whose map areas overlap only partially. There might be two L2 vowels occupying the area of a single L1 vowel, requiring a finer grained map for L2. Given such possibilities, the following questions emerge.

1. Does the listener have separate maps for each language, or is his original map merely modifed to accommodate the new language?
2. Does learning a second language alter the listener's perceptual map for sounds in his first language?
3. Might the answers to these questions be different depending on how much L2 experience the listener has, which languages are involved, or even which vowels are involved?

Let us examine the considerations of point 3 in more detail.

Amount of L2 Experience

Elman *et al.* (1977) used English-Spanish bilinguals in a study of perception of voice onset time (VOT) after voiced and voiceless stops. Subjects' fluency in each language was rated on the basis of recorded conversation and a list of test words. The lower of the two fluency ratings determined the strength of bilingualism for each subjects. The study found that strong bilinguals appear to have two separate perceptual targets for differentiating voiced from voiceless stops (voice onset time (VOT) after b and p), depending on whether they are listening in English or Spanish. Weak bilinguals differentiate voicing at the same VOT in both languages.

Bohn and Flege (1990) found that for L2 phones that do not have a close correlate in L1, experienced and inexperienced bilinguals may use different perceptual cues. Native German speakers with more English experience differentiated English [æ] (for which there is no similar German sound) from [ɛ] (for which there is a very similar German sound) primarily on the basis of formant frequencies, as do native English speakers. Native Germans with less experience did not make a strong frequency distinction, but differentiated the same pair of vowels on the basis of duration.

Which languages are involved (Familiar vs. unfamiliar phonetic cues)

Tees and Werker (1984) found that native English learners of Hindi could learn to perceive non-English contrasts such as that between the breathy voiced dental stop /dh/ and the voiceless aspirated dental stop /th/. Although the Hindi glottal state contrast does not correspond to anything in English, the VOT difference between the two phones falls across the English VOT boundary. The same learners could not perceive the contrast between the Hindi dental and retroflex voiceless stops /t/ and /ʈ/ because there was no English parameter that would give them a cue to retroflexion.

English listeners can relate Hindi /dh/ and /th/ to the two separate English sounds /d/ and /t/. But Hindi /t/ and /ʈ/ both relate to the single English sound /t/.

Munro (1989) found that monolingual English speakers used frequency cues, but monolingual Arabic speakers used duration cues, to categorize unfamiliar French vowels that were systematically varied in formant frequency and duration. The Arabic speakers probably paid more attention to duration cues because Arabic has phonemic contrasts in vowel length. When L2 learners have a choice of cues, they use the cue that is familiar from past language experience.

On the other hand, Flege and Bohn (1989) found that native Spanish speakers differentiated English [i] and [ɪ] on the basis of duration, even though duration is not contrastive in Spanish. Since both English phonemes were equated with Spanish [i] in terms of formant frequencies, they shared an area on the English learner's F1 and F2 map, requiring the use of additional cues to differentiate them.

Which individual segments (Presence vs. absence of similar segment in L1)

Flege and Hillenbrand (1984), investigating experienced native English speakers of French, found evidence that for L2 vowels that do not have a close correlate in L1 (such as the French [y]), the L2 learner develops a new perceptual area for the novel vowel. But L2 vowels that do have a close correlate in L1 (such as the French [u]) are judged to be equivalent to the L1 vowel, and thus share the L1 area on the listener's perceptual map instead of acquiring their own area, even when the shared area may be relatively inaccurate for the L2 vowel. L1 phonetic experience thus impedes the formation of an accurate L2 perceptual target for a similar sound. "Judging acoustically different phones as belonging to the same phonetic category seems to underlie the process of speech perception. The continued operation of this perceptual process in L2 learning may lead to inaccurate perceptual targets for L2 phones." (Flege and Hillenbrand, 1984, p. 719). Thus for similar L1 and L2 sounds, contrary to Elman *et al*. (1977), experienced bilinguals did not appear to have separate perceptual maps. (But note that Elman *et al*. investigated consonants, which may be less readily judged as equivalent across languages.)

To summarize, "similar" L2 vowels map onto L1 vowel areas. (It has not been established how close together the vowels must be in order to be categorized as similar, whether similarity resides only in spectral cues, or whether different sets of vowels are similar for different listeners.) When L2 users must distinguish sounds that share a map area, they resort to other cues, preferably those known from L1. Of the studies just described, only those of Flege used bilingual subjects listening to vowel sounds. Each of his studies was concerned with how bilinguals differentiate between two L2 vowels that have varying degrees of similarity to an L1 vowel. Table 1 summarizes the vowels studied.

Reference	L1	L1 vowel	L2	L2 vowels	
Flege and Hillenbrand	English	[u]	French	[u]	[y]
Bohn and Flege	German	[ɛ]	English	[ɛ]	[æ]
Flege and Bohn	Spanish	[i]	English	[i]	[ɪ]

Table 1. Vowels used in some previous bilingual studies

Topic of this study

It appears that no bilingual study has looked at more than two L2 vowels at a time. It would be interesting to have a more complete L2 perceptual map in order to determine how novel and non-novel vowels affect each other and how the L2 map relates to the L1 map for an individual speaker.

To explore a broader range of L2 vowels, this study used native speakers of a 5-vowel language (Spanish), who have had to acquire additional vowels when they learned a second language (English). By comparing the Spanish vowel perception of these listeners to that of monolingual Spanish listeners, we can ascertain whether the vowel maps for the two groups are different and thus infer whether a change has taken place in the bilinguals' L1 map due to second language acquisition. By comparing the bilinguals' English vowel perception to their Spanish vowel perception, we can discover differences and similarities between the L1 and L2 vowel maps and speculate on the perceptual strategies used in L2.

From the general hypothesis that L2 and L1 perception have no effect on each other, two specific hypotheses were to be tested:

- There is no change in the L1 vowel map due to L2 acquisition. Bilingual native Spanish speakers will exhibit perceptual areas for Spanish vowels that are centered at the same F1 and F2 values as those of monolingual Spanish speakers, (analysis 1).

- The L2 vowel map is not the same as the L1 vowel map. Vowel tokens with identical F1 and F2 will be identified by bilingual listeners as different vowels, depending on whether they are listening in Spanish or English, (analysis 2).

Having the subjects listen to stimuli over an entire vowel map would have made the test sessions unworkably long. Therefore the non-low back vowels, which in any case offer additional complications due to rounding, were omitted. To eliminate allophonic and coarticulatory effects that might be different across the two languages, the investigation was limited to isolated vowels. Only the frequencies of the first two formants (F1 and F2) were varied, since they are easy to measure and control experimentally, and are widely identified as major cues to vowel identification in English. Other formants and such other phonetic cues as formant bandwidth and amplitude, vowel duration, and inherent spectral change were not varied.

METHOD

Stimuli

Front vowel stimuli were created using a Klatt formant synthesizer (cascade branch). F1 and F2, the independent variables, were varied according to the experimental design (see Choice of F1/F2 variables, below). F3 was calculated from F1 and F2 using the formulas described in Nearey (1989). F4 through F6 were left at the synthesizer's default values. The bandwidths of F1 through F3 were 50, 100, and 104 Hz, respectively, as in Fox (1984). The bandwidths of F4 through F6 were left at the default values. Values of the invariant parameters are shown in Table 2.

Formant	Frequency	Bandwidth
F1		50
F2		100
F3		104
F4	3500	200
F5	4500	200
F6	4990	500

Table 2. Invariant parameters in this experiment.

The F0 frequency and amplitude of voicing were ramped for a more natural sound, with the same ramp for each stimulus. The F0 frequency was interpolated from 100 Hz at 0 msec to 110 Hz at 20 msec, then to 80 Hz at 200 msec. The amplitude of voicing was interpolated from 50 db at 0 msec to 58 db at 180 msec, then to 45 db at 200 msec. These values were arrived at by trial and error. After synthesis the peak amplitudes of the stimuli were equalized.

A pilot study testing durations for the vowels and interstimulus intervals had shown that naive subjects are very uncomfortable identifying stimuli of short duration (140 msec) or with short interstimulus intervals (2.5 sec), but reasonably confident with stimuli of 200 msec and an interstimulus interval of 3 seconds. The stimuli for this experiment were therefore set at 200 msec, with a 3 second interstimulus interval.

The stimuli were presented in blocks of 6; there was a 7 second pause between blocks. There were 36 stimuli, each presented 5 times within 5 pseudo-random orders, for a total of 180 tokens.

Choice of F1/F2 variables

In order to discover the monolingual Spanish and English regions of uncertainty for this kind of stimuli, a pilot study was run with 4 Spanish and 4 English speakers, using synthesized

Figure 3
Uncertainty regions for Spanish monolinguals

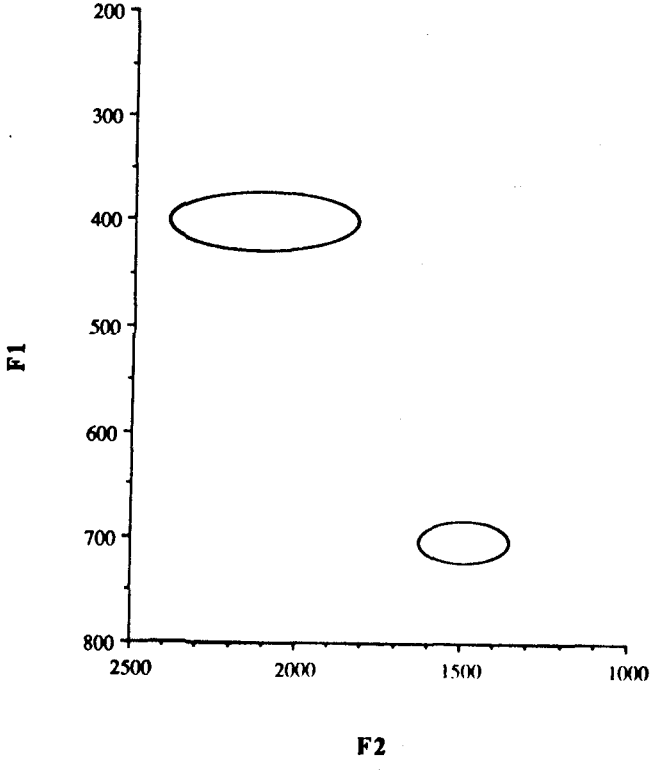
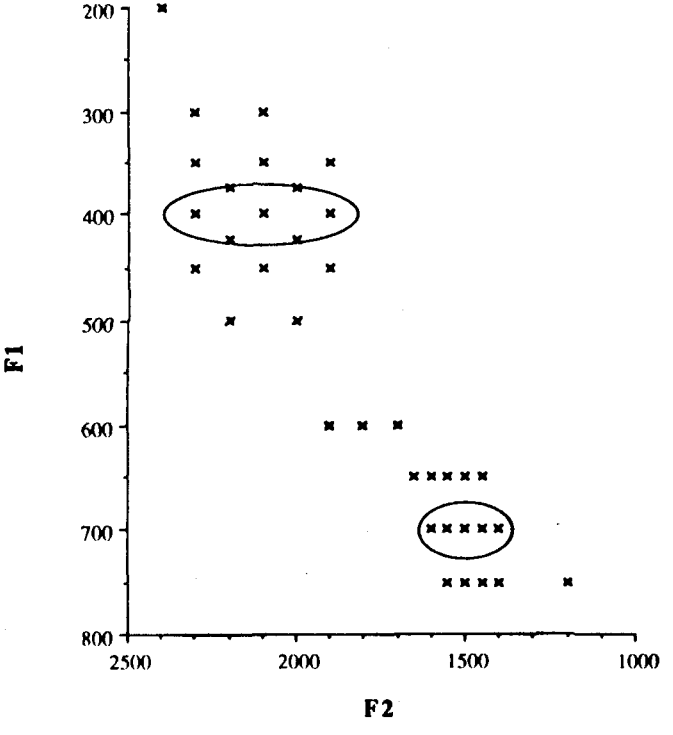


Figure 4
F1 and F2 values of the stimuli



front vowel stimuli where F1 and F2 varied uniformly by 100 Hz. Spanish monolinguals disagreed within and between subjects on vowel identifications in the two regions shown in Fig. 3: F1=400 Hz, F2=1900-2300 Hz and F1=700, F2=1400-1600 Hz.

The first of these regions of uncertainty corresponds to the Spanish [i/e] boundary, whereas for the four English listeners it was within the region for [e] or [ɛ] or at the [e/ɛ] boundary, depending on the listener. The second region corresponds to the Spanish [e/a] boundary and was in the middle of the [æ] region for the English listeners.

In acquiring English, a Spanish speaker would have to learn to identify some English vowels that are centered in regions where no Spanish vowels are centered. Thus bilinguals' responses to stimuli in such regions can tell us something about how the perceptual map has adjusted during L2 acquisition. Stimuli for the main experiment were therefore concentrated around the two regions of uncertainty revealed by the pilot experiment, with a few surrounding stimuli for fuller coverage. Figure 4 shows F1 and F2 values for all stimuli. Stimuli that are nearer the top of the figure are high vowels; those nearer the right are back vowels, as in the previous figures. The stimuli cover portions of the perceptual regions for /i/, /e/, and /a/, in Spanish, and /i/, /ɪ/, /e/, /ɛ/, /æ/, and /a/ in English.

The ideal spacing for stimuli in the areas of concentration would be the just noticeable difference (JND), about 20 Hz for F1 and 100 Hz for F2 in the high-vowel region, and about 30 Hz for F1 and 80 Hz for F2 in the low-vowel region for tasks similar to this one (Flanagan 1955). Since such density would yield an unworkably large number of stimuli in the high-vowel region, stimuli were added only at alternate JND positions in that part of the F1/F2 grid, creating the sparser pattern shown at F1=350-450 in Fig. 4.

In addition to the regions of concentration, stimuli were added at F1=600, F2=1700-1900 in order to locate the boundary between English [ɛ] and [æ] for bilingual speakers, a boundary that had not been revealed by the pilot experiment.

Tokens of one extreme [i] (F1=200, F2=2400) and of one extreme [a] (F1=750, F2=1200) were included as performance criteria. Any subject who failed to identify these stimuli as [i] and [a] respectively was eliminated from the analysis. (See Criteria, below.)

Subjects

Subjects who successfully completed the experiment (see Criteria, below) were twelve monolingual Spanish speakers and twelve bilingual speakers whose first language was Spanish and second language was English.

The monolinguals were twelve adult females born in Mexico who now reside in Los Angeles and have lived in the United States from 6 months to 21 years. All had grown up in Spanish-speaking homes and continued to speak only Spanish, although their children and some of their spouses use English outside the home. None had lived in any other countries or studied other languages.

The bilingual subjects were two adult males and ten adult females. One was born in Ecuador, one in Peru, eight in Mexico, and two in the United States. All had grown up in monolingual Spanish-speaking homes. Four of them had started learning English when they entered U.S. schools at age 5 or 6; the others had begun learning English when they moved to the United States at ages 7 through 23. They now speak both English and Spanish at home and in other daily contacts. None had lived in any other countries nor studied other languages.

Procedure

Each subject listened to a tape with the 180 stimuli, played on a Marantz PMD 340 cassette recorder with Calrad 15-118 earphones. To reduce order effects, four tapes with different random orders were used, with three monolingual and three bilingual speakers listening to each tape. Each tape lasted approximately 15 minutes.

The monolingual speakers listened to the tape one time, using a Spanish response sheet that had 'piso peso paso' (floor, peso, step) printed on each line. The lines were grouped in blocks of six with a blank line between blocks, to correspond to the spacing of the stimuli on the tape. Subjects were required to mark one word on each line, using one line for each sound they heard on the tape. They were instructed to mark the word that contained a vowel most like the stimulus sound, to try to respond to each stimulus, and to guess if necessary. They were assured that there were no right or wrong answers.

The bilingual speakers listened to the tape twice on separate days, using the Spanish response sheet on one day and an English response sheet with the words "read rid raid red rad rod" on the other. Half of the bilinguals used the English response sheet first and half used the Spanish response sheet first. Each bilingual heard the same tape for both sessions, so that any order effects would be comparable over the two sessions. But subjects were told that they were listening to "the Spanish tape" or "the English tape". This fiction, along with the language of the answer sheet, was intended to put the subject into the right perceptual set for each test. [1]

Prior to testing, each subject was interviewed informally and the answers were tape recorded. The interview covered language background, ages of residence in various linguistic situations, and current language environment. Subjects were also asked to read aloud the words from the response sheet as a criterion test (See Criteria, below.) Monolinguals were interviewed in Spanish, bilinguals in English.

A practice tape and response sheet in the appropriate language were provided at the start of each test session. Subjects were allowed to adjust the volume of the recorder during the practice. Subjects were not allowed to adjust the volume or to pause during the actual test. Three subjects who did not want to continue were excused at the end of the practice.

Criteria

A subject's responses were used if they passed the criteria listed below. The first two were designed to ensure that subjects had adequate hearing and reading skills and understood the test situation. The last two were intended to eliminate subjects whose linguistic skills were inappropriate, such as speakers of unusual Spanish dialects or bilinguals who hadn't acquired all of the English vowels.

1. Subject identified all instances of the stimulus F1=200, F2=2400 as [i] and all instances of the stimulus F1=750, F2=1350 as [a] when listening in Spanish.
2. Subject left no more than 6 answers blank.
3. Subject used all of the possible response categories: [i,e,a] in Spanish and (if bilingual) [i,ɨ,e,ɛ,æ,a] in English.
4. Subject had 3 Spanish phonemes and (if bilingual) 6 English phonemes when he read the words from the response sheet aloud.

All subjects met criterion 4. Of the monolingual subjects who completed the test, three failed to meet criterion 1, two failed criterion 2, and two lost their place on the answer sheet. Of the bilingual subjects, one failed criterion 1, two failed criterion 2, eight failed criterion 3, one failed to return for the second session, and one was accidentally assigned the wrong tape to listen to. These

---

[1] The inclusion of 'raid' in this set may seem curious, since the English /e/ phoneme is usually realized as a diphthong. In a pilot study with four native English speakers, stimuli around F1=300-400, F2=1900-2400, has elicited a majority of [I] responses when presented at a duration of 140 msec, but a majority of [e] responses at 200 msec. Another pilot study with two native English speakers showed them to be uncomfortable identifying 200 msec stimuli when there was no [e] option on the answer sheet. Since the use of naive subjects required the use of 200 msec stimuli (as explained in the Stimuli section above), [e] was included on the English answer sheet.

21 were replaced with other subjects, although those who failed on criterion 3 proved to be of interest in the light of subsequent results. The results of those eight subjects will be discussed below, but not included in any statistical calculations.

RESULTS

The responses were pooled into three groups, monolingual listening in Spanish, bilingual listening in Spanish, and bilingual listening in English. Analysis 1 compares the first and second groups; analysis 2 compares the second and third groups.

Data for each subject were plotted on an F1 X F2 grid, (see Fig. 5a for representative cases.) Only vowels that received 50% percent or more of the responses for a given stimulus were included.

Individual monolingual subjects tended to respond very consistently, often with no overlap between vowel areas. The largest overlap for adjacent vowel areas was 50 Hz for F1 and 100 Hz for F2. The responses of bilinguals listening in Spanish were less consistent, and showed overlaps as large as 100 Hz in F1 and 200 Hz in F2. Fig. 5a shows individual responses from two subjects, one a typical monolingual and the other a typical bilingual subject responding in Spanish. Fig. 5b shows the responses of the same bilingual subject listening in English, and the responses of a speaker of General California English in the pilot study. (Due to different designs, the two sections of the figure do not include an identical set of stimuli.) Small diamonds indicate stimuli for which no vowel received more than 50% of the responses.

The responses of individual bilinguals listening in English were even less consistent, with large areas of overlap, instances of a vowel occupying two separate areas, and occasional instances of three vowels sharing an area. It is clear that for these listeners, F1 and F2 are not the only cues used to determine vowel identity. Duration, intensity, and inherent spectral change are important phonetic cues that were not included in this experimental design. We have no evidence whether the bilingual listeners rely on such additional phonetic cues or on higher level (e.g., lexical) information to identify English vowels.

By contrast, native English speakers in the pilot study responded with clear areas for each vowel based on F1 and F2 alone (except for the duration cue that differentiates [i] from [e], mentioned in the Procedure section above.)

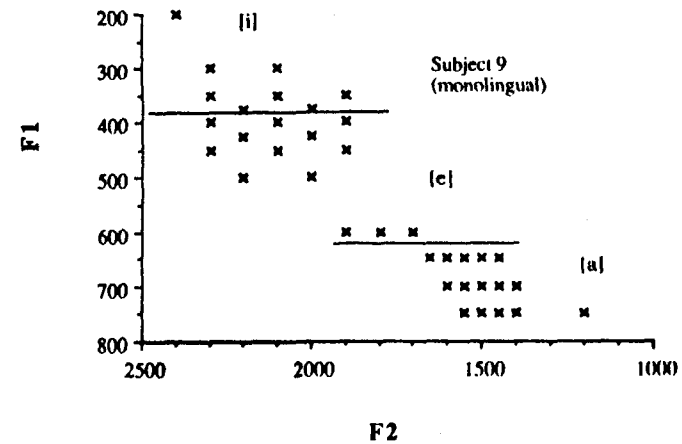
Analysis 1

The first experiment compares the responses of monolinguals and bilinguals listening in Spanish. Fig. 6 shows the total responses for both groups. Bilingual group responses in the ranges of F1=200-300 and F1=400-750 were identical to those of the monolingual group. In the range of F1=350-375, bilingual reponses included fewer [i]'s, resulting in a different [i] to [e] boundary.

To determine whether this difference was significant, the mean and standard deviation of F1 and F2 for all the [i] responses, all the [e] responses, and all the [a] responses in each subject group were calculated, as illustrated in Fig. 7. T-tests of the group means for each vowel showed no significant differences. Thus the response maps for the two groups are essentially the same.

A subject who hears a stimulus near the center of a vowel area on his perceptual map will respond consistently with the same vowel, while for a stimulus near a boundary he might respond to the various iterations with 2 or even 3 different vowels. The variation in responses averaged higher for the bilingual group than for the monolingual group on 30 of the 36 stimuli. This indicates that even though both groups have the same shape of response map for Spanish, on most stimuli the bilinguals found it more difficult to judge what the vowel was. This may indicate that English boundaries crossing the centers of Spanish vowel areas were causing a boundary-like inconsistency even in the Spanish responses.

Figure 5a
Sample response maps, Spanish



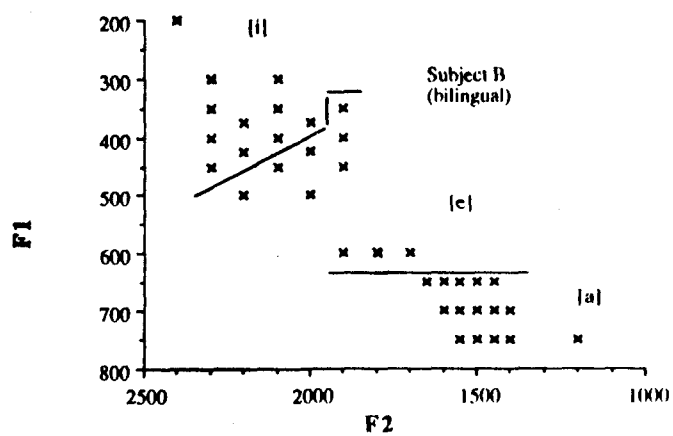
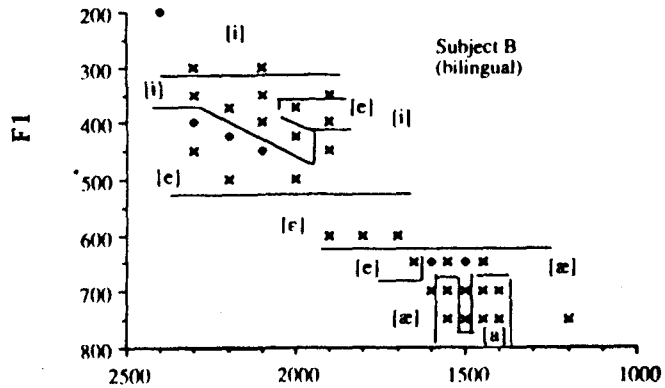


Figure 5b
Sample response maps, English



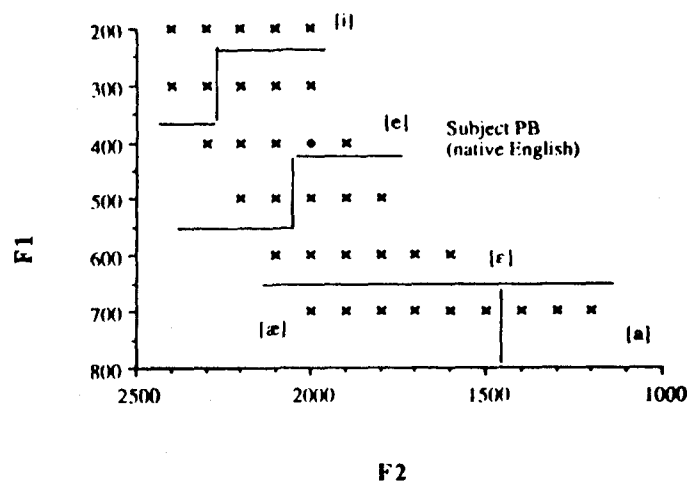


Figure 6
Monolingual and bilingual group responses, listening in Spanish

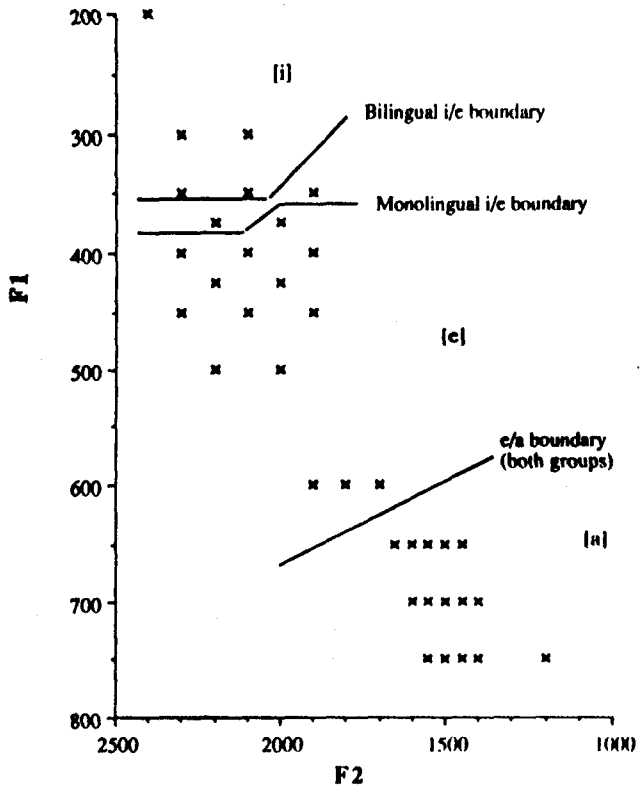
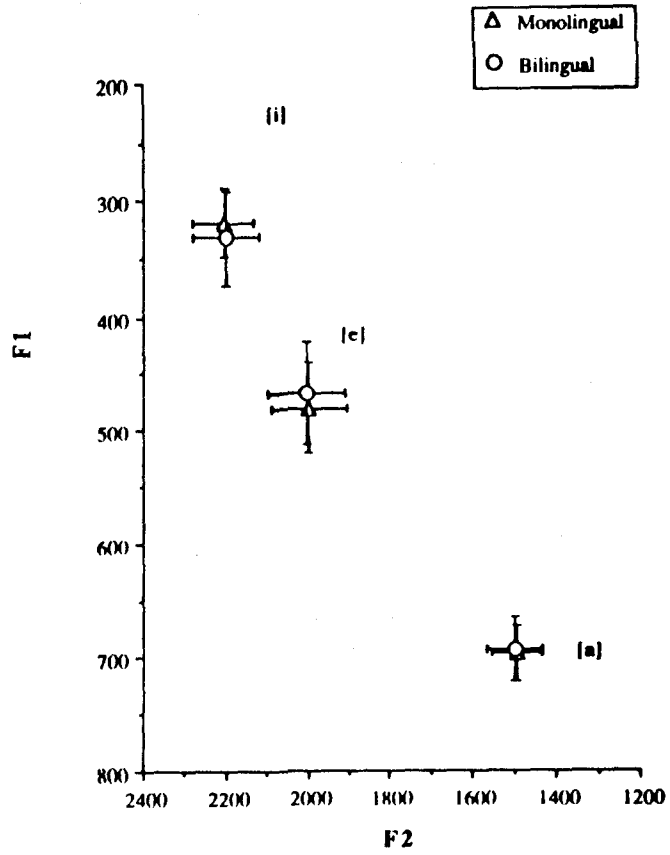


Figure 7
Group means and standard deviations, native speakers responding in Spanish



Analysis 2

The second analysis compares the bilinguals' Spanish responses with their English responses. As we saw in Fig. 5b, it is difficult to interpret the English responses by looking at their position on the F1 X F2 map. The map of means and standard deviations shown in Fig. 8 provides a way of understanding the English data.

It shows that there is nearly complete overlap between [i] and [ɨ], [e] and [ɛ], and [æ] and [a]. (The unexpected ordering between [e]/[ɛ] and [æ]/[a] will be discussed below.) By contrast, a similar map of two native English speakers from the pilot experiment (Fig. 9) shows distinct vowel areas with an overlap only between [e] and [ɨ], which in English are differentiated by duration (as shown in the pilot study discussed in the Procedure section above).

Although it is tempting to view the pattern in Fig. 8 as comprising only three vowels rather than six, the pattern does not occur for most of the individual subjects. Three of the individual maps contain six distinct vowel clusters, three of the maps have four clusters, and one of them has only two clusters. The remaining five maps do have three clusters, but only one of them groups the English vowels in pairs as in Fig. 8. Thus the figure does not represent the behavior of subjects as individuals, but only as a group.

It is well established that for native English listeners, F1 and F2 can be the primary determinants of vowel identity. But the lack of a distinct F1 X F2 mapping for each vowel by the bilinguals in this experiment indicates that additional parameters may be required by non-native listeners.

In order to determine the relationship of the Spanish and English vowels for bilinguals, English vowels whose F1 and F2 means were within 1 standard deviation of the means for any Spanish vowel were charted. Table 3 shows the vowel proximities for the combined bilingual group.

Spanish vowel	English vowel	Number of standard deviations
[i]	[i]	2
[i]	[ɨ]	2
[e]	[e]	2
[e]	[ɛ]	1
[a]	[æ]	1
[a]	[a]	2

Table 3. Proximity of English to Spanish perceptual vowels (bilingual group means).

Thus for this group the Spanish [e] has a strong tendency to map onto the English [ɛ] vowel area, and the Spanish [a] onto the English [æ] area. No other English vowels were within 2 standard deviations of a Spanish vowel. The vowel means for both languages are shown in Fig. 10.

Since the English vowels are in three clusters, it is tempting again to see the bilinguals' English pattern as a 3-vowel scheme based on the Spanish phonemes. But only five individual subjects displayed this pattern. Other patterns included English vowels that were not near (within 2 standard deviations of) any Spanish vowel, Spanish vowels that were not near any English vowel, and Spanish [e] coinciding with three different English vowels in the same listener's system.Table 4 shows vowels that were within 1 standard deviation for the individual listeners.

Where a Spanish vowel is near to one or more English vowels, it may be that the original Spanish map has been extended to fit the English sounds, or it may be that there is a separate map for each language, portions of which just happen to coincide. A Spanish vowel that is not near an

Figure 8
Group means and standard deviations, bilinguals responding in English

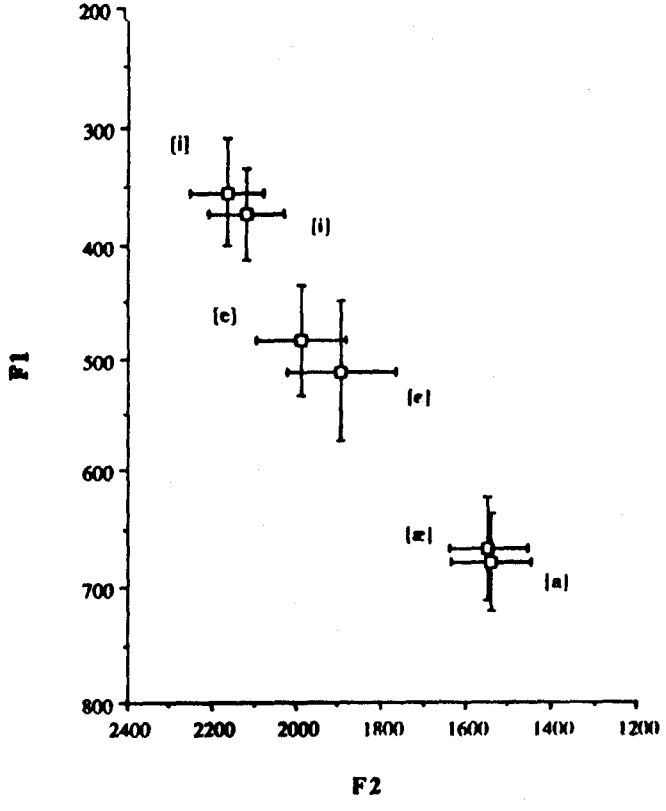


Figure 9
Group means and standard deviations, 2 native speakers of California English

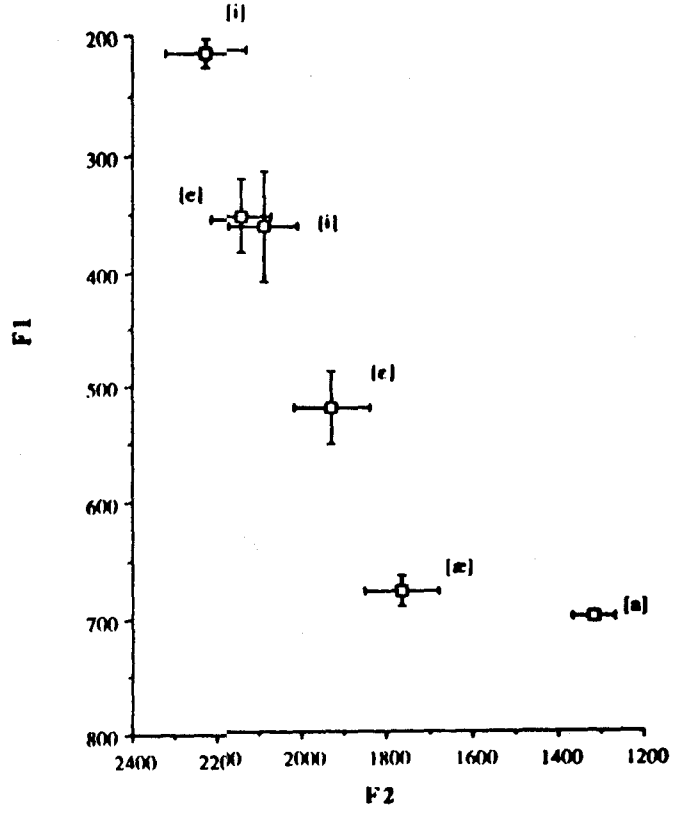


Figure 10
Mean responses from bilingual group

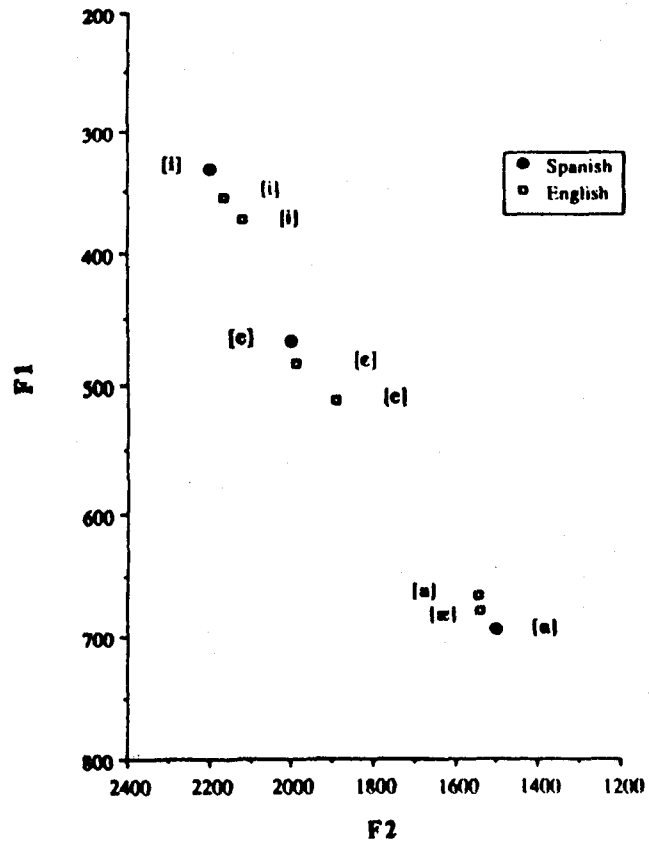
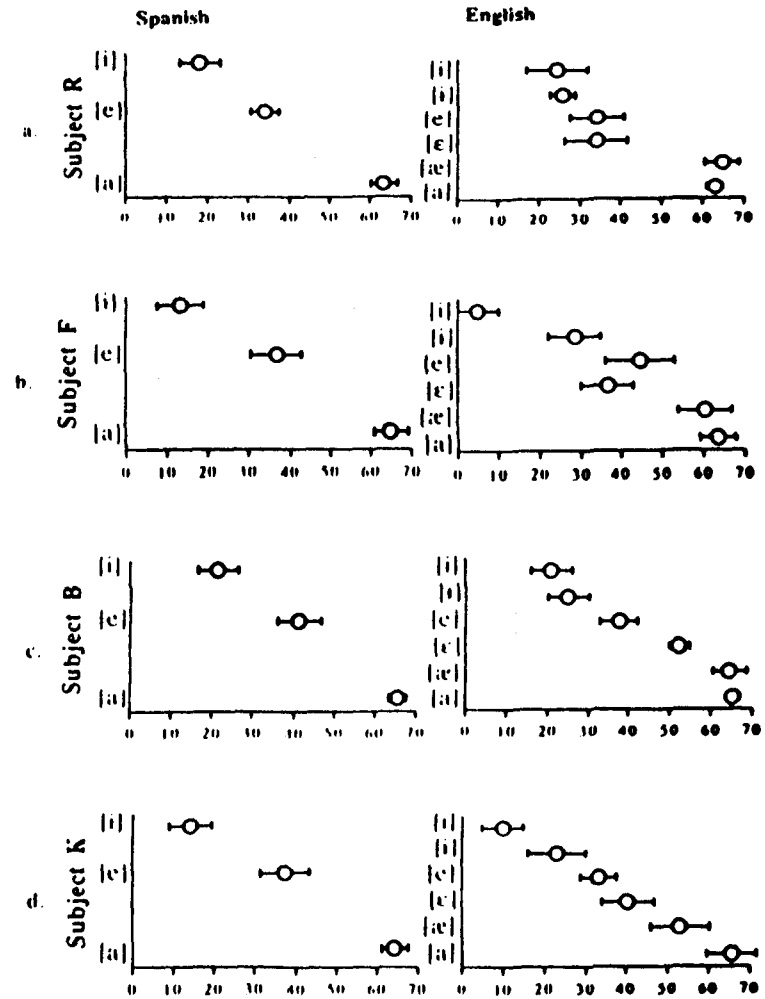


Figure 11
F2/F1 mean and standard deviation



English vowel provides evidence that a new map has been constructed for English.

To facilitate comparison between heavily overlapping vowel areas, each area was reduced to a single dimension represented by the difference between the logs of F2 and F1 (log F2/F1) (Miller, 1989), with a single mean and standard deviation. The resulting charts for four subjects are shown in Fig. 11.

Spanish vowel	English vowel within 1 s.d.	Number of subjects
[i]	[i]	1
[i]	[i] and [ɨ]	3
[i]	[ɨ]	2
[i]	none	6
[e]	[i] and [ɛ]	1
[e]	[ɨ] and [ɛ]	1
[e]	[e]	0
[e]	[e] and [ɛ]	3
[e]	[ɛ]	6
[e]	none	1
[a]	[æ]	2
[a]	[æ] and [a]	7
[a]	[a]	2
[a]	none	1

Table 4. English vowels within 1 standard deviation of a Spanish vowel (individual bilingual means).

The English portion of Fig. 11a is an example of a tightly clustered 3-vowel pattern. Two of the twelve bilingual subjects displayed this pattern. Using the standard deviation of each Spanish vowel as a unit of measure, the English [i] and [ɨ] are within 2 units of Spanish [i]; English [e] and [ɛ] are within 1 unit of Spanish [e]; and English [æ] and [a] are within 1 unit of Spanish [a]. The other subject in this group had English [ɛ] in the Spanish [i] cluster instead of the Spanish [e] cluster.

The English portion of Fig. 11b exemplifies a 3-cluster system where the center cluster is differentiated into three overlapping but separate vowels. The English [ɨ] and [e] are more than 1 unit away from the Spanish [e], but nearer to [e] than to the other Spanish vowels. Three of the subjects displayed this kind of pattern, although for one of them it was the Spanish [i] cluster rather than the [e] cluster that was differentiated into separate vowels.

The English portion of Fig. 11c shows a system that is no longer recognizable as three clusters. English [ɛ] is 2 units away from Spanish [e], midway between Spanish [e] and [a]. This is clearly an independent vowel rather than a member of the [e] or [a] cluster. Four of the subjects displayed this kind of pattern, although the independent vowel was different in each case. In one of them the independent vowel was English [ɨ], in one it was English [e] (with English [ɛ] occupying the Spanish [e] area), and in one it was English [a] (with English [æ] occupying the Spanish [a] area).

The English portion of Fig. 11d shows a definite 6-vowel pattern. The English [i] and [a] vowels are within 1 unit of their Spanish counterparts. The [e] and [ɛ] are within 1 unit of Spanish

[e] but on opposite sides of it and more than 1 unit from each other. The [ɨ] and [æ] are separate vowels that are distant from any Spanish vowel. Three subjects displayed this kind of pattern.

It must be stressed that regardless of the perceptual pattern, all bilingual subjects spoke English with 6 recognizably different vowels for [i], [ɨ], [e], [ɛ], [æ], and [a], not only in reading the word list, but also in conversation while answering interview questions.*Bilinguals' English vowel production*

Two native speakers of General California English (GCE) listened to the interview tape for instances of the six English vowels used in this study. The conversations provided an average of 2.4 [i], 4.5 [ɨ], 2.5 [e], 4.4 [ɛ], 3.4 [æ], and 1.5 [a] vowels in stressed position for each speaker. These vowels were judged to be reliably different from each other for each speaker, although they were not necessarily the same as GCE vowels. (See *Chicano English* below for a discussion of another possible L2 target language.) Speakers judged to have strong non-GCE accents produced [ɨ] and [ɛ] higher than is usually heard from native GCE speakers. One produced [æ] farther back than GCE [æ] in English words that have Spanish cognates (e.g., grammar, family, Spanish), but not in words without Spanish cognates (e.g., had, background, that, faster). Speakers judged to have light non-GCE accents generally varied from native GCE on the basis of consonant production only, although two of them also produced [e] and [o] without the native English diphthongs. In all other instances, the speakers with light accents produced vowels that were judged to be within the normal range for GCE, as did the speakers judged to have no non-GCE accent.

Rejected data

Since several of the bilinguals had patterns with fewer than six clusters, it is of interest to consider the eight bilingual subjects whose results were eliminated due to criterion 3, failure to use all six response categories on the answer sheet. Do these represent an earlier stage of L2 acquisition, with English vowel areas that are so similar to those of Spanish that not all the response categories (read rid raid red rad rod) can be distinguished?

One subject's results were invalid due to end effects: only the response words in the center of the page had been used. Subject 46, who had acquired English at age 6, displayed five distinct vowel areas similar to those in Fig. 11d, but heard no [e]'s among the 180 tokens. Subject 41, who had acquired English at age 6, displayed four vowel clusters like those in Fig. 11c, but heard no [i] among the 180 tokens. The remaining subjects, who had acquired English at ages 11 through 30, had patterns with three clusters; three were loosely clustered as in Fig. 11b and two were tightly clustered as in Fig. 11a.

Subject 16, the most extreme of the 3-cluster cases, had responses only on the three English words (read raid rod) whose vowels correspond to Spanish vowels, indicating that the English distinctions were not present for this listener. The other subjects whose results were eliminated span the same range of possibilities as do the subjects whose results were used. Except for subject 16, no unusual data were lost by eliminating the results.

Age and language experience

Table 5 groups the subjects according to the cluster patterns discussed in the previous section. It shows the ages at which subjects began using English, the number of years they have used it, and a rating (by two native English listeners) of the degree of Spanish accent (S for strong, L for light, and N for none.)

There appears to be no correlation between number of years of English usage and the pattern of the perceptual map. Except for subject K, the table shows good agreement between the average starting age, the strength of the accent, and the number of clusters in the perceptual map. Listener K is anomalous in being the only subject outside of the 3-cluster groups to have started using English as an adult, but she had English instruction in school as a child that may have given her more facility. Thus the kind of perceptual map that develops for L2 may be a function of the age at which L2 is acquired.

Fig. 11 group	Subject	Starting age	Total years	Accent	
a.	C	10	10	L	
a.	R	23	16	S	
		16.5	13		Group average
b.	F	9	12	L	
b.	P	15	13	S	
b.	Q	18	26	S	
		14	17		Group average
c.	B	11	10	L	
c.	E	5	15	N	
c.	I	7	13	N	
c.	L	5	15	L	
		7	13.25		Group average
d.	K	18	3	S	
d.	M	6	14	N	
d.	N	5	14	N	
		9.6	10.6		Group average

Table 5. Individual language background information.
Subjects are grouped by perceptual patterns.

Chicano English

Godinez and Maddieson (1985) have shown convincingly that Chicano English is not a foreign-accented English, but a native English dialect. Chicano English front vowels are produced with higher F1 and F2 values than the corresponding GCE vowels. This finding implies that subjects in the current experiment who learned English in a Chicano community might have acquired a different L2 from that of the subjects who learned English in a non-Chicano community. During the interviews, seven subjects said they had acquired English in a Chicano community, three said they had not, and the remaining two were not asked. T-tests of F1 and F2 for each vowel showed no significant difference between the Chicano and non-Chicano groups except for the [e] vowel. Godinez and Maddieson do not include [e] in their study; thus there is no way to determine whether the Chicano subjects in the present experiment had an [e] that was closer to Chicano English than to GCE. Since there were no other significant vowel differences between the two groups, this study will assume that the L2 target language was substantially the same for all subjects. The difference between the results of this experiment and those of Godinez and Maddieson may be due to two factors. First, Godinez and Maddieson studied production, while this experiment studies perception. Second, the subjects in Godinez and Maddieson were younger and less educated. The Chicano subjects in the current experiment were college students who had been exposed to a good deal of GCE on television and from teachers while they were young, and who continue to hear GCE daily in college. Some of these may be fluent in both dialects.

Unexpected phoneme order

Several bilingual subjects showed an unexpected reversal of phoneme positions on the vowel map, a pattern that is seen also in the group means of the non-high vowels in Fig. 8. The results for the twelve subjects included three instances of [i] lower than [ɨ] nine of [e] lower than [ɛ], and seven of [æ] lower than [a]. (None of the native English speakers in the pilot studies exhibited such reversals.) One possible explanation for the reversals is the orthography of the answer sheets. To a native Spanish reader the English word "red" might be closely associated with the sound sequence [reð] as it would be pronounced in Spanish. Likewise the English "rid" might be closely associated with the Spanish sounds [rið], and the English "rad" with the sounds [rað]. Two of the subjects did show a possible effect of orthography on speech production by reading the word "rod" as [rɔd], but only one of those showed any reversals on the perceptual map. If orthography were the cause of the reversals, one would expect the subjects who had started

English at age 5 or 6, and therefore not been exposed to Spanish reading and spelling prior to learning English, not to display such reversals. But the reversals are just as frequent among the young English learners as among the older ones.

Another explanation is random variation. If the listener made no distinction between, for example, [i] and [i̞], then [i] would be lower than [i̞] in about half of the responses. The reversals in this experiment do occur about half of the time (19 reversals out of 36 vowel pairs (3 pairs for each of 12 subjects)). Thus they may be caused by the bilingual subjects' not differentiating the phoneme pairs on the basis of F1 and F2. It would be interesting to know if this result is actually due to random mappings at the individual level, but there are not enough responses from each subject (5 responses per stimulus) to test for randomness.

DISCUSSION

Analysis 1

There is some evidence that L2 acquisition can change L1 production and perception. Flege (1987) found that native French speakers who were experienced in English produced French /t/ with a longer (i.e. English-like) VOT than did French monolinguals; experienced native English speakers of French produced English /t/ with a shorter (French-like) VOT than English monolinguals. Beckman (1986) found L2 effects on L1 perception of stress by native English speakers who had learned Japanese.

Analysis 1 was designed to see if there was an effect of L2 acquisition on L1 perception of vowels. No such effect was found. The monolingual and bilingual groups' maps for Spanish vowels were essentially the same, as shown in Fig. 7. Although the L1 maps of the individual bilinguals were somewhat different from each other, especially for [i], the mean location of each listener's vowels was within the range of the means for the monolingual Spanish subjects. There was no pattern of differences related to duration of English use or to age of English acquisition. Further studies are necessary to determine whether the lack of effect is specific to this experiment or more generally true for vowel perception. (Note that Flege (1987) investigated consonant production, while Beckman (1986) was concerned with the acoustic correlates of stress rather than individual segments.)

Analysis 2

The critical period hypothesis

Analysis 2 revealed a tendency for L1 and L2 maps to be very different for bilinguals who acquired L2 as children, but to be more similar for bilinguals who acquired L2 as teen-agers or adults. This finding is *a posteriori* and was not tested statistically. It appears that those who learned English at an early age were able to acquire, to a greater or lesser extent, separate perceptual maps for each language. Those who learned English at a later age stretched their L1 map in various ways to accommodate the vowel sounds of English, but were not always capable of differentiating English vowels on the basis of the F1 X F2 map alone. This finding supports any version of the critical period hypothesis which, among other things, predicts that extended exposure to a second language may not be able to overcome maturationally conditioned limits on the ability of adults to learn a new language.

Bohn and Flege (1990), studying native German speakers of English, found that experienced L2 listeners were more able than inexperienced listeners to distinguish between English [ɛ] and [æ] on the basis of formant frequencies. They state that this finding provides evidence against the critical period hypothesis, since the hypothesis predicts that adult L2 learners should not show an improvement based on duration of experience. They do not present any data, however, about the ages at which their subjects acquired English and whether acquisition age may have had an effect on their results. The present study found an effect due to age of L2 acquisition, but, contrary to Bohn and Flege, no effect due to the duration of L2 experience. Bohn and Flege's inexperienced group had lived in an English-speaking environment for an average of only 0.6

years, while the experienced group had an average exposure of 7.5 years. Except for one subject with 3 years' exposure, all bilingual subjects in the present study had lived 10 or more years in the United States. Thus there were no inexperienced L2 listeners as defined by Bohn and Flege.

Equivalence classification

Where the L1 and L2 maps for an individual listener were similar, two or more English vowels overlapped each other and a similar Spanish vowel (e.g., English [i] and [ɨ] sharing an area equivalent to that of Spanish [i].) This result supports the Flege and Hillenbrand (1984) hypothesis that when adult L2 learners classify a new phone as being equivalent to a known L1 phone, they use the L1 phone in L2; equivalence classification prevents their learning the L2 phone accurately. The 1984 study and subsequent related studies used French /u/ and /y/ as the L2 test vowels. Native English speakers had acquired the /y/ rather well, but /u/ remained inaccurate even among experienced L2 French speakers because its similarity to English /u/ prevented its being perceived as other than the English vowel. The present study elaborates on this point by finding cases where two (and for some listeners three) L2 vowels are so similar to an L1 vowel that they cannot be distinguished on the basis of frequencies.

It appears that at the level where such distinctions are made, the listener has access only to his phonemic classification of the vowel and not to the phonetic details that led to that classification. (Werker and Tees (1984) explore the possibility of a phonemic level of perception.)

Since Flege and Hillenbrand (1984) used only one L2 vowel (/u/) that was capable of being classed as equivalent to an English vowel, they were not aware of the interesting possibility that different listeners make different equivalence judgements. The present study found that some subjects could not perceive the contrast between English [i] and [ɨ], while others could not differentiate between [ɨ] and [e]. Since different patterns of equivalence classification would result in different inabilities to perceive L2 contrasts, we can infer that the former group classed English [ɨ] as equivalent to Spanish [i], while the latter group classed [ɨ] as equivalent to Spanish [e].

L1 vowel inventory

It is interesting that native Spanish bilingual listeners associated English [æ] with their L1 [a], whereas the native German listeners in Bohn and Flege (1990) associated [æ] with L1 [ɛ], even though German does possess an [a] phoneme. This difference again points to the influence of the L1 phoneme inventory on L2 perception. Since German [ɛ] is relatively close to English [æ], English [ɛ] and [æ] were perceived as a variants of the single L1 phoneme [ɛ]. But Spanish does not have an [ɛ] phoneme. The Spanish [e] area was apparently too distant to encompass the English [æ] sound. Instead, English [æ] was grouped with English [a] as variants of the L1 phoneme [a].

Differences in perception and production

Flege and Bohn (1989) found that when perceptual differentiation of L2 spectral contrasts is hindered by equivalence classification, L2 listeners differentiate vowels on the basis of duration, even when L1 does not use duration contrastively. The absence of duration differences in the stimuli of the present experiment may be responsible for the bilinguals' chaotic responses in English. Native GCE subjects in the pilot study showed clear vowel areas based on spectral information alone, as did both monolingual and bilingual subjects listening in Spanish in the main experiment. But native Spanish subjects listening in English were unable to identify the stimuli on the basis of spectral information due to equivalence classification, and had no durational or other cues to aid them in differentiating spectrally similar vowels.

Although the bilingual subjects were not always able to perceive differences in L2 vowels, they were capable of producing all the necessary L2 vowel contrasts both in reading and during the interview. Their spoken vowels were not identical to those of GCE, but two GCE speakers judged the vowels to be adequately differentiated from each other. It may be that the bilinguals create contrasts based on some other parameter such as duration, equivalent to their perception strategy. Since duration is a secondary cue to vowel contrasts in many forms of English, listeners could hear the appropriate distinctions on the basis of duration even in the absence of spectral contrast.

A more tantalizing possibility is that the bilinguals actually do produce the appropriate spectral contrasts without being able to perceive them. Strange and Dittman (1981), reporting on Japanese speakers learning the English /r/-/l/ distinction, and Tees and Werker (1984), reporting on English speakers learning Hindi retroflex /ʈ/, both found that perceptual difficulties persist even after L2 learners can produce the appropriate contrasts. Novel consonant segments are usually taught by demonstrating the articulation, not by addressing perceptual differences. There are certainly large numbers of phonetics students who have learned to produce a retroflex /ʈ/ without being able to distinguish it auditorily from a dental /t/.

Briere (1966) found that American subjects learning a pseudo-language made up of selected phonemes from Arabic, Vietnamese, and French learned to produce unfamiliar contrasts in both consonants and vowels before they could perceive them. "Subjects can produce the criterial attributes of a category without being perceptually aware that they are criterial." During the process of perception, subjects treat the same attributes as noise. It is worth noting that some of Briere's subjects did eventually master the perceptual task during the course of the experiment, unlike Tees and Werker's Hindi learners, who had been studying the language for a year at the time they were reported on. In the present experiment, the disjunction between L2 perception and production persists in subjects who have used L2 for up to 26 years.

Dialect researchers (Labov *et al.*, 1971; Harris, 1985) have found a similar disjunction when vowel distinctions merge, as has been observed with "source" and "sauce" in New York City, and "hock" and "hawk" in central Pennsylvania, among many other instances. At the final stage of a merger, some speakers reliably produce the vowel contrast but can no longer hear it. At some level the speaker must be monitoring his production, but he has no ability to identify the separate vowels on the basis of the sound pattern of his current dialect.

Harris suggests that the contrast must be lexicalized, since it appears in production, but that the contrast has no perceptual function. (He ventures no model of how perception takes place). He posits that the lexicalization occurs at an early age, when the child can make fine perceptual distinctions that are lost as he matures into the phonological system of his native language. This explanation cannot account for the disjunction in a second language learned later in life. Therefore the results of the present experiment are evidence against Harris' model.

Labov *et al.* think that the production contrast resides in low level output rules: "The abstract rule system of the language produces many features of the phonetic output which are not individually controlled or monitored for the direct contrast of meaning. This should not be surprising when we reflect on how completely and unconsciously a person learns his native "accent" -- a set of phonetic particulars which may be quite inaudible to himself and others in the process of communicating meaning." Applied to bilinguals, this theory says that the L2 vocabulary is stored in an L1 phonemic representation, but with some kind of diacritics telling the output system to modify the pronunciation during speech. The output process is akin to that of an actor using a foreign accent. The actor's phonemic representation of his native language does not change when he uses an accent; only the output is modified. A speaker who has some awareness of the sound system of a language can mimic that sound system, regardless of which language it is being imposed on. Perhaps the same strategy is used more pervasively (and less consciously) by L2 speakers with L1 lexical representations for their L2 vocabulary.

How then would perception operate? If the L2 listener decodes what he hears in the same way that he encodes what he says (for example, if a native Spanish listener hearing English [æ] changes it to [a] with a diacritic) before submitting it to his mental map for comparison, then we would see no disjunction between production and perception. We must assume that the diacritics are not stored in the lexicon but reside in the output mechanism. When the bilingual hears an L2 item, it is compared to the unmodified L1 "spelling" in the lexicon. In normal speech there is enough redundant information that a match can be found despite the disjunction. But isolated synthetic L2 vowels do not provide enough information to be identified accurately.

The results of the current experiment support such a model, with the added complication that there are varying degrees of bilingualism. For some bilinguals, particularly those who acquired their second language in childhood, some or all L2 phonemes are available in the mental lexicon. Thus an item in the L2 lexicon could contain a mixture of L2 phonemes requiring no modification at output and L1 phonemes with diacritics indicating output modifications.

SUMMARY

The results of this study support two earlier hypotheses about second language acquisition, that there is a criticial age after which a language cannot always be learned authentically and that equivalence classification interferes with acquisition of authentic L2 phonemes. The study also showed that the L1 vowel inventory determines the kinds of equivalences that can occur. No support was found for the notion that L2 acquisition affects L1 perceptual targets.

This study produced further evidence that distinctions can be produced even when they are not perceived . This finding is interesting both as it relates to second language acquisition and to dialect change. The literature of both fields points to several instances of this disjunction, but up to now there has been no adequate theory to explain it. This study offers an explanatory model: L2 vocabulary is stored in the lexicon in an L1 phonemic representation; production of L2 phones is accomplished by modification of the L1 phones during output. This is a worthy topic for further research.

A peripheral but important finding of this study is that F1 and F2, which are important tools in L1 vowel research, are less conclusive descriptors of L2 behavior.

References

Ainsworth, W.A. and J.B. Millar (1971). Methodology of experiments on the perception of synthesized vowels. *Language and Speech* 14(3):201-212.

Beckman, Mary E. (1986). *Stress and Non-Stress Accent*, Ch. 7, 179-202. Dordrecht, Foris.

Beddor, Patrice Speeter, and Sarah Hawkins (1990). The influence of spectral prominence on perceived vowel quality. *Journal of the Acoustical Society of America* 87(6):2684-2704.

Bohn, Ocke-Schwen, and James Emil Flege (1990). Interlingual identification and the role of foreign language experience in L2 vowel perception. *Applied Psycholinguistics* 11(3):303-328.

Briere, Eugene J. (1966). An investigation of phonological interference. *Language* 42(3):768-796.

Disner, Sandra F. (1983). Vowel Quality: the relation between universal and language-specific factors. *Working Papers in Phonetics* 58. UCLA Phonetics Laboratory, Los Angeles.

Elman, Jeffrey L., Randy L. Diehl, and Susan E. Buchwald (1977). Perceptual switching in bilinguals. *Journal of the Acoustical Society of America* 62(2):971-974.

Flanagan, James L. (1955). A difference limen for vowel formant frequency. *Journal of the Acoustical Society of America* 27(3):613-617

Flege, James Emil (1987). The production of "new" and "similar" phones in a foreign language: evidence for the effect of equivalence classification. *Journal of Phonetics* 15:47-65.

Flege, James Emil, and Ocke-Schwen Bohn (1989). The perception of English vowels by native speakers of Spanish. Paper presented at the 117th meeting of the Acoustical Society of America, Syracuse, N.Y.

Flege, James Emil, and James Hillenbrand (1984). Limits on phonetic accuracy in foreign language speech production. *Journal of the Acoustical Society of America* 76(3):708-721.

Fox, Robert Allen (1982). Individual variation in the perception of vowels: implications for a perception-production link. *Phonetica* 39:1-22.

Fox, Robert Allen (1984). Auditory contrast and speaker quality variation in vowel production. *Journal of the Acoustical Society of America* 77(4):1552-1559.

Godinez, Manuel (1984). Chicano English Phonology: Norms vs. Interference Phenomena. In Jacob Ornstein-Galicia (ed.) *Form and Function in Chicano English*, Ch. 4, pp. 42-48. Rowley, Massachusetts: Newbury House Publishers, Inc.

Godinez, Manuel, and Ian Maddieson (1985). Vowel differences between Chicano and General California English? *International Journal of Sociology and Language,* 1985:43-58.

Harris, John (1985). *Phonological Varistion and Change. Studies in Hiberno-English*, Ch. 5, pp. 297-340. Cambridge University Press

Johnson, Keith (1989). On the perceptual representation of vowel categories. Paper presented at the 118th meeting of the Acoustical Society of America, St. Louis.

Johnson, Keith (1990). An articulatory analysis of formant asymmetries in symmetric CVC sequences. Paper presented at the 120th meeting of the Acoustical Society of America, San Diego.

Labov, William, Malcah Yaeger, Richard Steiner (1972). *A Quantitative Study of Sound Change in Progress,* Ch. 6, pp. 229-257 University of Pennsylvania. (Printed and Distributed by The U.S. Regional Survey, Philadelphia.)

Miller, James D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America* 85(5):2114-2134.

Munro, Murray (1990). Attention to spectral and temporal cues in vowel perception among native speakers of Arabic and English. Paper presented at the 120th meeting of the Acoustical Society of America, San Diego.

Nearey, Terrance M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America* 85(5):2088-2113.

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

Peterson, Gordon E., and Ilse Lehiste (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America* 32(6):693-703.

Stevens, K.N., A.M. Liberman, M. Studdert-Kennedy, and S.E.G. Ohman (1969). Crosslanguage study of vowel perception. *Language and Speech* 12(1):1-23.

Strange, Winifred (1989). Evolving theories of vowel perception. *Journal of the Acoustical Society of America* 85(5):2081-2087.

Strange, Winifred, and S. Dittman (1981). Effect of discrimination training on the perception of /r/-/l/ by Japanese adults. Paper presented at meetings of the Psychonomic Society, Philadelphia.

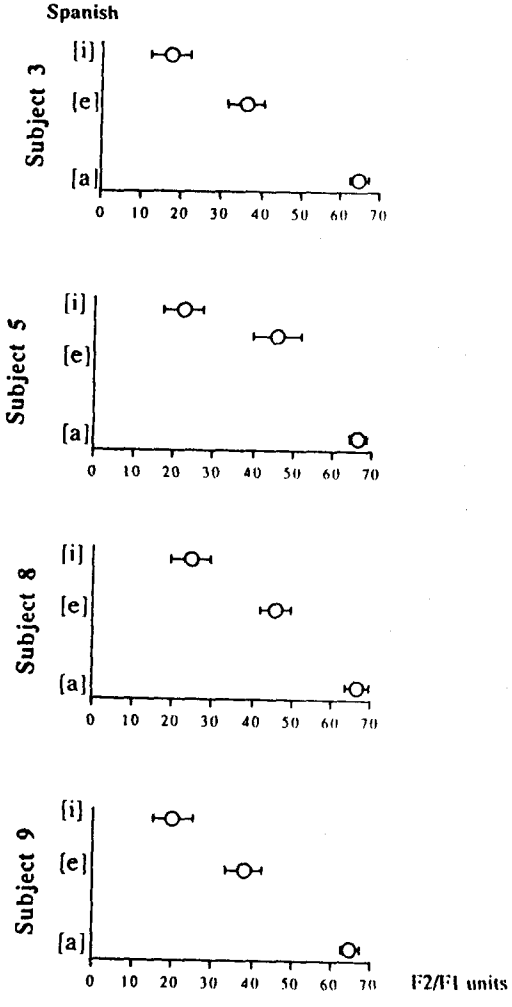
Tees, Richard C., and Janet F. Werker (1984). Perceptual Flexibility: Maintenance or Recovery of the Ability to Discriminate Non-Native Speech Sounds. *Canadian Journal of Psychology* 38(4):579-590.

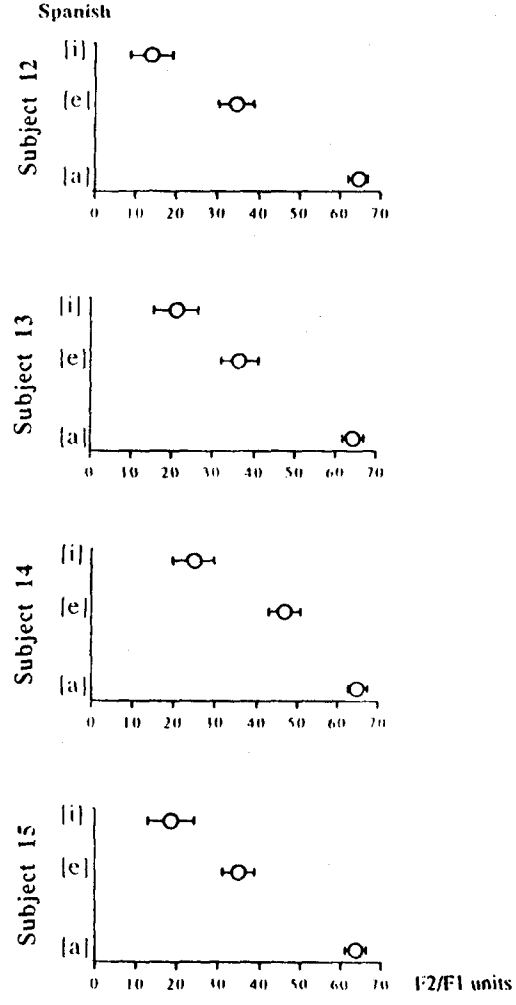
Terbeek, Dale (1977). A cross-language multidimensional scaling study of vowel perception. *Working Papers in Phonetics* 37. UCLA Phonetics Laboratory, Los Angeles.

Werker, Janet F., and Richard C. Tees (1984). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America* 75(6):1866-1875.

Willis, Clodius (1971). Synthetic vowel categorization and dialectology. *Language and Speech* 14(3):213-228.

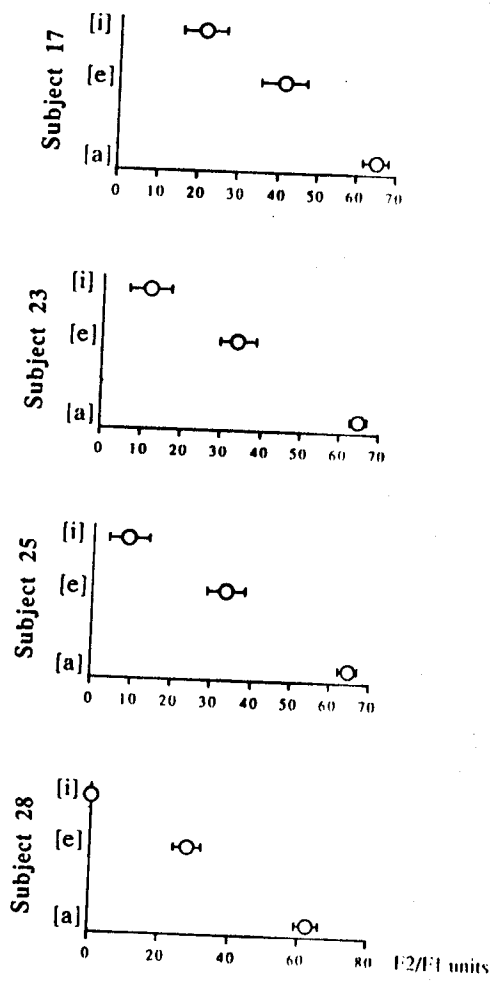
Appendix
F2/F1 mean and standard deviation, monolingual subjects





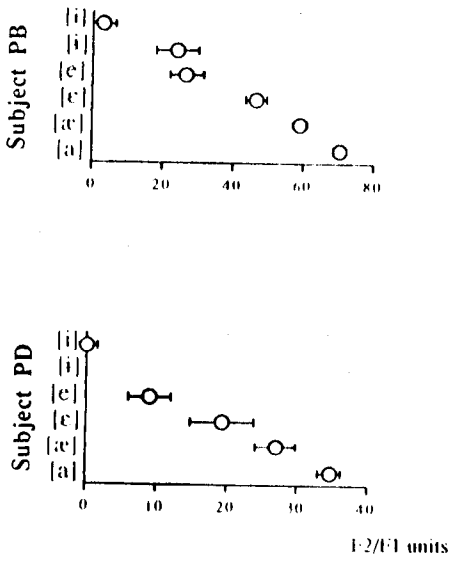
Appendix, continued
F2/F1 mean and standard deviation, monolingual subjects

Spanish

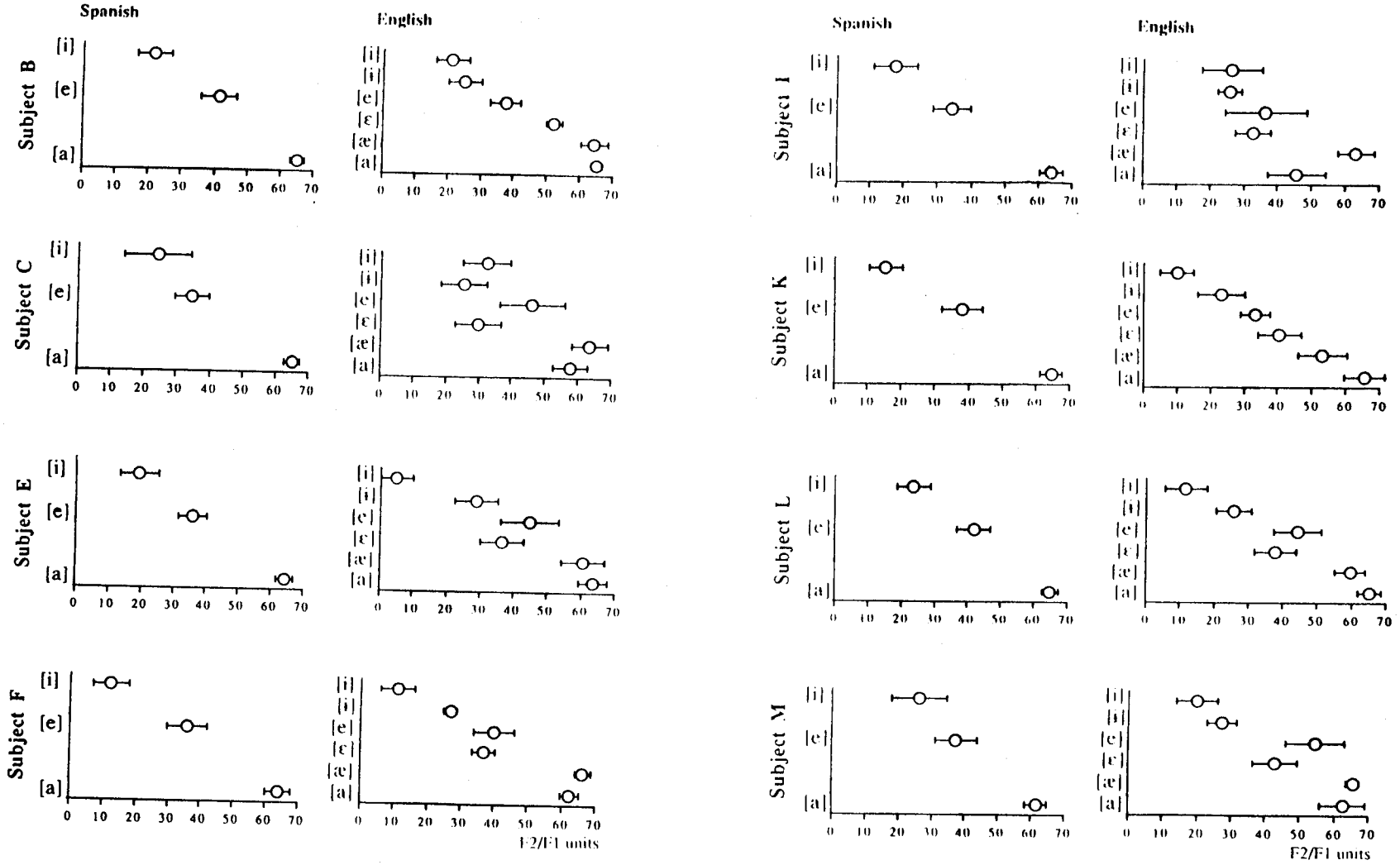


F2/F1 mean and standard deviation, native English speakers

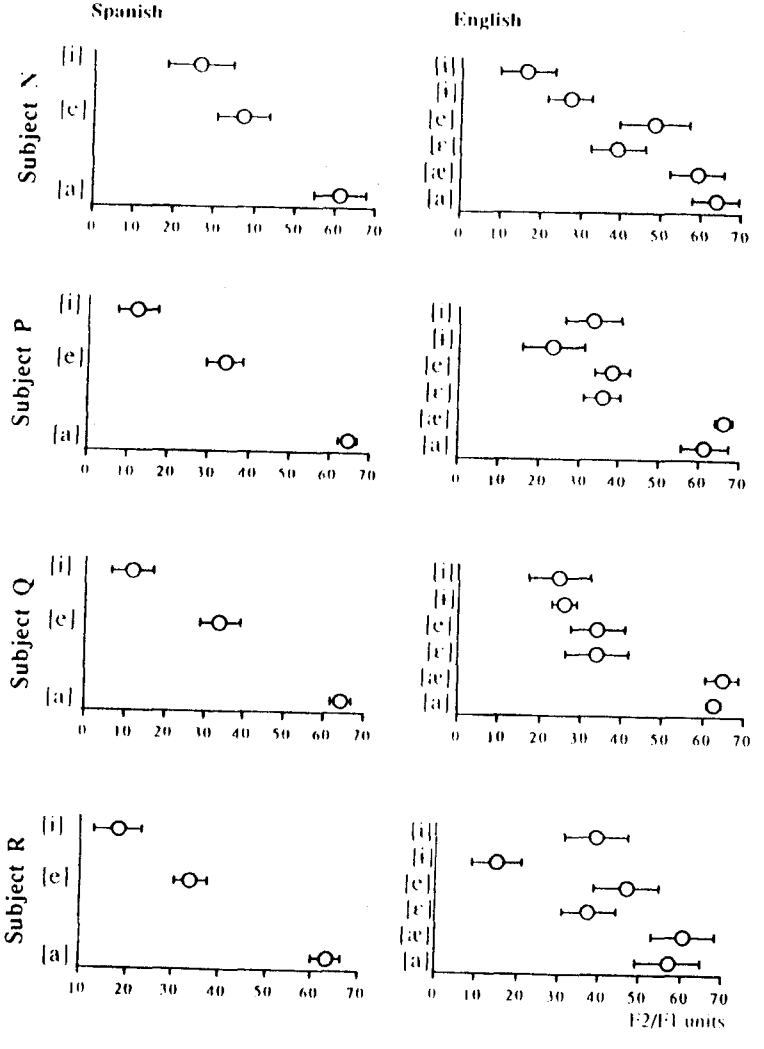
English



Appendix, continued
F2/F1 mean and standard deviation, bilingual subjects



Appendix, continued
F2/F1 mean and standard deviation, bilingual subjects



Units of intonation in discourse: Acoustic and auditory analyses in contrast

Stephan Schuetze-Coburn, Marian Shapley & Elizabeth G. Weber

Introduction

Intonation is in many respects a very well studied phenomenon. Detailed accounts of intonation — at least of English intonation — can be found from a variety of descriptive and theoretical viewpoints. At the same time, however, there is much that we do not know; intonation is in some sense the "least commented-on side of vocal communication" (Bolinger, 1986b, p. vii). One broad area where our understanding of intonational phenomena is particularly deficient is in natural discourse. As Altenberg (1987, p. 11) observes regarding English prosody, "what we know — or think we know — is often poorly supported by empirical data". The complexity of intonation necessitates simplification of the raw physical data to get at the system. But simplification is a form of abstraction, and although abstraction is a necessary step in the organization of the data, the ever-present danger of overabstraction whereby the constructed system acquires an inertia of its own must be vigilantly checked. One way to monitor this is to maintain a close tie to language in actual use. In this spirit, our orientation is decidedly empirical: In this study we examine a corpus of spontaneous interactional speech.

It is often useful to examine a phenomenon from different points of view. Intonation, with its many dimensions, offers the opportunity to break away from a perspective which makes use of only one type of data. In empirical studies of intonation, at least four data types can be identified: articulatory-proprioceptive, acoustic, auditory-perceptual, and perceptual-experimental. Here, we focus on the acoustic and the auditory(-perceptual). In addition, data sources differ. Data may consist of constructed phrases or sentences, or utterances taken from a corpus, with the prosody supplied by introspection, a reading of the prepared material, or a recording of the discourse. The data type and source together generally limit the scope of the investigation, as is evident from a brief review of the literature.

The use of auditory data

Descriptions of English intonation based primarily on auditory data include the early studies of Jones (1909), Palmer (1922), Armstrong and Ward (1926), as well as those of Kingdon (1958) and Bolinger (1964, 1972). This body of research describes intonational patterns in some preformed syntactic unit, usually clause or sentence, or occasionally sentence pair. In recent years, auditory investigations of intonation in longer periods of speech have become familiar. In such studies, the speech stream is segmented into prosodic units ('tone units', 'tone groups', 'intonation units', etc.), identified by various perceptual features, not all of which are necessarily limited to or motivated by prosody. Works of this kind include, for example, Quirk et al. (1964), Halliday (1967), Crystal (1975), and Svartvik and Quirk (1980).

Discourse-oriented researchers have relied mainly on the auditory analysis of intonation, due in part to the sheer volume of material that must be analyzed, and in part to numerous other limitations, such as poor audio quality of natural recordings, inadequate availability of instrumentation, etc., which make acoustic analysis burdensome or impossible. Conversational data pose additional problems, including speaker overlap, unidentifiable speakers, and parallel dialogue. Despite these hurdles, instrumental analysis of discourse data is possible, as Menn and Boyce (1982) have shown.

The use of acoustic data

Research on intonation in English which is based on fundamental frequency data includes that of Liberman and Sag (1974), Maeda (1974), O'Shaughnessy (1976), Pierrehumbert (1980), Cooper and Sorensen (1981), and Liberman and Pierrehumbert (1984). In these studies, the speech analyzed consists of simple and complex prepared sentences or paragraphs which are read aloud by native speakers. The scope of the intonational analysis is therefore delimited, at least implicitly, by syntax. In comparison, work by Willems (1982), de Pijper (1983), and 't Hart, Collier, and Cohen (1990), like that of some discourse analysts, is aimed first at deriving phonetic units, which may be related to grammatical units at a later point.

The use of combined approaches

Occasionally, both auditory and acoustic data figure in an analysis. Auditorily based evaluations have been used to verify the perceptual relevance of acoustic data, as Cohen and 't Hart (1967) and 't Hart, Collier, and Cohen (1990) do in their technique of analysis by synthesis, whereby simplified fundamental frequency (F_0) curves are judged against the (resynthesized) original utterances. Conversely, acoustic data are sometimes used by discourse researchers as a means of verifying auditory judgments. For example, Brazil (1978, p. 8), who compares F_0 traces with auditorily based transcriptions, reports that "both the analytical categories and the expectation of F_0 correlates seemed, in general terms, to be confirmed". Brown, Currie, and Kenworthy (1980) employ a balanced combination of auditory and acoustic analyses in their intonation studies of Edinburgh Scottish English. But here and elsewhere, there is no indication that a systematic comparison has been made. Indeed, systematic comparisons between acoustic and auditory data — whether for prepared sentences or spontaneous speech — are conspicuously absent from the literature.

In sum, aside from the earliest works, which are of course auditorily based, linguists investigating intonational structure in texts or long stretches of spontaneous speech have almost exclusively relied on auditory analysis, with a type of tone unit as the basic domain, while those analyzing short segments of prepared speech (typically assuming a syntactic domain of one or two sentences) have tended to rely on instrumental analysis. And in general, systematic acoustic measurement has accompanied studies of speech which is usually read aloud, while systematic auditory analysis has been characteristic of studies of narrative, conversational, or other unplanned talk.

The orientation of this study and its relevance to intonation research

While both the acoustic and the auditory approaches have been fruitful, there is as yet little detailed information on the relationship between the perception of auditory pitch units and their acoustic correlates, especially for larger periods of connected speech in a natural discourse situation. Although some empirical studies of intonation have used both auditory and acoustic data, one type in any given study always has ancillary importance. Here we wish to break from this tradition. Our intent is to treat both acoustic and auditory data on equal terms, not to take one kind of data as primary and use the other as a method of verification or corroboration, as has been done in the past. In this study, we perform two independent, parallel prosodic segmentations on a corpus of spontaneous conversational speech. We examine the relationship between an acoustic unit defined in terms of fundamental frequency declination and an auditory unit roughly the size of an intonational phrase, based on parameters which contribute to perceived prosodic coherence.

We hope to establish a connection between the two kinds of units, such that one could be expressed in terms of the other. If the auditory and acoustic units consistently coincide, that is share the same domain, then standard assumptions and practices by both auditory and acoustic researchers would be confirmed. On the auditory side, such agreement would lend validity to the use of auditory methods of analysis for discourse data, as it could be inferred that acoustic cues of intonational structure can be reliable perceived. The use of a phrase-level unit as the primary do-

main of intonational in the analysis of discourse would also be supported. On the acoustic side, such agreement would prove the applicability of acoustic models to discourse studies and underscore the helpfulness of the auditory dimension as input to acoustic analysis. If the units do not coincide in any consistent way, then a closer look at the methods and components of both types of modeling of intonation would be called for. Disagreement would imply that a phrase-level unit would be suspect as the basic intonational unit of discourse analysis. The representation of discourse intonation by a series of acoustically constructed declination units would be inadequate. Auditorily derived units would also be questionable as a source of units for acoustic modeling.

In the following sections, the viability of comparing auditory and acoustic data is briefly reviewed, and our specific hypothesis is given. We then explain the methodology of the current study and present the results. Finally, the implications and significance of our findings are discussed with reference to the intonational analysis of discourse and phonological accounts of declination.

Acoustics and perception of intonation

This study examines the coincidence of acoustic and auditory units in discourse and some of the implications of this coincidence. However, we should not assume without comment that acoustic and auditory measures of intonational phenomena can be directly compared. The exact relationship between F_0 and pitch is controversial. As many researchers exhort, "pitch perception is not to be equated with the perception of fundamental frequency of a periodic or quasiperiodic acoustic signal" (Krause, 1984, p. 243). Even though perceived pitch is related to F_0 in an approximately linear fashion at frequencies below 1000 Hz — well within the normal range of F_0 in human speech (Ladefoged, 1962) — the effect of other acoustic parameters on the perception of pitch is complex. Nevertheless, F_0 is usually taken to be a reasonable indicator of pitch, as experimental studies have shown that speakers compensate for nonlinguistic pitch perturbations (cf. 't Hart, Collier, and Cohen, 1990, Ch. 2). Here we will ignore the effects of such variation.

Given that F_0 contours and pitch movements can be compared, we expect, on the whole, that acoustic units and auditory units should align, rather than be mismatched. Specifically, we hypothesize that the initial boundaries of the auditory units will coincide with the initial boundaries of the acoustic units.

While the selection of the two units of this investigation was motivated not by any designed compatibility of the units, but by their relevance for acoustic or auditory analyses, our hypothesis is nevertheless strongly supported by the fact that the auditory units are determined partly on the perception of the features defining the acoustic units, i.e. the change in the rate of declination of F_0 over time and the presence of associated pauses. The first of the shared features is called 'reset', with F_0 declination lines (slopes describing the gradual fall in F_0 over time during a period of speech) reset acoustically and pitch reset auditorily. Although global pitch trends are not explicitly represented in our auditory analysis (as they are in some auditory analyses, e.g. Selting 1987), pitch reset — which we take to be the primary perceptual correlate of F_0 reset — is an important cue in the identification of the auditorily based units. The second prosodic feature the two units share is an aspect of timing. In both analyses, pauses in the speech stream contribute to the demarcation of unit boundaries, as will be described below.

Furthermore, as conversational data are involved, the situational variable of speaker change constitutes a de facto third shared feature. In practice, all new turns begin new auditory units. Interspeaker differences in voice quality, along with the regularly occurring prosodic features (for example, utterances typically begin on a higher pitch than they end; cf. Crystal, 1969), virtually guarantee the perception of new auditory units at turn boundaries. While there was no direct way of representing speaker change acoustically, we had reason to believe that turns affected acoustic unit boundaries similarly. The physiological mechanisms involved in F_0 production alone (cf. Lieberman and Blumstein, 1988) suggest that declination reset will normally occur at the beginning

of each turn. Obviously, if this is the case, the coincidence of the two units is assuredly high whenever a speaker starts talking.

Indirect support for boundary alignment can also be gleaned from the phonological literature. In many accounts of intonation, declination is said to operate within a clausal domain, corresponding prosodically to an intonational phrase. Thus we would expect declination lines to originate at intonational phrase boundaries.

On the other hand, some doubts about establishing a correspondence may be harbored due to reports in the literature regarding the difficulty of relating acoustic and auditory measures. While all researchers logically assume a connection between the physical speech signal and its perception, Brown, Currie, and Kenworthy (1980), for instance, were not able to identify regular correlates in the acoustic record of their tone group data obtained by ear. Our study, then, provides a controlled arena for testing assumptions and claims concerning the perception of intonation, the viability of auditory analysis, and the domains of intonational phenomena.

Methods

The data consisted of excerpts from multiperson conversations of the 'dinner table' type ranging from about one to two hours in length. In each case the speakers were told they would be taped. Excerpts used in this study were taken from the middle of the conversations and included one or more long turns of a story plus pre- and post-story exchanges of short turns, totaling about sixteen minutes of elapsed time. The selections thus contained speech from a number of participants and exhibited a wide range in turn length, from single word utterances to extended narrative passages. The conversations were recorded on consumer-quality cassette recorders in non-laboratory environments.

The selected excerpts were transcribed and segmented acoustically and auditorily. One author was responsible for the acoustic units (Shapley, 1989), while the other two authors segmented the conversations into auditorily based intonation units from the tapes. These two segmentations were arrived at independently; the judgments of intonation unit boundaries were made without the help of the instrumental data, and the acoustic unit boundaries were determined without knowledge of the perceptual analysis. A subsequent division of the conversations into turn units was necessary in order to ascertain the effect of speaker change. In the following sections the methods of prosodic segmentation will be described.

Acoustic analysis: Declination Units (DUs)

The acoustic analysis derived declination units or DUs, formed of periods of speech in which F_0 measures, plotted over time, shared a common declination line. The concept of declination, or gradual falling off of pitch during an utterance, was noted by Pike (1945) and Bolinger (1964) in perceptual data. Cohen and 't Hart (1965) gave the concept its current term in their acoustic studies. Since then, declination has been noted in a number of languages (cf. Bolinger, 1986a) and is a part of many models of intonation based on acoustic data (see 't Hart, Collier, and Cohen (1990, Ch. 5) for an excellent overview).

A digital sound spectrograph was used for the F_0 analysis. F_0 values were read off from narrow-band spectrograms using the tenth harmonic, or the closest clearly countable harmonic to the tenth available. (In addition, some of the F_0 points were checked computationally using a temporal structure analysis to determine the time delay from one glottal pulse to the next. We used *Signalyze*, Macintosh software written by Eric Keller, InfoSignal, Inc., for this purpose.) For every tenth of a second, an F_0 value or the occurrence of a pause or laughter was recorded. Wide-band spectrograms aided in the matching of words to F_0 values. The F_0 values in Hz for each speaker were first converted to semitone values and then to normalized Z-scores.[1] The results

were plotted as a function of time. The use of normalized values allows the pitch of different speakers to be directly compared in terms of variation from each speaker's mean, regardless of the speakers' individual pitch ranges. A semitone scale was used because the distribution of the logarithmic values was more nearly normal than that of the Hz values, increasing the validity of the normalized scores.

The resulting ragged plots of points were then stylized into simplified straight-line representations, and the DUs were delimited. DU boundary identification was straightforward for about 85% of the units, often because of the cooccurring pauses which separated the F_0 points. DUs were defined by segmenting the stylized plots into units sharing a common declination line. An envelope or grid in the shape of a parallelogram was superimposed on the points. The bottom line of the envelope connected low points, the top line connected the peaks, and a midline marked a middle level of pitch. The end boundary of a DU was located where the F_0 reached the speaker's lower bound, the bottom line of the parallelogram (the declination line) reached such a bound, or it was no longer possible to include high points in the envelope. This model was patterned after the work of Willems (1982). It is neither strictly a top-line model nor a bottom-line model, but a tipping or declination of the entire parallelogram containing the F_0 points. It was chosen because of its simplicity, because of its ability to group the data in the absence of a peak at the beginning of the unit (or in the absence of a low point at the end of the unit), and because it fit the data satisfactorily.

DUs were sensitive to the time scale. In fact, they are time based, because the slope of the declination line is a function of the length of the unit (see, for example, Cooper and Sorensen, 1981; Bruce, 1982; de Pijper, 1983; Thorsen, 1983; and 't Hart, Collier, and Cohen, 1990). Thus, the length of a DU was partly determined by the slope of the declination line. A pause following a sequence of F_0 points was often an indication that an end boundary should be drawn; to include points after the pause would have resulted in a slope uncharacteristically steep for a longer unit.

In the absence of any defining peaks, the height of the parallelogram (the height of the envelope) was taken to be ± 1 standard deviation from the midline, as this was the predominant range in the cases where defining peaks occurred. The height of the parallelograms (i.e. the range of variation in pitch) was fairly consistent for the great majority of units in the middle range, but varied with extremes of pitch, with the units having higher pitched declination lines also having the greater variation in height, and those with lower pitch having less variation. Outliers were disregarded, and in general every attempt was made to make the data fit into the typical envelopes for length and range of F_0. Although there were numerous cases where the procedure had to be carefully applied, few cases remained uncertain in the end.

In Figure 1, F_0 data of an excerpt from one of the conversations, plotted and stylized as described above, are shown to illustrate the derivation of DUs. The ordinate in the figure is in the normalized scale of semitone values; the abscissa is time. The corresponding text is given in (1) below.[2]

The transcript is arranged so that the text of each DU occupies its own line, with capitalization of the first word in the line an additional marker of a new acoustic unit. The relevant lines in (1) are labeled sequentially from (a) to (m) and matched to the DUs in Figure 1. Utterances for which there are no F_0 data — overlapped speech and low-volume backchannels — are left unlabeled.

In Figure 1, each envelope surrounding a set of F_0 points represents a separate DU. The beginnings and ends of DUs are indicated by the perpendicular sides of the parallelograms. For most of the DUs in the figure, the characteristic F_0 reset is noticeable immediately upon inspection in that the beginning edge of a parallelogram is shifted upwards with respect to the end of the previous one. This is true for DUs (c-f) and (h-m). The exceptions are DUs (b) and (g). Visually,

the envelope for DU (b) appears to constitute a continuation of (a). But in this case, a pause

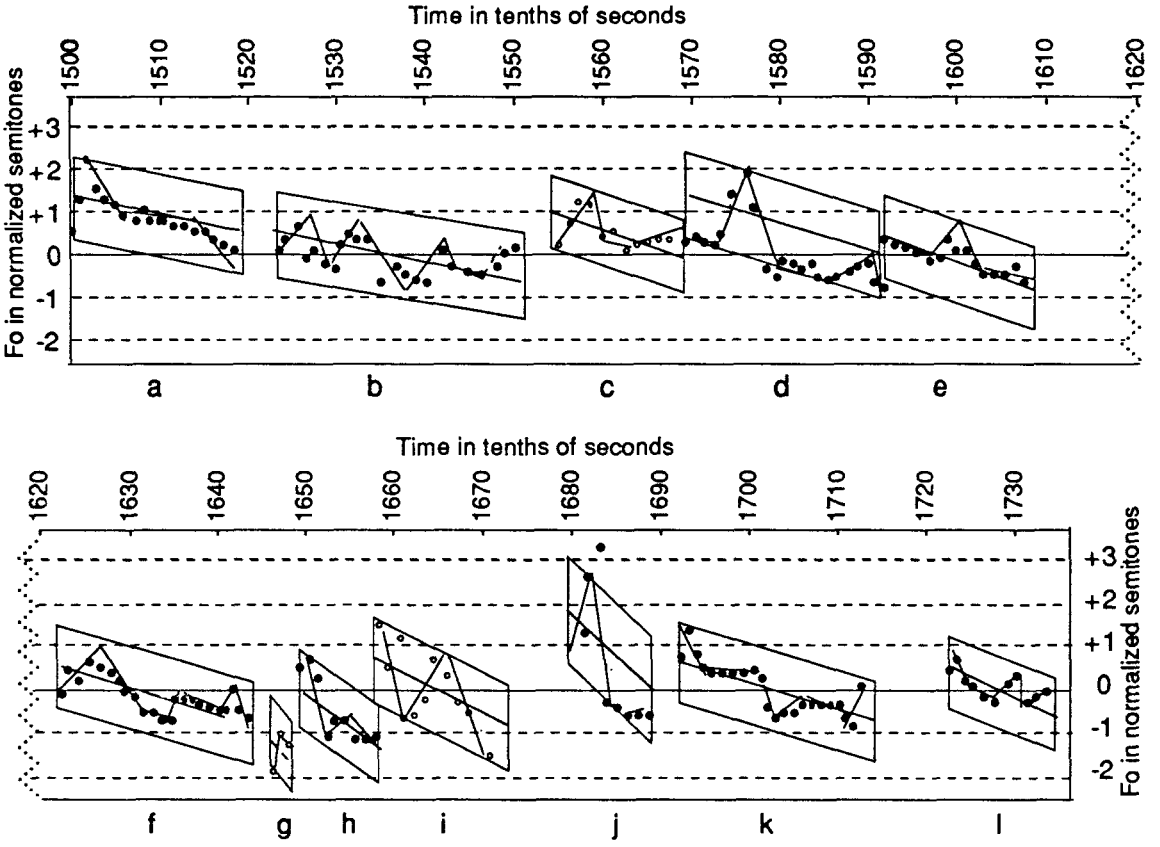


Figure 1. Derivation of DUs from the plotted and stylized F_0 data.

(1) (a) P: .. Uh nó Í- Í .. I gá- I méan
(b) ... Í- I assúmed- I was [rélatively] cálm in the sénse that I [[figured
S: [«yes»]
N: [[@@@@ (H)]]
(c) C: Áfter an hóur]] [and a hálf?
(d) P: [Thát á=fter a r- .. a réasonably shórt período
(e) I mean befó=re the tóes begin to frée=ze
X: @@@@
(f) P: (H) That sómebody would appéar .. I mean from ínsíde
S: .. yeah
(g) N: .. Mhm
(h) P: [I méan that-
S: [«it's»
(i) N: [[‹F Théy'd gét úp F›
S: [[«he fígures he's»
(j) Léft a way óut ányway=
... of course
(k) P: .. It wás the way óut
(l) .. But [it túrned óut thére was anóther way out .. that I didn't ‹@ réalize @›
X: [@@
P: [[@@@
S: [[oh
(m) P: [‹@ I shóuld've been «lóoking» óut @›

follows after the F0 points of (a) reach the baseline of the parallelogram. This configuration results in (b) being identified as a separate DU. (Also, DU (b) is subordinate to DU (a). We do not address DU subordination directly in this paper. For a detailed discussion, see Shapley 1989). Finally, DU (g) illustrates a very short backchannel utterance for which there are minimal F0 data. In this instance, the change in speaker determines the DU boundary, as the apparent continuation is most likely an artifact of the normalization process.

Auditory analysis: Intonation units (IUs)

The auditory analysis defined a prosodic unit which we call intonation unit, or IU, after Chafe (1987). Following Du Bois, Schuetze-Coburn, Paolino, and Cumming (1991), the speech stream was exhaustively segmented into IUs from the tapes by auditory means. In basic terms, an IU consists of a stretch of speech by a single speaker uttered with a "coherent intonation contour" (Chafe, 1987, p. 22). Although the exact perceptual correlates of this notion remain somewhat elusive, numerous prosodic cues have been identified which are used to determine IU boundaries (cf. Crystal, 1969, pp. 204-207; Cruttenden, 1986, pp. 35-42; Du Bois et al., forthcoming, Ch. 22). The cues are of two main types: those concerned with the pitch pattern of the utterance and those related to the timing. The perception of coherence in the pitch pattern is influenced at least by two factors: degree and direction of pitch movement on a stressed syllable and change in pitch relative to the speaker's preceding utterance (pitch reset). Timing cues which contribute to perceived IU unity include an acceleration in tempo on initial unstressed syllables, prosodic lengthening of the final syllable(s), and a noticeable pause (0.3 second or greater) between IUs. Aspects of voice quality, such as laryngealization, also play some role. Not all prosodic features are individually notated in the transcript (but they are taken into account during the auditory segmentation). Instead, each boundary marker represents the sum of the cues for locating the boundary at that point. IU boundaries are explicitly marked by a combination of punctuation (',' or '.' or '?') and diagonal line ('/' or '\'). The lines indicate perceived terminal pitch direction, and punctuation indicates 'transitional continuity' class, that is, the equivalence class of pitch contours expressing "the degree of continuity which occurs at the transition point between one intonation unit and the next" (Du Bois et al., forthcoming, Ch. 6). Comma signifies the class of 'continuing' contours; period signifies 'finality', and question mark 'appeal'. Regarding terminal pitch direction, a slash signifies a nonfall (rise or level), a backslash a fall. Truncated (uncompleted) IUs are indicated with a double hyphen ('--').

The relative importance of the cues may differ — pitch reset, for example, is arguably more central than tempo modulation — but none alone defines an IU boundary per se; rather, a conjuncture of cues is usually required for an IU to be perceived. One can say that the prototypical IU exhibits all of these cues, yet seldom are all actually present in any given instance. That is, most IUs deviate from the prototype to some degree. Thus, a given IU may exhibit pitch reset and a definite contour, but none of the other features. In practice, our IU is similar to the 'tone group' or 'tone unit' of other researchers (e.g. Halliday, 1967; Crystal, 1969; Svartvik and Quirk, 1980). Altenberg (1987, p. 47) describes the tone unit prosodically as "a coherent intonation contour optionally bounded by a pause and containing (among other things) a salient pitch movement (the nucleus), normally at the end of the unit". IUs, however, need not contain one or more prominent ('accented') syllables (cf. Chafe, 1991), as do units based on the presence of a 'nucleus'. Short responses or backchannel utterances, especially when low pitched, frequently have no prominence, but are nevertheless considered IUs. The DU (g) discussed above in (1) — being also an IU — is one such example. Uncompleted IUs are also considered separate units, i.e. they are not treated as 'residue' which can be incorporated into some other IU or ignored.

Nonprosodic factors

Turns. An interactional analysis of the conversational excerpts derived speaker turns. We identified turn boundaries from the audio tapes and classified each prosodic unit as beginning a

turn, continuing a turn, or as constituting a backchannel utterance. Of course, the determination of speaker turns is by no means a straightforward process. However, we cannot go into the details of the procedure here; for a discussion of the parameters involved, see Sacks, Schegloff, and Jefferson (1974) and Oreström (1983). In outline, a new turn began whenever a speaker attempted to gain the floor; backchannel utterances were considered utterances made without such intention. As previously noted, any change of speaker was expected to result in both acoustic and auditory reset — and thus a cooccurrence of DU and IU boundaries — so it was necessary to be able to consider the turn-initial and turn-medial prosodic boundaries separately.

In addition, the relative frequency of DU-IU boundary alignment was compared for short and long turns, as we suspected that turn length might unduly influence the correspondence of prosodic boundaries. Due to the interactional nature of the talk, our data contained many short exchanges between speakers. Yet if short turns predominated in our excerpts — turns which due to their length alone were likely to consist of exactly one DU and one IU — the results of a blind comparison of DU and IU boundaries would be biased in favor of alignment. To check the extent of this bias, we assessed the effect of turn length. The question arose as to how turn length should be calculated, i.e. whether by absolute duration, number of syllables, or some other measure. As it would be difficult to determine what constituted a 'short' turn in terms of seconds or syllables, we opted for tabulation in terms of DUs per turn and IUs per turn. A short turn was thus naturally one containing the fewest number of DUs or IUs. Occurrence of boundary alignment according to turn length was then computed.

Syntax. The connecting thread between the acoustic and the auditory data was, of course, the text: The words of the text were used to line up the DUs with the IUs. However, as we wanted to focus on the prosody of the conversational excerpts, effort was made to minimize the effect of syntactic structure on the prosodic segmentations. In the acoustic analysis, the text was matched with the F_0 plots for the most part after DUs had been delimited. In identifying IUs, syntactic boundaries were also disregarded, to the extent that this is possible using an unfiltered signal. Experimental work has consistently shown that prosodic judgements are 'distorted' to some extent when the segmental information is included. Nevertheless, there are numerous reasons why we choose to analyze the whole speech signal. First of all, the evaluation of some prosodic features (e.g. final lengthening) simply requires a linguistic context. Secondly, the effect of syntax is not uniform (syntactic boundaries vary in their strength), so that one can consciously guard against common analytical pitfalls (e.g. 'hearing' a prosodic boundary at every clause juncture). Thirdly, the design of the analytical system has to be taken into consideration: The portions of intonation contours relevant for this study are specified in terms of binary distinctions ('final' vs. 'nonfinal', 'fall' vs. 'nonfall'), which increases the reliability of the auditory judgements. Finally, because the analyst looks for a cluster of prosodic factors, IU boundary identification is also adequately consistent (Schuetze-Coburn, in progress). Syntactic structure contributes (at most) one additional component, thus it is unlikely that any given boundary is *determined* solely by syntax. Estimating the role of syntax is important because it is often claimed (or more commonly, assumed) that syntactic structure aligns with — or determines — prosodic structure. While in fact a regular correspondence between syntax and DUs has been ascertained in discourse data (Shapley, 1989), as has a correspondence between syntax and auditorily based prosodic units (Altenberg, 1987), we will only briefly touch on the role of syntax here, since we discuss its relationship to both DUs and IUs in another paper (Schuetze-Coburn and Shapley, forthcoming).

Comparison of acoustic and auditory units

Following the complete segmentation of the texts into DUs and IUs, the rate of coincidence between the two prosodic units was calculated. For methodological reasons we focused on the coincidence of initial boundaries. In principle, it should not matter whether the beginning or the end of a unit is noted, as the end of one unit simultaneously locates the start of the next. For the auditory analysis, this proved to be true: The final boundary of one IU was invariably the beginning boundary of the next (assuming pauses are ignored; boundary pauses were, however, consid-

ered to occur *between* units). However, for the acoustic data, it was slightly preferable to use initial boundaries. During the course of an utterance, the amplitude of the speech sometimes trailed off near the end of the unit, making the final boundary less easy to identify. Furthermore, in many cases where the speech of two speakers overlapped, it was possible to discern the initial unit boundary of the second speaker, due to the strength of the interrupting signal, whereas the ending boundary of the first speaker had to be inferred. (An example of such overlap can be found in 1c-d.) In addition, the decision to use initial boundaries was influenced by the requirements of the syntactic analysis in Schuetze-Coburn and Shapley (forthcoming). In order to maintain future compatibility, we choose to compare beginning prosodic boundaries in the present paper. Thus, all comparisons discussed below will be described in terms of the coincidence of unit beginnings.

As the acoustic analysis depended on the measurement of the physical parameter (F_0), the comparison of acoustic and auditory units — the focus of this paper — also depended on its availability. Consequently, IUs for which we had no corresponding F_0 readings were excluded from the counts. These included 84 nonsegmental units (primarily laughter, coughing, etc.), 48 units with inaudible portions, and 160 other units audible, but without F_0 values. Of these 160 units, 69 contained overlapping speech, 56 were auditorily marked as having very low pitch or volume, and of the remaining 35 units, 27 consisted of very short (1-2 syllable) utterances. Example 2 below shows two instances of omitted lines.

(2) K: .. Sínce then it's been bólted ... clósed, \ éver since. \ .. but uh --
...(.8) (TSK) The nèxt time thèy came in, /
X: ... ‹P «when was that» P›? /
K: ... [‹@ «Thèy can tell that» @› we bólted the dóo=r, \ (H) uh=, /
X: [‹p «XX» p› [R237-244]

The third and fifth lines of (2) show utterances made by an undetermined speaker. Both stretches of speech were very low in volume (thus bracketed by '‹P P›'). While the first utterance was audible and perceived as a complete IU, it did not show up on the acoustic record. The fifth line was auditorily inaudible and acoustically unmeasureable. Both lines were eliminated from the present data for the purposes of tabulation.

Results

The conversational excerpts were segmented into 455 DUs and 807 IUs for a total of 1262 prosodic unit boundaries. Overall, the size of both units varied considerably. A general characterization of their duration is as follows. DUs averaged 1.6 seconds in length, with a range of 0.1 to 6.0 seconds. IUs averaged 0.7 second and ranged from 0.1 to 3.5 seconds. DUs were also longer in terms of syllables, averaging 8.4 syllables, compared to 4.7 syllables for IUs. A short excerpt from the analysis which indicates the variation encountered is presented in (3). Recall that each line corresponds to a single DU, with IU boundaries indicated by a punctuation mark and slash symbol complex.

(3) (a) K: ... So hé went úp, / and tóok a náp, \ ... and wóke úp, \
(b) ... And lóoked úp, / .. to sée what tíme it was, \ and sáid, / ...(.8) "hm". \
(c) L: ...(2.0) "The clóck [radio is góne". \
(d) K: [The clóck radio isn't thére. \ [R375-385]

Thus (3) illustrates a sequence of four DUs uttered by two speakers. The first DU contains three IUs; the second, four IUs; the third, one IU and the fourth, again one IU. The acoustic data are shown plotted in Figure 2. In the figure the envelopes define the DUs, and the vertical dotted lines demarcate the IUs, which are shaded grey. The white areas correspond to the pauses of the section.

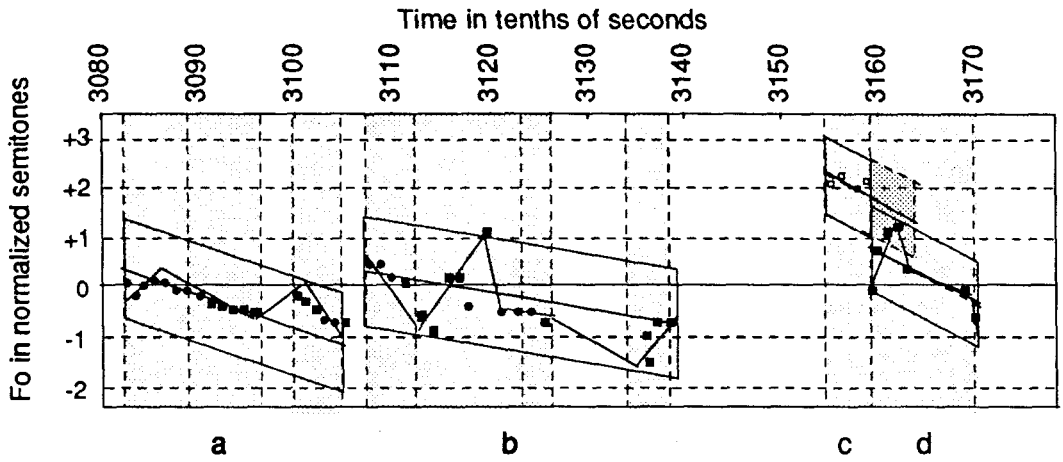


Figure 2. Acoustic data for a sequence of four DUs uttered by two speakers.

Coincidence of DU and IU boundaries

As both the totals reported above and the DUs and IUs of example (3) suggest, DUs were typically the larger unit (in time), and they frequently contained more than one IU. Thus, IU boundaries were much more numerous than DU boundaries. Even so, a definite relationship between the two was apparent. When an initial DU boundary was present (N=455), an initial IU boundary was also present 99% of the time. In other words, given a new DU, the likelihood of finding a new IU is almost certain. Example (3) above gives an indication of this usual pattern: Each DU boundary in the example corresponds to an IU boundary.

The converse relation, of course, does not hold. When an initial IU boundary was present (N=807), an initial DU boundary was copresent in just 450 cases, or about 56% of the time. That is, given a new IU, a new DU is likely to occur only about half the time. Again, (3) exemplifies the regular occurrence of DU-internal IU boundaries. The first DU in (3) contains two internal IU boundaries, while the second DU contains three.

The fact that the coincidence of DUs with IUs is high means that there are very few IU-internal DU boundaries. Indeed, our data exhibited only 3 instances. Such mismatches may simply be attributable to error in either analysis. On the other hand, we note that in all three cases, a single speaker was involved who was apparently having some difficulty in expressing herself at the point of the nonalignments. The auditory interpretation of such disfluency can be problematic, especially if it extends over a lengthy stretch of speech.

Relation of DUs to IUs

Since DUs were in nearly every case as long as or longer in time than IUs, the relationship between the lengths of the units can be conveniently expressed in terms of IUs per DU. For the corpus as a whole, the number of IUs per DU ranged from less than 1 (i.e. partial IU) to 8, with a mean of 1.8. Instances of minimal DUs would include the few cases of nonalignment just discussed. One-word responses and backchannel utterances are more typical cases of short DUs, albeit ones which correspond to IUs. DU (g) of example (1) is one such instance. Contrast with this the long DU illustrated in (3). The second DU of the example contains four IUs (but note that three of them are fairly short). DUs of this length were quite rare. In our data the vast majority of DUs (96%) contained 1-3 IUs, with a median of 2.

Factors contributing to the coincidence of DUs and IUs

Earlier we noted that factors other than reset play some role in the perception of prosodic unit boundaries. We identified aspects of utterance timing and speaker turns — in particular,

pause, new turn, and turn length — as likely having an influence on the DU-IU relationship. In this section we consider the occurrence of these factors in conjunction with DUs, IUs, and with jointly aligned (i.e. coinciding) prosodic unit boundaries.

Pause. A total of 305 pauses 0.3 second or greater were measured from the spectrograms and waveforms of the speech signal. Of these, 293 or 96% preceded either acoustic or auditory unit boundaries or both ('boundary pauses'). Boundary pauses occurred before jointly aligned DU and IU boundaries in 225 cases; before nonaligned boundaries, pauses occurred in 2 cases with respect to DUs and in 66 with respect to IUs. Thus, the presence of a pause would imply the beginning of at least one new prosodic unit boundary. However, prosodic units regularly occurred both with and without boundary pauses. Table 1 summarizes the location of all pauses in relation to prosodic unit boundaries.

Table 1. Occurrence of pauses in relation to prosodic unit boundaries. N is the number of units. Boundary pauses are pauses between units. The occurrence of internal pauses is independent of boundary pauses.

	N	with preceding boundary pause		without boundary pause		with internal pause	
DUs	455	227	*50%*	228	*50%*	76	*17%*
IUs	807	291	*36%*	516	*64%*	12	*1%*
Total	1262	518	*41%*	744	*59%*	88	*7%*

It is noteworthy that just 518 (41%) of the 1262 prosodic unit boundaries were preceded by pauses. In other words, pauses played no role in boundary identification for a substantial portion of prosodic units, with half of the DUs and nearly two-thirds of the IUs showing no boundary pause. Furthermore, one sixth of the DUs had unit-internal pauses. (Most of these, though, occurred at IU boundaries, as IUs generally lacked internal pauses.) In sum, whenever a pause occurs, a DU boundary is possible, and an IU boundary is likely. But it is more likely that a boundary will occur without an associated pause.

New turn. The tabulated data included a total of 214 speaker turns for which there were F_0 readings and 117 backchannel utterances. Only the prosodic units that were part of actual turns, however, could be included in the analysis. Unfortunately, F_0 readings were not available for nearly all (97%) of the backchannels, thus they had to be omitted from the final counts. As the representativeness of the remaining 4 instances for which there were acoustic data was in question, they were likewise omitted.

By definition, all turns began new IUs and DUs. (While there were 4 turns in which, in the acoustic analysis, the initial IU(s) could have been included in a DU begun by the previous speaker, we took these apparent interspeaker units to be an effect of the normalization process. In these few instances, speaker change alone defined the DU boundary.) Yet the majority of prosodic boundaries (66%) did not occur at turn junctures. Viewed differently, at most 47% of the DU and 27% of the IU boundaries could possibly be attributed to their turn-initial status. Thus, while a new prosodic unit boundary accompanies each instance of speaker change, it is more likely that such a boundary will occur during a turn, rather than at its beginning.

Turn length. The data included a broad range of turn lengths, as measured in either DUs or

IUs. Turn length measured in IUs varied from 1 to 55 IUs; measured in DUs, the length varied from 1 to 25 DUs. A good many turns were of minimal length: 35% (69) were only one IU long, while 67% (132) were one DU in length. Since the number of prosodic units was not equally distributed across turns of all sizes, the DU-IU correspondence was scrutinized according to turn length. Given that a DU boundary implied an IU boundary (and thus boundary alignment at just those points), it sufficed to compute the average number of IUs per DU for short turns versus long turns. We calculated (in IUs) the mean length of DUs for turns 1 DU in length versus all other turns. These short turns averaged 1.5 IUs/DU (±0.8); the longer turns averaged 1.9 IUs/DU (±1.0). The difference is statistically significant (in a one-tailed t-test, $t = 4.12$, $p < .001$). Thus, frequent speaker change does somewhat inflate the total number of joint prosodic unit boundaries present in the data, and it does marginally increase the overall rate of boundary coincidence (to 1.8 IUs/DU). However, the consequences of this bias for our data are limited. A prominent DU-IU correspondence still holds in extended stretches of talk by one speaker. As noted above, the majority of prosodic unit boundaries occurred turn internally, despite the high percentage of one-DU turns. Moreover, the slightly lower IU/DU ratio for one-DU turns is balanced by the relatively constant ratio for longer turns. That is, for turns longer than one DU, the mean IU/DU ratio does not continue to increase as turn length increases, but seems to reach a ceiling and varies within a very limited range.

The relative effect of pause, turn, and reset. While it is evident that pause and turn both exert a measure of influence in the occurrence of prosodic unit boundaries, we have not yet examined the combined influence of these factors and their effect on boundary coincidence in relation to reset (or, for IUs, other prosodic features). Consider the distribution of these factors, presented in Table 2.

Table 2. Cooccurrence of factors determining prosodic unit boundaries. Here, 'pause' is boundary pause; 'reset' is F_0 reset; 'other prosodic' refers to prosodic factors other than pause which define IUs. Factors are tallied separately for DUs and IUs. 'Aligned' refers to coinciding initial DU and IU boundaries.

	DU aligned		DU only	DU total		IU aligned	IU only	IU total
Turn, pause & reset	97	23%	-	97	Turn, pause & other prosodic	(97)	-	97
Turn & reset	117	24%	-	117	Turn & other prosodic	(117)	-	117
Pause & reset	128	31%	2	130	Pause & other prosodic	(128)	66	194
Reset only	108	22%	3	111	Other prosodic only	(108)	291	399
Total	450	100%	5	455	Total	(450)	357	807

Table 2 lists the occurrence of prosodic units with respect to the presence or absence of boundary pause and speaker turn. DU and IU tallies are given separately. Coinciding prosodic unit boundaries are tallied in the 'aligned' columns. (The figures are, of course, identical, but are listed under both units to indicate that alignment involves both the acoustic and the auditory dimensions.) Unaligned boundaries are tallied in the 'DU only' and 'IU only' columns. Note that any combination of features which includes new turn will result in boundary alignment.

Again, a bit under half of the cases of boundary alignment (23% + 24%) are associated with new turns, and a bit more than half (23% + 31%) are associated with boundary pauses. But a substantial minority (22%) of jointly aligned boundaries have no connection with either pause or turn. In these instances, we can attribute the alignment of DU and IU boundaries to the feature of

reset, as this is the only other prosodic factor influencing the formation of DUs. In other words, although a majority of units exhibit boundary cues which make it impossible to ascertain the contribution of reset alone in the perception of the unit boundaries, a substantial minority of prosodic units occur without the pause or turn cues.

To summarize, while the occurrence of pause or speaker turn can be seen to favor the coincidence of DU and IU boundaries, there were far fewer pauses or turns than aligning boundaries. Specifically, although a majority (78%) of boundary pauses did fall at joint prosodic unit boundaries, unit-internal pauses were not uncommon. And although a change in speaker implied a joint prosodic unit boundary throughout the data, boundary alignment also occurred regularly in stretches of speech by one speaker.

Discussion

Our hypothesis has been proven in that a correlation between DUs and IUs exists. The boundaries of the acoustically and auditorily based units overlap to a large extent. Of course, this finding — that the instrumental and perceptual measures of intonation do not conflict — was not unexpected. Assuming any relationship between the physical signal and its perception, we would expect a better than chance alignment of initial DU and IU boundaries, as reset (and pause) are features both prosodic units share. But our results show that the DU-IU correspondence is much more than casual. Virtually all new DUs began new IUs as well. Clearly, the acoustic facts of DUs are directly related to their perceptual equivalents in IUs. In light of this, we can point to a specific acoustic measure which correlates with IU boundaries, that is, we may infer that F_0 reset is an acoustic correlate of pitch reset.

The DU-IU correspondence and the auditory analysis of intonation

The fact that we can identify an acoustic correlate of IUs leads us to conclude that there is a definite phonetic basis to perceived intonation units of the type outlined in Du Bois et al. (1991). While this conclusion in and of itself is perhaps unremarkable — especially for those researchers who customarily work with instrumental data in experimental situations — it raises issues which are in some sense still controversial. Specifically, our results directly challenge commonly held views on the unreliability of perceptual judgments of intonation by trained linguists.

Ever since the publication of Lieberman's (1965) famous experiment, many linguists have been skeptical about the use of auditory data, insisting that if they are to have any value at all, they must be verified either by experiment or machine. Brown, Currie, and Kenworthy (1980, p. 48) reflect this view in justifying the necessity of acoustic as well as auditory data: "The dangers of relying on auditory analysis alone are well known". Such overgeneralizations are, however, due in part to a misrepresentation of Lieberman's results. His conclusions apply in fact to one particular system of intonational transcription — the Trager-Smith representation of pitch levels, stress, and juncture — not to all forms of auditory analysis. Our systematic comparison of acoustic and auditory analyses makes this abundantly clear. Auditory judgments of prosodic features can be consistently made, given an appropriate analytical framework.

Skepticism towards auditory analysis has also resulted from certain expectations regarding the nature of prosodic units. Those who have sought acoustic correlates of standard auditory units without success understandably have little reason to be confident about perceptual judgments. Again, let us take Brown, Currie, and Kenworthy (1980) as a well known example. In their intonational studies, they were compelled to reject the use of tone groups, having encountered difficulty segmenting the data "in a principled way" (p. 46). In other words, they found that no purported feature of the notion 'tone group' could unambiguously delineate the speech stream. Instead of tone group, they employed a pause-defined unit, finding pause an easily measured boundary marker. Although preferring a single, relatively transparent parameter in defining a unit of intonation may be methodologically satisfactory, it does not guarantee meaningful results.

Pauses in connected speech, though measurable with comparative ease, are multifunctional; speakers pause for interactional, cognitive, and rhetorical, as well as for intonational and grammatical reasons (cf. Deese, 1980; Goodwin, 1981). As we have seen, though frequently present in our data, a boundary pause need not accompany every IU. A unit based *only* on pause cannot be considered the primary domain for intonation (cf. Couper-Kuhlen, 1986, p. 75).

Moreover, in preferring such a parameter, the assumption is made that units of intonation can be categorically defined on the basis of one prosodic feature. What the DU-IU correspondence outlined here indicates, however, is that auditorily based units of intonation should not be expected to be definable on one criterion alone. Not all IUs exhibit reset. (In our corpus, 44% of the IUs do not have the initial pitch shift required to trigger declination reset.) Thus, singling out one feature — such as reset — will suffice to delimit only a subset of units, albeit a large one.

Doubts about the value of instrumental analysis of prosody are also expressed in the literature. Skeptics here point to the inability of the machine to distinguish between linguistically relevant and irrelevant aspects of the acoustic signal. As Crystal (1969: p. 13) states, "Instrumental analyses produce pictures of speech which are too sensitive to detail to provide any clear pattern". Beyond the issue of sensitivity, questions are sometimes raised regarding the interpretation of the acoustic record. Of course, before instrumental data is meaningful, certain interpretative steps must be made, introducing the very subjectivity the instrumental method seeks to avoid. Nevertheless, the results of our comparison of DUs and IUs demonstrate that the judicious use of acoustic data readily produces relevant patterns of organization. While discerning a 'coherent intonation contour' from the raw acoustic facts may indeed be problematic, a combination of prosodic and nonprosodic phenomena — F_0 declination change as observed in normalized data, pauses, and speaker turns — can serve as a reliable boundary-identifying metric.

The scopes of DUs and IUs

Although the correspondence between DUs and IUs was found to be high, the two units were often not coextensive. A range in the number of IUs per DU was observed, and on the average IUs occurred about as twice as often as DUs. Obviously, DUs and IUs differ in scope. This, of course, has to do directly with the way DUs and IUs are defined and delimited. A perfect alignment between DUs and IUs cannot be expected, because the auditory features corresponding to the acoustic parameters defining DUs constitute only two of a larger set of features used in the identification of IUs. Besides pitch reset (acoustically, the measure of change of the F_0 declination line) and perception of a measurable pause, numerous other prosodic features were said to cue the presence of an IU boundary, including pitch prominence, prosodic lengthening, and tempo acceleration — features which have little, if any, direct effect on the overall declination. Yet there is no requirement that an IU invariably exhibit any particular feature or set of features. As a unit defined in terms of a prototype, the more features that coalesce at any point, the stronger ('more prototypical') the boundary will be, but an IU boundary may also be perceived when only one or two features occur. Accordingly, F_0 reset need not be present. For instance, the pattern illustrated in (4) commonly occurs at the beginning of utterances.

(4) P: We=ll, / ... uh=m, /
...(1.1) You- you sort of cóme dówn through this=, / ... you know, / this= cástle. \
[V58-63]

In this example, the combination of prosodic lengthening and nonfinal (here level) pitch on the first word of the utterance, together with the following pause, serves to distinguish *well* as a separate IU without the occurrence of a pitch reset following the pause.

Prosodic domains and declination

Taking account of declination has been an important aspect of modeling F_0 contours, whether the phenomenon is incorporated directly into the model (e.g. Pierrehumbert, 1980) or is explained

by the interaction of other components (e.g. Liberman and Pierrehumbert, 1984). In dealing with acoustic phenomena such as declination, it is common for explicit models of intonation to assume a prosodic domain the size of an 'intonational phrase' (IP). Our comparison of acoustic and auditory units in natural discourse, however, casts doubt on this assumption. (While our analysis has been in terms of IUs instead of IPs, they are clearly of the same order of magnitude: IPs have been compared, for instance, to tone units (Pierrehumbert, 1980, p. 64). For the purposes of this discussion we will treat them as equivalent.)

The fact that declination is not necessarily reset at each IU boundary, but that it commonly extends over a sequence of two or three (and occasionally more) IUs, poses some problems for standard phonological accounts. First of all, it is not clear that the declination data in our corpus could be properly described. Had we a priori restricted the domain of declination to an IU-sized unit, declination would normally have been observable, but we would never have been able to observe its full extent. Moreover, F_0 reset would not be apparent in each of these smaller units. Studies such as Lieberman, Katz, Jongman, Zimmerman, and Miller (1985) illustrate the point being made. Their sample of spontaneous speech — nineteen "short, simple, declarative sentences" (p. 650) extracted from recordings of one conversation and two lectures — had extreme syntactic limitations imposed on it. Consequently, it is not surprising that they found declination in only two thirds of the extracted sentences. In our corpus, only 56% of IUs showed a reset in pitch large enough to accompany a new declination line.

Finally, we wonder whether explanations of the mechanics of declination which make reference to the internal operations of IU-sized (or syntactic) units can appropriately deal with a phenomenon that we have shown often transcends the domain of the unit.

If not the IP, what, then, is the domain over which declination operates? In fact, this crucial question is seldom asked — or at least not answered. (Sorensen and Cooper (1980, p. 421) ask the question, but decline to address it directly. Indeed, they work with the same sentence or sentence-pair examples of most other researchers (see Ladd (1986) for an exception). Moreover, it is not quite true that "the kinds of inferences drawn from experiments with read speech exhibiting [declination] would also apply to spontaneous speech" p. 408.) Much as in syntax with the 'sentence', the adequacy of the IP as a unit is rarely examined critically. In experimental situations where this prosodic domain is given in advance, an illusion of perfect alignment may occur, so there is little impetus for such examination. (Compare, however, Bruce (1982) and Thorsen (1985; 1986), who observed separate declination effects for clauses and clause sequences in readings of short passages of Swedish and Danish. Thorsen thus argues for a layered or hierarchical system of domains, with smaller prosodic units subordinated to larger scale ones.) But in connected discourse, a somewhat larger domain for declination is apparently called for. What is this domain? In the following paragraphs we examine three possibilities for reconciling the differences in scope of our units: A perceptual domain larger than the IU, possibly corresponding to a DU; an acoustic domain smaller than the DU, possibly corresponding to an IU; and a domain of IUs and DUs connected by other common factors not measured here.

Auditory units beyond the IU. Since the results of our comparison of DUs and IUs show that the two units are not simply equivalent units in different analytical dimensions (acoustic vs. auditory), the question arises as to whether we can more closely relate the acoustically derived DUs to any other auditory units. That is, are there perceptual units which are larger than IUs? The relevant literature does mention several possible candidates. Those defined in terms of perceptual units include: 'pitch sequence' (Brazil, 1978), 'minor paratone' (Yule, 1980), 'major phrase' (Ladd, 1986), and '[intonational] sentence' (Chafe, 1987). However, upon inspection a basic problem with each of these candidates is encountered: Each is too narrowly defined to account for the comparatively wide variation exhibited in DUs.

For instance, a pitch sequence is defined as "any stretch of language which ends with low ter-

mination and has no other occurrences of low termination within it" (Brazil, 1978, p. 18). In terms of the notation system used here, the boundary between pitch sequences would be most closely identified by the presence of a final fall (more specifically, the last IU of the sequence will exhibit this pattern). In our texts, IUs exhibiting final fall (transcribed '.\') do strongly tend to occur at the end of DUs (114/131 or 87%) — however, not all do. Some final falls are DU internal. More importantly, though, only a quarter of all DUs end in a final fall (114/455). The actual intonational structure of DUs is not limited to some theoretically correct pattern. Example (5), an extended turn by the main speaker, illustrates this point.

(5) K: Nó=.\ .. bút I thínk -- .. théy were --
...(.9) Wé thínk they were uh=, / ... cáse- -- ... gétting the upstáirs, / when we came báck,\
... And that we'd scáred them óff,\ but there's no good ... évidence on that.\
... I mean there ís good évidence,\
.. The sécond time they bróke in, / and got nothing,\
... Dóug came back,\
... An=d uh ...(.9) scáred them ó=ff.\ [R52-65]

While the DUs of lines 3 and 7 end in final falls, the remainder do not; furthermore, line 1 contains a DU-internal final fall.

A similar lack of correspondence in the definition of the other units and the prosodic structure of DUs is observed. Minor paratones and intonational sentences always end in a fall at or near the bottom of the speaker's pitch range. As just stated above, this is not true for DUs. In the case of the major phrase, it is said to be "set off by audible prosodic breaks" (Ladd, 1986, p. 316) of the type that correspond more to the properties of IUs: pauses, syllable-final lengthening, and boundary tones. These features all occur DU internally, as well as at DU boundaries. Thus, it would appear that these larger-scoped auditory units are not well matched to DUs.

Finally, DUs could possibly be related to the 'π-frames' of Gibbon (1984), which serve as the domain for global prosodic features such as declination. While the scopes of DUs and π-frames appear to be similar in many respects, features other than declination as measured here are associated with the latter. Furthermore, it is unclear how linear DUs would map to these process-oriented, recursive structures. However, the relationship of IUs to DUs — unnoticed in data from reading sentences aloud — would seem to provide indirect support for Gibbon's prosodic frames. Gibbon notes that the distinction between π-frames and γ-frames (corresponding roughly to IUs) is typically not drawn, "since π-frames and γ-frames tend to be co-extensive in such data" (p. 184).

Acoustic units within the DU. Another approach to approximating the scope of acoustic and auditory units would be to locate smaller acoustic units within the DU. Some research on declination has indeed identified such internal acoustic structure. Sorensen and Cooper (1980), for example, found small ('partial') F_0 resets between syntactically connected clauses. Research on languages other than English has produced similar results. Collier (1985) found resets between clause boundaries in Dutch. Other evidence of acoustic subunits is provided by Thorsen (1985; 1986). Working with Danish, she found that the acoustic record for a sequence of sentences may exhibit a distinct declination line for each sentence in the sequence, in addition to the 'global' declination line.

Unfortunately, the syntactic orientation of most experimental work on declination makes comparison difficult, considering the limitations of this study. Results based on our DUs would seem to be compatible with those based on declination lines derived from readings of sentence sequences, in that the domain of (some aspects of) declination may be longer than a simple declarative clause. But in our data we do not notice the systematic partial reset reported. Whether or not such reset occurs between parts of an utterance may be largely a function of the type of data used. That is, differences in the data could dictate the nature of the declination observed, for the differ-

ences are indeed striking. In spontaneous speech, for instance, it is the exception to find simple clauses containing more than one full noun phrase argument (Du Bois, 1985). But most experimental data — including that in the above-mentioned references — consists of constructed sentences with two, three, or even four full noun phrases. It is well known that full noun phrase constituents usually bear intonational prominence. Hence experimenters have frequent F_0 peaks within short periods of speech with which to construct top line declination lines, as well as low points for bottom line constructions, something not usually available to analysts of natural data. The lack of smaller DUs in our data could well be a function of the fewer instances, relatively speaking, of intonational prominence.

A second approach to defining DUs of smaller scope would be to make a more detailed acoustical analysis of the data, rather than looking for declination subunits. Adding aspects of the F_0 curves such as location of peaks and contour shape, or correlates of the timing features used here to delineate IUs, would serve to segment the speech stream more finely. But the difficulty of defining such features solely on an acoustic basis, as noted for example by Hadding and Studdert-Kennedy (1964), would make this approach seemingly difficult to implement. Still, we expect that a closer study of some features, especially relative speech rate, could shed light on the location of many of the IU boundaries not coinciding with DU boundaries.

In sum, the possibility of smaller acoustic units defined by declination lines cannot be dismissed. It is conceivable that small pitch 'resets' do occur — and are perceived — in our data at most IU boundaries, but that the corresponding F_0 variations do not result in the resetting of the declination line. With the present acoustic data, however, evaluating such finer discriminations would be speculative at best.

It is also possible that there is no easily defined large auditory unit, or no small acoustic unit, and that the domains of both IUs and DUs are determined by other common features connecting the two kinds of units, the most obvious being the text, or syntax. The comparisons in this study were text-free as far as they could be made, but that is not to say that the text does not influence the perception of prosody. The interactional nature of the data also speaks for a closer consideration of the function individual IUs have in the conversations. Further investigation into the internal structure of DUs with respect both to the syntax of constituent IUs and to their interactional status seems warranted.

On the other hand, there may be no one appropriate phonological prosodic domain for declination. Brazil (1978, p. 18) has said informally what Lieberman (1967) proposed in a phonetic model: "Common-sense and rudimentary physiological considerations might lead one to expect a progressive fall in pitch over any considerable stretch of speech". If the regular decline of F_0 in DUs which we have observed for spontaneous speech can be attributed to the mechanics of speech, not only would its pervasiveness be accounted for, but also any apparent difficulty in fitting DUs to larger perceptual units would be more palatable. This view has been echoed by numerous researchers: "Where declination is systematic, it may be due not to the speaker's phonetic plan" — as most prosodic units presumably are — "but rather to physiological factors such as diminishing subglottal air pressure" (Levelt, 1989, p. 399). Other research seems to indicate that speakers have at least partial control over declination; Ladd (1984, p. 67) summarizes this point: "There is plenty of evidence that the *amount* of resetting after a phrase boundary is subject to linguistic constraints of one sort or another". Clearly, the position of declination in a phonological model is a complex problem that will be occupying our attention for some time to come.

Conclusions

In our analysis we determined the correspondence between acoustically based declination units (DUs) and auditorily defined intonation units (IUs). The degree to which the two kinds of units shared common initial boundaries served as our measure of correspondence. In comparing the ini-

tial boundaries of acoustic and auditory units, the issue in question was whether the two units were describing the same prosodic phenomenon, but in different dimensions. Our results indicated a two-part answer.

Because virtually all of the acoustic unit boundaries were also identified as auditory unit boundaries, we concluded that the specific acoustic features of DUs (F_0 reset and pause) had perceptible auditory correlates. Moreover, as the auditory analysts were consistently identifying these correlates when segmenting the corpus into IUs, the validity of using both acoustic and auditory methods in intonation analysis is bolstered.

Because the units differed in scope, with one to three auditory units occurring for every acoustic unit, we concluded that declination in normal American English conversation frequently extends beyond the scope of standard prosodic phrasing. The adequacy of the 'intonational phrase' as the primary domain for declination is thus called into question. In addition, due to this scope difference, an analysis in terms of DUs, while lacking the detail afforded by the IU analysis, adds a measure of cohesiveness not available from IUs. Neither unit alone (as defined here) is sufficient to completely describe our intonational data. Although we have focused in this paper on reset as a cue for identifying prosodic boundaries, it is hoped that the particulars of the DU-IU correspondence presented here can provide the first step in a proper acoustic description of IUs as well as an understanding of the internal structure of DUs.

Acknowledgements

We would like to thank Peter Ladefoged for use of the UCLA phonetics laboratory. Thanks also to Alan Cruttenden, Jack Du Bois, and Sandy Thompson for their comments and criticism on an earlier draft of this paper, and to members of the UCSB intonation seminar, 2 July 1990. In addition, we want to express our appreciation for the very helpful suggestions of Nina Grønnum and J. 't Hart.

Footnotes

[1] Conversion to normalized semitones was as follows. First the F_0 (Hz) values, which are on a linear scale, were converted to semitones (octave values), measured on a logarithmic scale. This was done by dividing the natural logarithm of the Hz value by the natural logarithm of the twelfth root of 2, or ln(2)/12, which is approximately 0.0578. (An octave interval, which doubles the pitch, results when ln(2) is added to the ln(Hz) value; since there are twelve semitones in an octave, a semitone interval is achieved by an increase of 1/12 ln(2). Assuming a zero point of 16.35 Hz for the semitone scale (a reasonable zero point for the threshold of pitch, cf. Graddol 1986: 228), the semitone value is actually:

$$\mathrm{Ln}\left(\frac{\mathrm{Hz}}{16.35}\right) / .0578$$

The normalized (Z) scores were then computed for each speaker from the mean and standard deviation of that speaker's semitone scores. The procedure was to subtract each value from the mean of the speaker and divide the result by the standard deviation of that speaker. That is, the score is expressed in terms of number of standard deviations from the mean for the speaker. This results in a mean of 0 for each speaker and a standard deviation of 1.

[2] Briefly, transcription conventions are as follows: Three dots '...' mark pause, with the length of pauses over 0.7 s given in parentheses; '..' marks a brief break in speech rhythm; '=' indicates prosodic lengthening; (H) indicates inhalation; @ is laughter. Vertically aligned pairs of brackets '[' or '[[' mark speaker overlap; '« »' indicates uncertain segmental material. Prominent syllables are

marked with an acute accent. Voice quality features are represented by an angle bracket notation: ‹P P› is soft, ‹F F› is loud speech; ‹@ @› is speech accompanied by laughter. For detailed information, consult Du Bois et al. (1991). A preliminary version is available as Du Bois, Cumming, and Schuetze-Coburn (1988).

References

Altenberg, B. (1987). *Prosodic Patterns in Spoken English.* Lund: Lund University Press.

Armstrong, L.E., and Ward, I.C. (1926). *Handbook of English Intonation.* Leipzig: G.B. Teubner.

Bolinger, D. L. (1964). Intonation as a Universal. *Proceedings of the IXth International Congress of Linguistics,* Cambridge 1962. The Hague: Mouton.

Bolinger, D. L. (ed.) (1972). *Intonation: Selected Readings.* Harmondsworth: Penguin.

Bolinger, D. L. (1986a). Intonation across languages. In J. Greenberg, C. Ferguson, and E. Moravcsik (eds.), *Universals of Human Language,* Vol. 2, *Phonology* (pp. 471-524). Stanford: Stanford University Press.

Bolinger, D. L. (1986b). *Intonation and its Parts.* Stanford: Stanford University Press.

Brazil, D. (1978). *Discourse Intonation II.* Birmingham: English Language Research.

Brown, G., Currie, K.L., and Kenworthy, J. (1980). *Questions of Intonation.* London: Croom Helm.

Bruce, G. (1982). Textual aspects of prosody in Swedish. *Phonetica,* **39**, 274-287.

Chafe , W. (1987). Cognitive constraints on information flow. In R.S. Tomlin (ed.), *Coherence and Grounding in Discourse* (pp. 21-51). Amsterdam: Benjamins.

Chafe, W. (1991). Prosodic and functional units of language. To appear in J.A. Edwards and M.D. Lampert (eds.),*Talking Data: Transcription and Coding for Discourse Research.* Hillside, NJ: Erlbaum.

Cohen, A., and 't Hart, J. (1965). Perceptual analysis of intonation patterns. *Proceedings of the Vth International Congress on Acoustics,* Liège 1964, paper A 16.

Cohen, A., and 't Hart, J. (1967). On the anatomy of intonation. *Lingua,* **19**, 177-192.

Collier, R. (1985). The setting and resetting of the baseline. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics,* **19**, 111-132.

Cooper, W., and Sorensen, J. M. (1981). *Fundamental Frequency in Sentence Production.* New York: Springer-Verlag.

Couper-Kuhlen, E. (1986). *An Introduction to English Prosody.* Tübingen: Niemeyer.

Cruttenden, A. (1986). *Intonation.* Cambridge: University Press.

Crystal, D. (1969). *Prosodic Systems and Intonation in English.* London: Cambridge University Press.

Crystal, D. (1975). *The English Tone of Voice.* London: Arnold.

Deese, J. (1980). Pauses, prosody, and the demands of production in language. In H.W. Dechert and M. Raupach (eds.), *Temporal Variables in Speech* (pp. 69-84). The Hague: Mouton.

Du Bois, J.W. (1985). Competing motivations. In J. Haiman (ed.), *Iconicity in Syntax* (pp. 343-365). Amsterdam: Benjamins.

Du Bois, J.W., Cumming, S., and Schuetze-Coburn, S. (1988). Discourse transcription. In S.A. Thompson (ed.), *Discourse and Grammar* (pp. 1-71), Santa Barbara Papers in Linguistics, Vol. 2.

Du Bois, J.W., Schuetze-Coburn, S., Paolino, D., and Cumming, S. (1991). Outline of discourse transcription. To appear in J.A. Edwards and M.D. Lampert (eds.), *Talking Data: Transcription and Coding for Discourse Research.* Hillside, NJ: Erlbaum.

Gibbon, D. (1984). Intonation as an adaptive process. In D. Gibbon and H. Richter (eds.), *Intonation, Accent and Rhythm: Studies in Discourse Phonology* (pp. 165-192). Berlin: de Gruyter.

Goodwin, C. (1981). *Conversational Organization: Interaction between speakers and hearers.* New York: Academic Press.

Graddol, D. (1986). Discourse specific pitch behavior. In C. Johns-Lewis (ed.), *Intonation in Discourse* (pp. 221-237). London: Croom Helm.

Hadding, K., and Studdert-Kennedy, M. (1964). An experimental study of some intonation contours. *Phonetica* , **11**, 175-185. Reprinted in D. Bolinger (ed.) (1972). *Intonation: Selected Readings* (pp. 348-358). Harmondsworth: Penguin Books.

Halliday, M.A.K. (1967). *Intonation and Grammar in British English.* The Hague: Mouton.

't Hart, J., and Collier, R. (1975). Integrating different levels of intonation. *Journal of Phonetics*, **3**, 235-255.

't Hart, J., Collier, R., and Cohen, A. (1990). *A Perceptual Study of Intonation.* Cambridge: Cambridge University Press.

Jones, D. (1909). *Intonation Curves.* Leipzig: G.B. Teuber.

Kingdon, R. (1958). *The Groundwork of English Intonation.* London: Longmans, Greene and Company.

Krause, M. (1984). Recent developments in speech signal pitch extraction. In D. Gibbon and H. Richter (eds.), *Intonation, Accent and Rhythm: Studies in Discourse Phonology* (pp. 243-252). Berlin: de Gruyter.

Ladd, D.R. (1984). Declination: A review and some hypotheses. *Phonology Yearbook,* **1**, 53-74.

Ladd, D.R. (1986). Intonational phrasing: The case for recursive prosodic structure. *Phonology Yearbook,* **3**, 311-340.

Ladefoged, P. (1962). *Elements of Acoustic Phonetics*. Chicago: University of Chicago Press.

Levelt, W.J.M. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Liberman, M., and Sag, I. (1974). Prosodic form and discourse function. *Papers from the Tenth Regional Meeting, Chicago Linguistic Society,* 416-427.

Liberman, M., and Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff and R.T. Oehrle (eds.), *Language Sound Structure* (pp. 157-233). Cambridge, MA: MIT Press.

Lieberman, P. (1965). On the acoustic basis of perception of intonation by linguists. *Word,* **21**, 40-54.

Lieberman, P. (1967). *Intonation, Perception, and Language*. Cambridge, MA: MIT Press.

Lieberman, P., and Blumstein, S. (1988). *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge: Cambridge University Press.

Lieberman, P., Katz, W., Jongman, A., Zimmerman, R., and Miller, M. (1985). Measures of the sentence intonation of read and spontaneous speech in American English. *Journal of the Acoustical Society of America,* **77**, 649-657.

Maeda, S. (1974). A characterization of fundamental frequency contours of speech. MIT Research Lab in Electronics Quarterly Progress Report No. 114. (XVI Speech Communication).

Menn, L., and Boyce, S. (1982). Fundamental frequency and discourse structure. *Language and Speech,* **25**, 341-383.

O'Shaughnessy, D. (1976). Modelling fundamental frequency and its relation to syntax, semantics and phonetics. PhD Dissertation, MIT.

Oreström, B. (1983). *Turn-taking in English Conversation*. Lund: CWK Gleerup.

Palmer, H.E. (1922). *English Intonation, with Systematic Exercises*. Cambridge: Heffer.

Pierrehumbert, J. (1980). The phonology and phonetics of English intonation. PhD Dissertation, MIT.

Pijper, J.R. de (1983). *Modelling British English Intonation*. Dordrecht: Foris.

Pike, K.L. (1945). *The Intonation of American English*. Ann Arbor: University of Michigan.

Quirk, R., Svartvik, J., Duckworth, A.P., Rusiecki, J.P.L., and Colin, A.J.T. (1964). Studies in the correspondence of prosodic to grammatical features in English. In *Proceedings of the IXth International Congress of Linguistics*. The Hague: Mouton. Reprinted in R. Quirk (1968).

Essays in English Language: Medieval and Modern (pp. 120-135). London: Longman.

Sacks, H., Schegloff, E.A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking. *Language*, **50**, 696-735.

Schuetze-Coburn, S. (In progress). Information flow in discourse. PhD dissertation, UCLA.

Schuetze-Coburn, S., and Shapley, M. (Forthcoming). The coincidence of acoustically and auditorily based prosodic units with syntactic structure.

Selting, M. (1987). Descriptive categories for the auditive analysis of intonation in conversation. *Journal of Pragmatics*, **11**, 777-791.

Shapley, M. (1989). Fundamental frequency variation in conversational discourse. PhD dissertation, UCLA.

Sorensen, J.M., and Cooper, W.E. (1980). Syntactic coding of fundamental frequency in speech production. In R.A. Cole (ed.), *Perception and Production of Fluent Speech* (pp. 399-440). Hillsdale, NJ: Erlbaum.

Svartvik, J., and Quirk, R. (eds.) (1980). *A Corpus of English Conversation*. Lund: CWK Gleerup.

Thorsen, N. (1980). A study of the perception of sentence intonation—Evidence from Danish. *Journal of the Acoustical Society of America*, **67**, 1014-1030.

Thorsen, N.(1983). Standard Danish sentence intonation—Phonetic data and their representation. *Folia Linguistica,* **17**, 187-220.

Thorsen, N. (1985). Intonation and text in Standard Danish. *Journal of the Acoustical Society of America*, **77**, 1205-1216.

Thorsen, N. (1986). Sentence intonation in textual context—Supplementary data. *Journal of the Acoustical Society of America*, **80**, 1041-1047.

Willems, N. (1982). *English Intonation from a Dutch Point of View*. Dordrecht: Foris.

Yule, G. (1980). The speakers' topics and major paratones. *Lingua,* **52**, 33-47.

Modified version of a report commissioned by the Swedish Social Sciences Research Council. The opinions expressed are those of the author.

Phonetics and Phonology in Sweden

Peter Ladefoged

Phonetics is the study of speech sounds. Phonology is the study of the patterns of sounds that occur in languages. In some countries, such as the United States, departments of linguistics place major emphasis on the study of the formal constructs of phonology, and there are more phonologists than phoneticians. In Sweden this is clearly not the case. There are comparatively few people whose main interest is in formal phonology, and a considerable number who are concerned with the physical realization of speech sounds. In order to understand why this is so we must remember that speech involves much more than language, and part of phonetics is outside linguistics. As well as purely linguistic information, speech conveys sociolinguistic information about the social and regional background of the speaker, and additional information about the speaker's emotional state. It may also serve simply as an indicator of the speaker's identity. Phonetics is concerned with all of these aspects of speech sounds. In addition, phonetic studies are closely linked with work in speech pathology, psychology, and communications engineering.

It is possible to regard phonetics as a field in its own right; and, indeed, this has been the case at times in Sweden. Nowadays, however, phonetic research in Sweden is conducted mainly in linguistics departments or in institutes of technology. There are no Departments of Phonetics. Nevertheless in Sweden the wide ranging nature of phonetic studies is fully recognized, and phonology is regarded as the linguistic part of phonetics, rather than phonetics being regarded as the final stage of a linguistic description. In this chapter we will begin with a discussion of work in phonology, and then continue with an examination of various aspects of work in phonetics.

Phonology

There are many interesting phonological phenomena within Swedish itself. One of the best known is what is variously called supradentalization, or postalveolarization, or retroflexion. This involves the effect of a preceding /r/ on one of the consonants /t,d,n,s,l/ each of which is usually made on the teeth. When /r/ precedes there is a merger at the phonetic level into a single consonant made further back in the mouth, but with the same manner of articulation as the original dental consonant. The effect can occur both within words, as in the pronunciation of fort 'fort' with the final consonant being a merger of /r/ and /t/, and between words, as in för tunn 'too thin' with a similar merger, as well as in other circumstances, as has been shown by Eliasson (1986). (A similar effect occurs in many forms of American English, in which the /d/ at the end of the word hard is made further back due to the preceding /r/. This effect may spread to subsequent consonants in the same word; such as /n/ in harden, and both /n/ and /d/ in hardened and even to the initial consonants in a subsequent word, such as the /t/ in hardened to it.) This phonological process

is very prevalent in Swedish and has many ramifications, some of which have not yet been fully investigated.

Other basic phonological research on Swedish concerns such things as the number of vowels. Are there 18 vowels — 9 short and 9 long — or are the vowel length alternations completely predictable from the surrounding sounds so that Swedish has only 9 underlying vowels? As Eliasson (1985a) remarks in an interesting discussion of this problem: "... it is still possible, even in languages with a long research tradition, to discover extensive structural evidence whose significances for the outstanding issues in the description of the language has not yet been at all explored."

Swedish is one of the smaller languages of Europe, and Sweden has long realized the necessity of studying other languages. As a result there have been a number of studies of English and French phonology, some of which demonstrate Swedish contributions to new views of phonology. In the last decade linguists in many countries have been much concerned with non-segmental phonology and metrical structure. A typical problem concerning English has been the correct formulation of the rhythm rule whereby words such as 'thirteen,' which usually have the stress on the second syllable when pronounced in isolation, have the stress on the first syllable in phrases such as 'thirteen men.' Horne (1990) has provided some interesting data on this problem.

Swedish linguists have also considered phonological problems in a wide range of other languages, such as consonant deletion in Turkish (Eliasson 1985b), the structure of words in Maori (Eliasson 1989), Mongolian palatalization (Svantesson 1991), the origin of tones in Mon Khmer (Svantesson 1989), Welsh vowel length and syllable structure (Wood 1988) Lule-Sami voicing and duration patterns (Engstrand 1987a, 1987b), Old Icelandic i-umlaut (Braroe 1979) and Sanskrit aspiration(Ejerhed 1981). There have also been studies of other languages in which the emphasis has been more on the phonetic data than on points of phonological theory, including work on nasal mora in Japanese (Nagano-Madsen 1989), Wu (Chinese) consonants and tone (Svantesson MS), Greek intonation (Botinis 1991), Udehe creaky and breathy voice(Radchenko MS), Bulgarian vowel reduction (Pettersson and Wood 1987), and the interlanguage used by Korean learners of Swedish (Pyun 1987). Interesting work on sound change in perception and production has been reported by Janson (1983), who also has more recently been conducting investigations of the phonology of a wide range of languages (Janson 1991). But with the exception of work on pitch accent, which will be considered later, Swedish research in the more theoretical aspects of phonology has not been extensive. Unfortunately, Linnel, whose early work was very interesting (Linell 1979, 1982), no longer works in this area. It is good, however that at least some Swedish researchers are abreast of new developments in phonology and are making significant contributions in this area.

General Phonetics

A central concern of phonetics is the relation between the acoustics of sounds and the articulations that produced them. The first giant steps in this field came with Fant's Acoustic Theory of Speech Production (1960), and Fant and his colleagues at KTH have continued to lead the way from that time on. The recent work by Lin and Fant (1991) on vocal tract modeling continues this tradition, bringing a great deal of new knowledge to this field. Other KTH work on vocal cord vibrations and phonation types (Anantapadmanabha and Fant 1982) also has many implications for linguistic studies of languages that contrast more than just voiced and voiceless glottal states.

Until recently Sweden had another major contributor in general phonetic theory. Lindblom's ideas on the balance of forces shaping sound systems have received world wide attention. He has suggested (Lindblom, MacNeilage and Studdert-Kennedy, forthcoming) that languages are self-organizing systems in which the various patterns of sounds arise through different resolutions of the tensions between the needs of listeners for auditory distinctiveness, of speakers for articulatory ease, and of communities for social cohesion. Swedish phonetics has lost a lot now that ideas such as these are no longer available through Lindblom's teaching. The phonetics group that he headed is still producing noteworthy research. They have been much concerned with immigrants in Sweden (Cunningham-Andersson and Engstrand 1990). Engstrand's work on word accents (Engstrand 1989) and Traunmüller's extensive studies of vowels are also worthy of mention (Traunmüller 1981, 1988; Traunmüller and Lacerda 1987).

Researchers in other institutions in Sweden have also made notable contributions in general phonetics. Early insightful work by Öhman (1966, 1967) is still widely cited. It is a pity that he has not continued to contribute in this field. Wood has made extensive use of X-ray data to capture new generalizations about the articulations of vowels (e.g. Wood 1979, 1986, and 1991a). Much of this work is well known internationally, but some phoneticians express reservations as to whether his results involve oversimplification of the tongue shapes in vowels, and whether, in some papers, too much is inferred from a single speaker of each language. Recently he has been extending these studies beyond the steady state positions involved in vowels to considerations of the temporal coordinations of speech gestures (Wood 1991b), which is a useful contribution to current discussion on this topic.

Intonation and Prosody

It is difficult to know where to draw the line between phonology and phonetics; and this is particularly true when it comes to studies of intonation. A sentence may consist of a particular sequence of words, built up of certain speech sounds; but these words can be said with many different intonations, some of which convey linguistic information, and some of which do not. There obviously may be a linguistic difference between two sentences which differ in syntax. In the written language these differences may be marked by punctuation, but in the spoken language they are usually conveyed by the

intonation. An example of a pair of English sentences differing in syntactic structure is: "When danger threatens your children, call the police" compared with "When danger threatens, your children call the police." Comparable sentences in Swedish have been analyzed by Bruce, Granström and House (1991).

Other differences in intonation, however, may merely reflect differences in the attitude of the speaker to the person being addressed, or to the topic under discussion, or even to the world in general. A speaker may sound condescending to the listener, or sarcastic about the topic, or even angry at life. In between the two extremes of syntactic and attitudinal differences of intonation there are differences that may, or may not, be considered to be linguistic involving, for example, differences in emphasis on a particular word.

Intonation studies in Swedish are complicated by yet another point. In Scandinavian languages some words have lexically distinctive accents, so that, for example, anden 'duck' and anden 'spirit' are the same except for the pitch accent. The definitive work on this topic is that of Bruce (1977), which is internationally regarded as an insightful contribution to phonological studies of this kind. Perhaps partly because the pitch variations in Swedish words have long been recognized as an interesting facet of the Swedish sound system, (Meyer 1937) Sweden has long been a leader in prosodic studies of all kinds.

It is difficult to quantify pitch changes and make abstract formal descriptions of intonation because the range and the absolute values of the pitch changes in sentences are very dependent on the individual speakers. Strides in normalizing pitch records were made by Gårding (1983; see also Gårding 1991), who devised a system for placing a grid on a pitch curve and interpreting pitch changes in a standardized form. Variants of this technique have been used to describe tone and intonation in Hausa (Lindau 1986), Chinese (Gårding, Zhang and Svantesson 1983), and in French (Touati 1987, 1991). Current work at Lund on the exploitation of pitch in natural dialogue uses some of the most up to date methodology (Bruce 1991). Sophisticated computer systems are used not only for extracting physical variables such as the fundamental frequency (pitch) and duration, but also for testing the proposed phonological analyses. Thus observed prosodic patterns are being related to the linguistic structure of a spontaneous dialogue, with all its interruptions and turn-taking implementations. The Lund researchers are also comparing the prosodic patterns in spontaneous dialog with those in laboratory recordings of written material (Bruce, Granström and House 1991). At the moment, these two types of speech turn out to be not very different from the point of view of the prosodic structure. Further studies have shown how speakers of different dialects (British and American English) differ in their use of pitch to signal whether a new topic is being introduced (Horne 1991). To be valid, such work requires the analysis of the speech of a large number of speakers, which has not yet been done.

There is good cooperation in intonation studies between the researchers at Lund and those at KTH (e.g. Bruce, Granström and House 1990). But the

emphasis of much of the work on prosodic aspects of speech at KTH has been on durational properties rather than on pitch. The aim has been to describe the temporal organization and rhythm of Swedish, particularly with regard to prose reading in different styles (Fant and Kruckenberg 1989, Fant, Kruckenberg and Nord 1991a,b,c). This work has greater importance than its immediate practical relevance to studies of Swedish. It is an excellent example of how to get experimental evidence to bear on phonological questions by putting considerable thought into what are the correct things to measure. A large number of interesting statistics have been discovered. For example, when reading fast, content words such as nouns and adjectives shorten to about 75% of their duration in isolation; but function words such as pronouns, conjunctions and articles are considerably shorter, with articles being only about 21% of their length in isolation. It is also clear that there are greater changes in the lengths of the pauses rather than in the individual sounds.

These studies of the temporal organization and rhythm of Swedish have been extended to studies of poetry as well as prose (Nord, Kruckenberg and Fant 1991, Kruckenberg, Fant and Nord 1991); this work is highly regarded by phonologists. These researchers have also conducted cross-linguistic studies of Swedish, French and English. It is apparent that "the smaller contrast between stressed and unstressed syllable durations in French compared to English and Swedish is both a matter of a smaller contrast in syllable complexity and a lower degree of stress induced lengthening. In addition, the relative precision and low degree of vowel reduction in French reduces the stressed/unstressed contrast." (Fant, Kruckenberg and Nord, 1991). However, these results should be regarded with caution, as almost all of them are based on studies of a single speaker of each language. We hope that this rather severe defect will be corrected shortly.

Other work on pauses is in progress at Umeå. In this case most of the results are based on recordings of a text read by 10 speakers, each recording this text at a slow, normal and fast speed. This considerable body of data has led to a number of results (Strangert 1990a, b, 1991a, b,c). Pauses turn out to be marked in complex ways. "The acoustic correlates of pauses, in addition to silence, include prepausal lengthening, resetting of intensity and F_0, and voice quality irregularities. In general, the higher the rank of the boundary, the stronger and more varied were the acoustic correlates." (Strangert 1991a)

Speech synthesis

Some of the greatest challenges facing phonetics today are in the realms of speech synthesis and speech recognition. High quality synthetic speech is needed for many purposes. An obvious example is for use in reading machines for the blind. Of more commercial importance (regrettably) is the need for synthetic speech in answering machines that will supply over the telephone information from catalogues, timetables, telephone directories, stock exchange prices or any kind of data base that needs to be constantly updated. The task of producing good quality speech for any of these purposes requires the talents of both linguists and engineers

Nearly all the early stages of the process of going from text to speech involve linguistic research. First all the abbreviations and numbers have to be converted, so that the computer can pronounce "SEK285.4" as "two hundred and eighty five Swedish kronor and forty öre". Then the text has to be converted into a phonological transcription by using a set of rules, with comparatively few words being handled as exceptions. Again this is a task for which training in linguistics is helpful. Next all the phonological processes have to be taken into account, particularly those that adjust the dictionary forms of words to the forms required in particular contexts in connected speech. Stress has to be assigned correctly, bearing in mind differences such as in "He's now fifteen" compared with "Fifteen men." Most difficult of all is the assignment of intonation, which interacts with all levels of linguistic analysis, as well as involving knowledge of the speaker's attitude to the world in general, and to the topic under discussion. Still within the realm of phonetics is the task of taking the narrow phonetic transcription and translating it into a set of acoustic parameters (such as formant frequencies) for generating sounds. Swedish researchers have been in the forefront of many of these endeavors.

Research on speech synthesis in Sweden is centered at KTH, but there is notable collaboration with linguistics departments in other institutions, not only in Stockholm but also in Lund and Umeå. This collaboration makes this research an important part of linguistics in Sweden. It is also noteworthy that speech synthesis research in Sweden is multilingual in nature, and puts considerable emphasis on basic research. The text-to-speech system developed at KTH has always been designed to be multilingual, with the language-specific parts being formulated mostly in a notation close to the one commonly used in generative phonology (Carlson, Granström and Hunnicutt, 1990). The aim is to provide a developmental environment that can be easily used by linguists and others without the need to know about computer programming. This has led to the rapid development of multi-lingual speech synthesis systems. "Versions of the [KTH] text-to-speech program are now commercially available in British and American English, German, French, Italian, Spanish, Norwegian, Danish and Swedish" (Carlson, Granström and Hunnicutt, 1990). It should be noted, however, that intelligibility tests have shown that the American English version of this speech synthesis system compares rather poorly with the better monolingual American systems (Logan, Greene and Pisoni, 1989).

There are two kinds of basic research involved in the Swedish speech synthesis effort, one being concerned with general phonetic problems, and the other with more specific issues involving particular sounds and prosodies. The general phonetic problems include studies on articulatory synthesis (Lin 1990, Lin and Fant 1990), female voice sources (Karlsson 1991), and new methods of measuring segmental intelligibility (Carlson, Granström and Nord, 1990a, 1990b). This concern with basic research is evident even in the design of the KTH speech data base (Carlson, Granström and Nord, 1990c), which is intended for acoustic phonetic research rather than (as other similar data bases) primarily for evaluating speech recognition systems. Researchers at KTH have also been concerned with ways of explaining speaker characteristics

and speaking styles with respect to synthesizing different emotional and attitudinal dimensions. This work is not within the realm of linguistics; but it is certainly part of phonetics.

More linguistic offshoots of work on speech synthesis may be exemplified by research on Swedish "sonorants" /r,l,v,j/ by Carlson and Nord (1991), who discuss acoustic correlates of the allophones of these sounds and provide interesting reasons why they should be regarded as a phonological class, despite both /v/ and /j/ being weakly fricative. Going beyond phonology, morphological considerations are seen to play an important role in converting written text to a phonetic transcription suitable for subsequent speech synthesis.

Speech recognition

In many countries, most applied phonetic research is directed towards improving speech recognition systems. In Sweden, the emphasis is on speech synthesis rather than recognition, with the tacit understanding that speech researchers do not yet have sufficient knowledge to make large steps in speech recognition. This may be a wise decision in that there has not been any dramatic progress in speech recognition in the last few years. For a long time we have been able to use computers to distinguish single words, such as the digits zero through nine. More recently, several systems have been developed that can recognize limited sets of words in task-specific situations, in which the computer can structure the dialog. For example, in an airline reservations system, the computer can ask "Where do you want to go? Which day of the month do you wish to travel? At what time? On what airline? " For each of these questions there is only a limited set of possible answers. Computers can successfully recognize and process speech in these circumstances; but they cannot as yet interpret normal conversational speech as spoken by people with a wide range of accents and different personal characteristics—which any of us can do. Before we can build systems that can go any further in this task there is a great deal of linguistic research to be done.

Some research on speech recognition is being carried out in Sweden, but most of it is concerned with the less linguistic aspects of the task. There are, however, interesting phonetic implications to the hypothesis that speech recognition of variant pronunciations could be improved by taking into account the fact that differing realizations of an utterance often depend on variations in the synchrony between two or more articulatory gestures (Blomberg 1991). It is also of phonetic interest that speech recognition is improved if the system is allowed to adapt to the speaker's individual voice source spectrum (Blomberg 1989a, 1989b). At the moment, however it is hard to find any linguistic or phonetic implications in an artificial neural network that recognizes phonemes (Elenius and Takács, 1990), which is the basis for much research of this kind.

Other phonetic research

It is hard to know where to draw the line in the discussion of linguistic aspects of phonetic research in this report. There is a considerable amount of work in speech pathology (often called phoniatrics in Sweden), much of it by people in linguistics departments. Some of this is clearly within the realm of

linguistics. For example, Magnusson and Nauclér (1990) have shown that language dis-ordered pre-school children may not have the linguistic and meta-linguistic prerequisites that are needed for learning to read and write adequately. Their work has also thrown interesting light on the argument over whether knowledge of phonemes can only be developed as an effect of learning to read and write in an alphabetic system. Magnusson and Nauclér (1988) have shown it is possible to become aware of phonemes without knowing the alphabet.

Other work in speech pathology is less evidently linguistic, although it is valuable in its own way. In Stockholm Almé, Öberg and Engstrand (1989) have discussed the speech of a glossectomized speaker, and Nord and Britta (1989) and Nord, Hammarberg and Lundström (1991) have reported on speech without a larynx. In Lund there is research that has interesting implications for the hearing impaired (House 1990a, 1990b). There is also a great deal of cooperation with the Department of Phoniatrics.

Conclusion

It is clear that Sweden is a world leader in phonetics, and has some well known work in phonology. Students from abroad will no doubt continue to come to study phonetics at the laboratories in Lund and Stockholm universities, as well as at KTH, all of which have high international reputations. Phonetics at Umeå is also strong, and there are facilities with good potential at Gothenburg and Uppsala. Even with the retirement of Fant from the professorship (though not from active research; see Fant 1989, 1990, 1991), the loss of Lindblom to the United States, and the move of Öhman to other interests, there remain notable senior scholars such as Granström and Bruce. There are also many middle level and younger scholars who are establishing international reputations.

References

Almé, A., E. Öberg and O. Enstrand. (1989). "An acoustic-perceptual study of Swedish vowels produced by a subtotally glossectomized speaker." *Phonetic Experimental Research, Institute of Linguistics, University of Stockholm* 10: 157-184.

Anantapadmanabha, T. V. and G. Fant. (1982). "Calculation of true glottal flow and its components." *Speech Communication* 1: 167-184.

Blomberg, M. (1989a). Adaptation to a speaker's voice in a speech recognition system based on synthetic phoneme references. Eurospeech 89.

Blomberg, M. (1989b). "Synthetic phoneme prototypes in a connected-word speech recognition system." *Proc. ICASSP 90* : 687-690.

Blomberg, M. (1991). Modelling articulatory inter-timing variation in a speech recognition system. XII International Congress of Phonetic Sciences. 4: 466-469.

Botinis, A. (1991). Intonation patterns in Greek discourse. XII International Congress of Phonetic Sciences. 4: 286-289.

Braroe, E. (1979). "Exceptions to Old Icelandic I-umlaut." *Studia Linguistica* 33(1): 43-56.

Bruce, G. (1977). "Swedish word accents in sentence perspective." *Travaux de l'institut de Linguistique de Lund* XII:
Bruce, G. (1991). The exploitation of pitch in dialogue. XII International Congress of Phonetic Sciences. 1: 271-274.
Bruce, G., B. Granström and D. House. (1990). Prosodic phrasing in Swedish synthesis. ESCA Tutorial Day and Workshop on Speech Synthesis.
Bruce, G., B. Granström and D. House. (1991). Strategies for prosodic phrasing in Swedish. XII International Congress of Phonetic Sciences. 4: 182-185.
Carlson, R., B. Granström and S. Hunnicutt. (1990). IMultilingual text-to-speech development and applications. *Advances in speech, hearing and language processing.* A.W. Ainsworth, ed. London, JAI Press. 269-296
Carlson, R., B. Granström and L. Nord. (1990a). "Evaluation and development of the KTH text-to-speech system on the segmental level." *Speech Communication* 9: 271-277.
Carlson, R., B. Granström and L. Nord. (1990b). "The KTH speech database." *Speech Communication* 9: 375-380.
Carlson, R., B. Granström and L. Nord. (1990c). Segmental intelligibility of synthetic and natural speech in real and nonsense words. 1990 International Conference on Spoken Language processing.
Carlson, R. and L. Nord. (1991). "Positional variants of some Swedish sonorants in an analysis-synthesis scheme." *Journal of Phonetics* 19: 49-60.
Cunningham-Andersson, U. and O. Engstrand. (1990). "Perceived strength and identity of foreign accent in Swedish." *Phonetica* 46: 138-154.
Ejerhed, E. (1981). "The analysis of aspiration in Sanskrit phonology." *Nordic Journal of Linguistics* 4: 139-159.
Elenius, K. and G. Takács. (1990). "Acoustic-phonetic recognition of continuous speech by artificial neural networks." *STL-QPSR* (2-3): 1-44.
Eliasson, S. (1985a). "Stress alternations and vowel length: New evidence for an underlying nine-vowel system in Swedish." *Nordic Journal of Linguistics* 8: 101-129.
Eliasson, S. (1985b). "Turkish k-deletion: simplicity vs. retrieval." *Folia Linguistica* 19(3-4): 291-309.
Eliasson, S. (1986). Sandhi in peninsular Scandinavian. *Sandhi Phenomena in the languages of Europe.* H. Anderson, Ed. Berlin, Mouton de Gruyter.
Eliasson, S. (1989). "English-Maori language contact: code-switching and the free-morpheme constraint." *Reports from Uppsala University, Department of Linguistics* 18: 1-28.
Engstrand, O. (1987a). "Durational patterns of Lule Sami Phonology." *Phonetica* 44: 117-128.
Engstrand, O. (1987b). "Preaspiration and the voicing contrast in Lule Sami." *Phonetica* 44: 103-116.
Engstrand, O. (1989). "Phonetic features of the acute and grave word accents: data from spontaneous speech." *Phonetic Experimental Research, Institute of Linguistics, University of Stockholm* 10: 13-37.
Fant, G. (1960). *Acoustic theory of speech production.* The Hague, Mouton.
Fant, G. (1989). "Quantal theory and features." *Journal of Phonetics* 17: 79-86.
Fant, G. (1990). "Speech research in perspective." *Speech Communication* 9: 171-176.

Fant, G. (1991). Units of temporal organization, stress groups versus syllables and words. XII International Congress of Phonetic Sciences. 1: 247-250.
Fant, G. and A. Kruckenberg. (1989). "Preliminaries to the study of Swedish prose reading and reading style." *STL-QPSR* 2:
Fant, G., A. Kruckenberg and L. Nord. (1991a). Language specific patterns of prosodic and segmental structures in Swedish, French and English. XIIth International Congress of Phonetic Sciences. 4: 118-121.
Fant, G., A. Kruckenberg and L. Nord. (1991b). Tempo and Stress. Fonetik '91.
Fant, G., A. Kruckenberg and L. Nord. (1991c). Temporal orangization and rhythm in Swedish. XII International Congress of Phonetic Sciences. 1: 251-256.
Fant, G., A. Kruckenberg and L. Nord. (forthcoming). "Durational correlates of stress in Swedish, French and English." *Journal of Phonetics* .
Gårding, E. (1983). A generative model of intonation. *Prosody: models and measurements.* Berlin, Springer.
Gårding, E. (1991). Intonation parameters in production and perception. XII International Congress of Phonetic Sciences. 1: 301-304.
Gårding, E., J. Zhang and J.-O. Svantesson. (1983). "A generative model for tone and intonation in Standard Chinese." *Working papers, Lund University, Department of Linguistics* 25: 53-65.
Granström, B. and L. Nord. (1991). Ways of exploring speaker characteristics and speaking styles. XII International Congress of Phonetic Sciences. 4: 278-281.
Horne, M. (1990). "Empirical evidence for a deletion formulation of the rhythm rule in English." *Linguistics* 28: 959-981.
Horne, M. (1991). Phonetic correlates of the 'new/given' parameter. XII International Congress of Phonetic Sciences. 5: 230-233.
House, D. (1990a). "On the perception of mood in speech: Implications for the hearing impaired." *Working papers, Lund University, Department of Linguistics* 36: 99-108.
House, D. (1990b). *Tonal perception in speech.* Travaux de l'institut de linguistique de Lund. Lund, Lund University Press.
House, D. (1991). A model of optimal tonal feature perception. XII International Congress of Phonetic Sciences. 2: 102-105.
Janson, T. (1986). Sound change in perception. *Experimental Phonology.* Orlando, Academic Press, Inc.
Janson, T. (1991). Comments on Maddieson: investigating linguistic universals. XII International Congress of Phonetic Sciences. 1: 355-358.
Kruckenberg, A., G. Fant and L. Nord. (1991). Rhythmical structures in poetry reading. XII International Congress of Phonetic Sciences. 4: 242-245.
Lin, Q. (1990). *Speech production theory and articulatory speech synthesis.* Stockholm, Dept. of Speech Communication and Music Acoustics, KTH.
Lin, Q. and G. Fant. (1990). "A new algorithm for speech synthesis based on vocal tract modeling." *STL-QPSR* (2-3): 45-52.
Lindau, M. (1986). "Testing a model of intonation in a tone language." *Journal of the Acoustical Society of America* 80: 757-764.
Lindblom, B., P. MacNeilage and M. Studdert-Kennedy. (forthcoming). *The Biological Basis of Spoken Language.* San Francisco, Academic Press.

Linell, P. (1979). *Psychological reality in phonology.* Cambridge, Cambridge University Press.
Linell, P. (1982). "The concept of phonological form and the activities of speech production and speech perception." *Journal of Phonetics* 10: 37-72.
Logan, J. S., B. G. Greene and D. b. Pisoni. (1989). "Segmental intelligibility of synthetic speech produced by rule." *Journal of the Acoustical Society of America* 86(2): 566-580.
Magnusson, E. and N. Kerstin. (1988). "How to become aware of phonemes without knowing the alphabet." *Working papers, Lund University, Department of Linguistics* 33: 163-171.
Magnusson, E. and N. Kerstin. (1990). "Reading and spelling in language-disordered children – linguistic and metalinguistic prerequisites: report on a longitudinal study." *Clinical Linguistics and Phonetics* 4(1): 49-61.
Meyer, E. A. (1937). *Die Intonation im Schwedischen.* Stockholm, University of Stockholm.
Nagano-Madsen, Y. (1989). "Mora and temporal-tonal interaction in Japanese." *Working Papers, Department of Linguistics and Phonetics, Lund University* 35: 121-131.
Nord, L. and H. Britta. (1989). "Analysis of Laryngectomee speech - a progress report." *Eurospeech* 2: 493-496.
Nord, L., B. Hammarberg and E. Lundström. (1991). Phonetic aspects of speech produced without a larynx. XII International Congress of Phonetic Sciences. 4: 322-325.
Nord, L., A. Kruckenberg and G. Fant. (1990). "Some timing studies of prose, poetry and music." *Speech Communication* 9: 477-483.
Öhman, S. (1966). "Coarticulation in VCV utterances: spectrographic measurements." *Journal of the Acoustical Society of America* 39: 151-168.
Öhman, S. (1967). "Numerical models of coarticulation." *Journal of the Acoustical Society of America* 41: 310-320.
Pettersson, T. and S. Wood. (1987). "Vowel reduction in Bulgarian." *Folia Linguistica* 21(2-4): 263-279.
Pyun, K.-S. (1987). *Korean-Swedish Interlanguage phonology.* Koreanological Studies 2. Stockholm, Institute of Oriental languages, University of Stockholm.
Radchenko, G. (MS). Acoustic features of creaky and breathy voice in Udehe.
Strangert, E. (1990a). Pauses, Syntax and Prosody. Nordic Prosody V. 294-305.
Strangert, E. (1990b). Where do pauses occur in texts read aloud? Proceedings from the Twelfth Scandinavian Conference of Linguistics. Reykjavikw.
Strangert, E. (1991a). Pauses in texts read aloud. XII International Congress of Phonetic Sciences. 4: 238-241.
Strangert, E. (1991b). "Perceived pauses, silent intervals and syntactic boundaries." *Phonum* 1: 35-39.
Strangert, E. (1991c). Phonetic characteristics of professional news reading. Fonetik '91.
Svantesson, J.-O. (1989). "Tonogenetic mechanisms in Northern Mon-Khmer." *Phonetica* 46: 60-79.
Svantesson, J.-O. (1991). Vowel palatalization in Mongolian. XII International Congress of Phonetic Sciences. 5: 102-105.

Svantesson, J.-O. (MS). Initial consonants and phonation types in Shanghai.
Touati, P. (1987). *Structures prosodiques du suédois et du français.* Travaux de l'institut de linguistique de Lund. Lund, Lund University Press.
Touati, P. (1991). Analyse de la prosodie de la parole spontanée en Suédois et en Français. XII International Congress of Phonetic Sciences. 4: 282-285.
Traunmüller, H. (1981). "Perceptual dimension of openness in vowels." *Journal of the Acoustical Society of America* 69: 1465-1475.
Traunmüller, H. (1988). "Paralinguistic variation and invariance in the characteristic frequencies of vowels." *Phonetica* 45: 1-29.
Traunmüller, H. and F. Lacerda. (1987). "Perceptual relativity in identification of two-formant vowels." *Speech Communication* 6: 143-157.
Wood, S. (1979). "A radiographic analysis of constriction locations for vowels." *Journal of Phonetics* 7: 25-43.
Wood, S. (1986). "The acoustical significance of tongue, lip, and larynx maneuvers in rounded palatal vowels." *Journal of the Acoustical Society of America* 80(2): 391-401.
Wood, S. (1988). "Vowel quanitity and syllable structure in Welsh." *Working Papers, Lund University, Dept of Linguistics* 33: 229-236.
Wood, S. (1991a). Vowel gestures and spectra: from raw data to simulation and applications. XII International congress of phonetic sciences. 1: 215-219.
Wood, S. (1991b). "X-ray data on the temporal coordination of speech gestures." *Journal of Phonetics* in press:

Dynamic aspects of English vowels in /bVb/ sequences

Keith Johnson

Abstract

Analysis of X-ray microbeam recordings of 5 speakers pronouncing /bVb/ sequences revealed that vowels in midwestern American English differ from each other in terms of lip, tongue, and jaw dynamics as well as in terms of "target" positions. The data supported Lehiste & Peterson's (1961) distinction between short and long-nucleus vowels. Short-nucleus vowels /ɪ, ɛ, ʌ/ had shorter deceleration phases during the lip opening movement and shorter acceleration phases during the lip closing movement. A kinematic analysis of the consonant opening and closing movements suggested that in a spring/mass model of articulator movement these short vowels would be characterized by greater spring stiffness. /ɔ/ had less spring stiffness during the opening gesture and /oᵁ, eᴵ, æ/ and /ɔ/ had less spring stiffness during the closing gesture. A canonical discriminant analysis of articulator positions across time found consistent patterns of tongue movement which separated the vowels into the same groups found in the kinematic analysis of lip movement. Each of the four factors found in the analysis was associated with a movement pattern. Additionally, the first two factors were associated with gross tongue location differences and the third factor was associated with tongue bunching. These analyses suggest that the dynamic control of the lip gestures in /bVb/ sequences is coordinated with tongue movement patterns for vowels.

1. Introduction

In addition to spectral distinctions (Peterson & Barney, 1952), American English vowels differ in acoustic duration and formant trajectory patterns. Low vowels such as the vowel in "hod" have longer durations than do high vowels such as the vowel in "heed", and the vowels in "hid", "head", "hood", and "hud" have shorter durations, respectively, than those in "heed", "heyed", "who'd", and "hod" (Lehiste & Peterson, 1961). Lehiste & Peterson also noted that the vowels in "hid", "head", "hood", and "hud" had relatively shorter F2 steady-states than did the other vowels. Therefore, they classified /ɪ, ɛ, ʊ/ and /ʌ/ as short-nucleus vowels, /i, u, eᴵ, oᵁ, ɑ, ɔ, æ, aᵁ, aᴵ/ and /oᴵ/ as long-nucleus vowels. The long-nucleus vowels could be further divided into simple- and complex-nucleus vowels, and among the complex-nucleus vowels, Lehiste & Peterson identified /eᴵ, oᵁ/ and /ɚ/ as single-target vowels and /aᵁ, aᴵ/ and /oᴵ/ as double-target vowels. The double-target vowels typically had two F2 steady-states, while the others did not. Stevens, House & Paul (1966) found that vowels produced in CVC sequences in which the initial and final consonants were identical did not have symmetric F2 trajectories. So, for example, the F2 trajectory of [gu] was not a mirror image of the F2 trajectory of [ug]. They reasoned that these formant trajectory asymmetries reflected the existence of offglides in the vowel. They suggested, for example, that /i/ and /u/ had peripheral offglides and could be transcribed [iʲ] and [uᵂ], while Lehiste & Peterson's (1962) short-nucleus vowels had central offglides and could be transcribed [Vᵊ]. The distinction between simple and complex nucleus vowels has also been a feature of linguistic descriptions of American English since at least 1933 (Bloomfield, 1933, p. 104, 124; Trager & Smith, 1951, p. 12 ff.; Chomsky & Halle, 1968) although the particulars of each successive analysis vary.

There is increasing evidence that listeners expect and make use of dynamic information in vowel perception (Strange, 1989). Huang (1986) and DiBenedetto (1989a,b) found that the temporal location of the peak in the F2 trajectory has an impact on the categorization of some American English vowels. Strange, Jenkins & Johnson (1983) found that listeners' vowel identification performance was only slightly affected when the vowel centers in CVC syllables were replaced by silence. Parker and Diehl (1984) confirmed this finding. When the middle 70% of the vowel was replaced by silence, vowel identification error rates were about 20% for short

vowels and less than 10% for long vowels (the comparable error rates in the full vowel condition were 5% and 3%). Even with 90% of the vowel replaced by silence (only about 10-15 ms of the vowel onset and 10-15 ms of the vowel offset remaining audible) Parker & Diehl found that listeners' performance was well above chance. These results were taken to indicate that vowel formant transitions provide valuable perceptual information, which listeners readily use. This conclusion was strengthened by Verbrugge & Rakerd's (1986) silent-center vowel study. They constructed silent center vowel stimuli by splicing together initial (or final) transitions taken from vowels produced by men, with final (or initial) transitions produced by women (and an appropriate amount of silence between the two portions). They found that these hybrid male/female silent center vowels were identified just as accurately as were single-speaker silent center vowels. This result suggests that 'target' formant frequencies may be less important in vowel perception than are the direction and rate of change of formant trajectories, and thus that spectral change in an important perceptual property for vowels in English. Nearey and Assman (1986) also came to this conclusion. They constructed stimuli from naturally produced isolated vowels by extracting a 30 ms portion from early in the vowel and another 30 ms portion from late in the vowel. Listeners could correctly identify the vowels when these vowel portions were played in the original order, but if the first portion was played twice or the two portions were played in the opposite order the listeners' performance dropped dramatically.

The two-target representation proposed in linguistic descriptions of American English fits several aspects of these acoustic and perceptual studies of vowels. First, some of the duration differences among American English vowels may reflect a difference between vowels with two vowel targets and vowels with one vowel target. Similarly, Lehiste & Peterson's (1961) distinction between short-nucleus and long-nucleus vowels can be described in terms of the number of articulatory targets involved in producing the vowel. Note, however, that this is not Lehiste & Peterson's (1961) interpretation. They distinguished between three types of long-nucleus vowels only one of which, in their view, had two vowel targets (the "true" diphthongs [a^{I}, a^{U}] and [o^{I}]). We will return to this point in the conclusion. A two-target model of vowel articulation in American English is obviously relevant for Nearey and Assman's (1986) perceptual study, and may also provide an explanation for the relative importance of vowel edges as opposed to vowel centers found in the silent center studies reviewed by Strange (1989). In addition, a phonetic distinction between one and two-target vowels corresponds to a distinction which must be made in view of some phonological phenomena. For instance, phonetically long vowels may occur in open syllables such as "bee", "bay", "spa", "law", "go", "do", and [bæ] (the noise a sheep makes), and in open upbeat syllables with secondary stress such as in the words "recede", "Daytona", "tautology", "rotation", "bubonic" and "Camay" while phonetically short vowels may not occur in these environments.

In contrast to the acoustic and perceptual studies which suggest that changes in formant trajectories during American English vowels are linguistically significant, many previous studies of vowel production have focussed on articulatory target positions during vowels and various sources of variability for these targets assuming that a vowel in American English is specified by a single articulatory target. For instance, Kent & Netsell (1971) reported the effects of linguistic stress on tongue, jaw and lip positions at the acoustic midpoints of vowels with some illustrative data on articulatory dynamics of stress distinctions. Kent & Moll (1972b) studied vowel-to-vowel coarticulation, reporting movement trajectories from one vowel target to another with various consonants or linguistic boundaries intervening. Ladefoged, DeClerk, Lindau & Papçun (1972) studied individual differences in tongue shapes for vowel targets, and their data were further analyzed in terms of tongue shape factors by Harshman, Ladefoged & Goldstein (1977). Gay (1974) noted the effects of consonant / vowel coarticulation, vowel-to-vowel coarticulation, and speaking rate on the positions of the tongue, lips and jaw at the point of maximum articulator displacement during vowels. Perkell & Nelson (1982) investigated variability in tongue positioning during vowel production as a function of the place of maximal constriction, looking at

one time slice for each vowel (the point of extreme movement toward the vowel target). Jackson's (1988) cross-linguistic study was based on tongue shape data taken at the vowel midpoint. All of these studies have in common that they characterize the articulation of American English vowels in terms of a single vowel target.

So, acoustic phonetic studies indicate that the vowels of English have discernable dynamic properties and perceptual studies indicate that these dynamic properties are important for speech perception. Also, the traditional linguistic analysis of American English vowels (going back to Bloomfield, 1933) suggests that the vowels can be separated into those which have two vowel targets and those which have only one. Yet, the dynamics of vowel articulation in American English have not been extensively studied. The experiment reported here addresses this issue by analyzing (1) the kinematics of lower lip movement for different vowels in /bVb/ sequences as produced by speakers of American English, and (2) tongue body movements during those vowels.

2. Method

The data were collected by Peter Ladefoged and Mona Lindau at the x-ray microbeam facility at the University of Wisconsin (Fujimura, Kiritani & Ishida, 1973; Kiritani, Itoh & Fujimura, 1975; Abbs, Nadler & Fujimura, 1988). Some aspects of these data have been reported previously (Lindau & Ladefoged, 1989,1990; Johnson, Ladefoged & Lindau, submitted).

2.1 *Subjects*

Five speakers (3 females and 2 males) of northern midwestern American English served as speakers for the experiment. They were paid a small sum for their participation and were recruited by the staff at Wisconsin from the university community. The subjects were unaware of the specific purposes of the experiment, and reported no history of speech or hearing deficiencies and had no dental fillings. The speakers were screened for dialect homogeneity by having them read several sets of dialect diagnostic words (e.g. "merry", "Mary", and "marry"). One of the male speakers (RP) did not distinguish between /ɔ/ and /ɑ/, so he did not read the /ɔ/ words. For these speakers, /æ/ was diphthongized and could be transcribed as [ɛᵊ] or [eᵊ], and /ɛ/ was transcribed as somewhat lower than in other dialects of American English.

2.2 *Materials*

The speakers read sentences containing symmetric C_iVC_i sequences with the consonants /d, b, s/ and the vowels /i, ɪ, eᴵ, ɛ, æ, ɑ, ɔ, ʌ, oᵁ, ʊ, u/. Not all of these sequences were real words in English and so the subjects were instructed in the pronunciation of the non-English sequences by pointing out words which rhyme with the test sequences. For instance, "beb" [bɛb] rhymes with "Deb" in this dialect. In an attempt to balance the demands of using actual words with the desire for a factorial experimental design, some of the words had CVC structure and some had CV structure with the following word in the carrier phrase supplying the final C of the sequence. For example, instead of being asked to read, "Say beeb (/bib/) between", the subjects read, "Say be between". The different syllable structures complicated the interpretation of lip closing kinematics as discussed below. A full list of the sentences is presented in Table I. Analyses of the /bVb/ sequences are reported below. This subset of utterances was chosen on the assumption that in them tongue movement would be relatively free from consonant effects and thus more easily interpretable as reflecting vowel articulations (Engstrand, 1988; Nord, 1975).

2.3 *Procedure*

Small (2.5 mm) gold pellets were glued to the speakers' lips, teeth, and tongue along the midline of the vocal tract (Figure 1). Two additional pellets tracked head movement. These pellets were glued to the bridge of the nose and to the border of the upper incisor and gums. One pellet was glued to the border of the lower incisors and the gums and indicated the location of the jaw. The lip pellets were glued to the borders of the vermilion ridges of the upper and lower lips. The tongue pellets were placed at intervals of approximately 15 mm on the protruded tongue with the

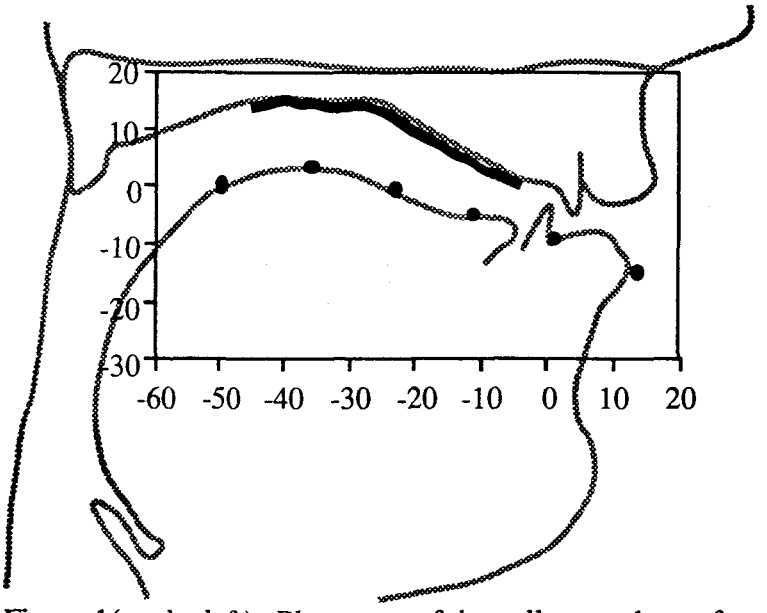
Table I
List of materials.

Say	dee	to me.	Say	bee	between.	Say	see	serenely.
	did			bib			sis	
	day			bay			say	
	dead			beb			cess	
	dad			bab			sass	
	Dodd			bob			soss	
	daw			baw			saw	
	doe			boe			sew	
	dood			----			soos	
	do			boo			sue	
	dud			bub			suss	

first pellet about 8-10 mm behind the tongue tip. Figure 1 shows the locations of the pellets at the midpoint of the vowel averaged across all vowels, consonants, and speakers and shows that when the tongue was not protruded the pellets were on average about 10 mm from each other. As the talkers read the experimental materials the movements of the pellets were tracked by a computer controlled x-ray system (Nadler, Abbs & Fujimura, 1987). A small beam of x-ray tracked each pellet, and the locations of the pellets (in both the vertical and horizontal dimensions) were recorded at intervals of 10 ms (for tongue, lower lip and nose) or 20 ms (for jaw and upper lip). Accuracy of the measurements was on the order of fractions of a millimeter. Additionally, the speech wave form was simultaneously sampled at a rate of 10 kHz.

Each sentence was repeated three times in a given recording run and the entire procedure was performed twice by each subject giving a total of six repetitions of each CVC sequence per subject. Thus, the total number of possible utterances was 960. Of this number 300 contained /bVb/ sequences and only 202 utterances (67% of the total) were available for analysis due to various types of experimental error. The statistical analyses therefore were based on unequal numbers of observations of the different vowels. Most of the missing observations were due to missing data collection runs, therefore for some of the tokens for some subjects only three observations per vowel were available.

After the data had been collected, the nose and upper incisor pellet traces were used to correct for head movements, rendering the other pellet traces in terms of movement relative to the speaker's occlusal plane rather than absolute movement. Five events were located in the two dimensional movement trajectory of the lower lip for each CVC sequence (Figure 2): (1) the point of maximum displacement toward consonant closure during the initial consonant, (2) the point of maximum speed (the change in displacement in two dimensions per unit time) from consonant closure to vowel opening , (3) the point of maximum vowel opening, (4) the point of maximum speed from vowel opening to the final consonant closure and (5) the point of maximum displacement toward consonant closure during the final consonant. A computer program located these articulatory events for each vowel utterance and recorded (1) the times of the events, (2) the locations of the pellets at each event, (3) the magnitudes of the opening and closing gestures, and (4) the peak speed values of the opening and closing gestures. This program has been previously described (Johnson, Ladefoged & Lindau, submitted) and will be briefly summarized here. For each utterance, the time of the acoustic onset of the vowel had been previously identified by eye in a digital wave form display and stored in a computer file. The measurement program looked at the lower lip trajectory in a window of time around the acoustic onset of the vowel and found the locations of maximum displacement and speed. The trajectories were evaluated in two dimensions, so maximum consonant displacement was defined as the point at which the lower lip was furthest



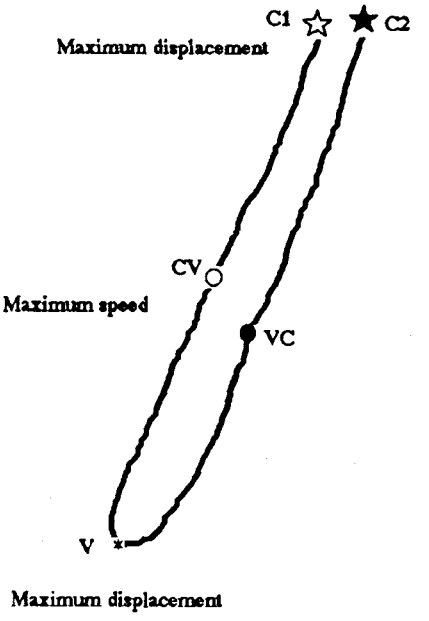


Figure 1(on the left). Placement of the pellets on the surfaces of the vocal tract.
Figure 2 (on the right). Schematic representation of the articulatory landmarks at which pellet locations were measured. This figure represents the locations of maximum displacement and speed of the lower lip. C1 = point of maximum displacement during the initial /b/. V = point of maximum displacement during the vowel. C2 = point of maximum displacement during the final /b/. CV = point of maximum speed during the opening movement. VC = point of maximum speed during the closing movement.

up and forward and the maximum vowel displacement was defined as the point at which the lower lip was furthest down and back. The time of the maximum vowel displacement measured in this way was not reliably different from previous measurements made from visual displays of the lower lip vertical movement (Lindau & Ladefoged, 1990).

3. Lower lip movement

This section describes vowel effects in the movement trajectories of the lower lip in /bVb/ sequences. Each trajectory can be summarized in terms of the durations of the component movements and the displacement amplitudes and peak velocities of the lower lip.

Figure 3 shows the relative times of the articulatory landmarks illustrated in Figure 2 for each vowel averaged across speakers. Each opening movement is composed of an acceleration phase (C1 to CV) and a deceleration phase (CV to V). Similarly, each closing movement is composed of an acceleration phase (V to VC) and a deceleration phase (VC to C2). The vowel nucleus can be defined as the portion of the trajectory extending from CV to VC. Based on the data in Figure 3 the vowels of English can be divided into two classes; short-nucleus /ɪ, ɛ, ʌ/ and long-nucleus vowels /i, e^{I}, æ, ɑ, ɔ, o^{U}, u/ (see Lehiste & Peterson, 1961 for acoustic evidence for this distinction, and that /ʊ/ is also a short-nucleus vowel). The short-nucleus vowels had shorter overall durations and also shorter nuclei. Within short-nucleus vowels, the opening acceleration (C1 to CV) was longer than the opening deceleration (CV to V) and the closing deceleration (VC to C2) was longer than the closing acceleration (V to VC). The opposite pattern occurred in long-nucleus vowels (longer opening deceleration and longer closing acceleration). Across vowels, the duration of the opening acceleration was correlated with movement amplitude and did not separate the vowels into short versus long-nucleus. The duration of the opening deceleration, on the other

hand, was not correlated with movement amplitude; the short-nucleus vowels having shorter opening decelerations than the long-nucleus vowels. The closing decelerations of short-nucleus vowels were longer than those of the long-nucleus vowels. This seemed to be true regardless of syllable structure; "bab" and "Bob" both had short closing decelerations as did "baw", "bow", "boo", etc. rather than long closing decelerations as in "bib", "beb", and "bub". The long opening deceleration for /ɔ/ seems to have been the result of was more lip rounding early in the vowel than late [ɔᵊ], pushing back the point of maximum lip opening (V). Conversely, the maximum displacement (V) during /oᵘ/ was also not in the center of the vowel nucleus, suggesting greater rounding at the end of the vowel than at the beginning.

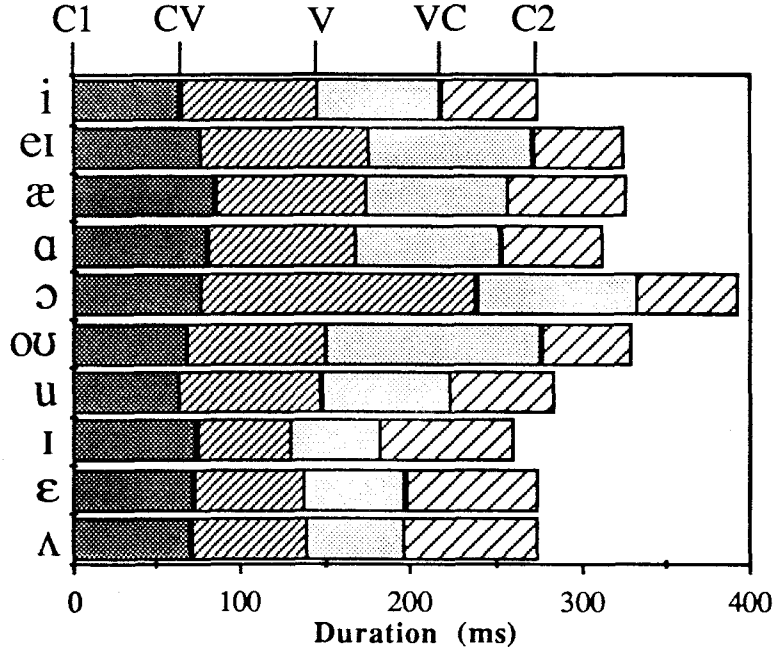


Figure 3. Average durations of lower lip trajectories during /bVb/ sequences. The labels at the top of the figure refer to the articulatory landmarks illustrated in Figure 2. The opening gesture extends from C1 to V, the closing gesture from V to C2. The opening acceleration is C1 to CV, the opening deceleration is CV to V. The closing acceleration is V to VC, and the closing deceleration is CV to C2.

Figure 4 shows kinematic data from the opening and closing movements. The vowel symbols in this figure represent the average values for that vowel. Following Beckman, Edwards & Fletcher (1991) these data can be interpreted in terms of a spring/mass dynamic system (Saltzman & Munhall, 1989). The positive correlation between opening displacement amplitude and peak speed (left panel) agrees with Kent & Moll's (1972a) observation that the further an articulator must move the faster the movement, and suggests that the opening movements differ primarily in terms of an underlying amplitude parameter for the opening gesture. In addition, the distinction between /ɔ/ and the other vowels appears to involve articulator stiffness. /ɔ/ lies below the regression line indicating that on average its opening movement was accomplished more slowly than was the opening movement for other vowels with similar displacement amplitudes. In a spring/mass model this pattern can be simulated by reducing spring stiffness. The short-nucleus vowels /ɪ, ɛ, ʌ/, on the other hand, can be characterized as having *greater* spring stiffness than the others. /ɪ, ɛ, ʌ/ lie above the regression line in the left panel of Figure 4. So, the duration differences illustrated in Figure 3 can be interpreted in terms of a spring/mass dynamic model. Differences in the duration of the opening acceleration were associated with differences in

movement amplitude in the model (which is at odds with a model in which the vowels are distinguished solely by gestural amplitude), and differences in the duration of opening deceleration were associated with spring stiffness in the model.

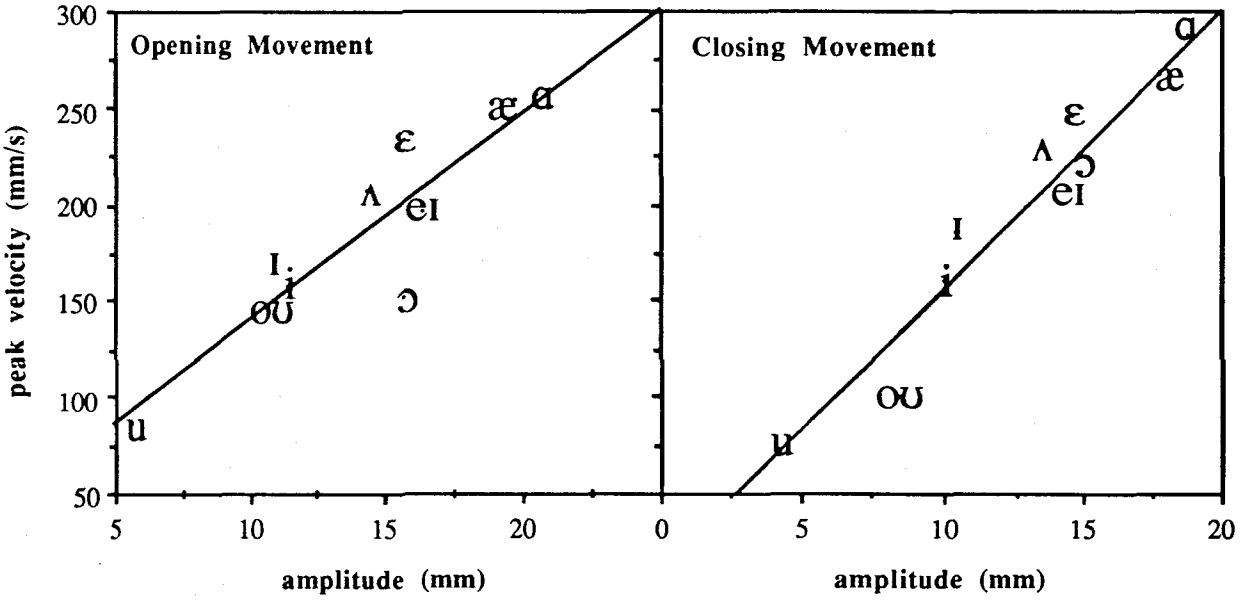


Figure 4. Opening movement and closing movement kinematics. Each symbol represents the values for the indicated vowel averaged across speakers. The regression lines were calculated from the vowel averages.

The right panel of Figure 4 shows kinematic data from the closing movements. As before, the vowels appear to differ primarily in terms of gestural amplitude, but also as before there appear to be some interesting stiffness differences. Like the opening movement of /ɔ/, the closing gesture of /oᵁ/ appears to be characterized by lower stiffness; it had low speed relative to its displacement amplitude. To a lesser degree this seems to be the case also for /eᴵ, æ/ and /ɔ/. The closing gestures of these three vowels all had lower velocities than other vowels having about the same displacement amplitudes (i.e. they lie below the regression line in the right panel of Figure 4). In addition, the closing movements of the short-nucleus vowels had higher velocities than other vowels having about the same displacement, and so could be characterized as having greater stiffness in a spring/mass model. Whereas it was possible to associate changes in stiffness with the duration of the opening deceleration and changes in gestural amplitude with the duration of the opening acceleration, the relationship between the kinematic model parameters of the closing gestures and the durations of the closing movement's components is not obvious. Increased stiffness for /ɪ, ɛ, ʌ/ was associated with longer closing decelerations and shorter closing accelerations, while decreased stiffness for /oᵁ, eᴵ, ɔ, æ/ tended to be associated with longer closing accelerations. Gestural amplitude was not correlated with total closing gesture duration nor with either of the movement phases. The interpretation of these data is complicated by the fact that syllable structure is a confounding variable in this study (some sequences having CV structure and some having CVC structure). Also, there may have been some aspects of tongue movement which constrained the lip gestures and placed vowel-dependent constraints on their temporal structure. We turn now to this topic.

4. Vocal tract configurations

This section describes the results of a factor analysis of articulatory movements in /bVb/ sequences. The acoustic and perceptual studies mentioned in the introduction suggest that dynamic

information may be important in maintaining vowel distinctions. The factor analysis was designed to identify reliable patterns of movement during the vowels. The locations of six pellets at three times during each vowel were entered into a canonical discriminant analysis (Kshirsagar, 1972; SAS Institute, 1982). The analysis gave a derived vowel space and factor loadings for each of the dimensions of the derived space. These factor loadings can be translated back into the articulatory space and can be interpreted as abstract articulatory patterns involved in vowel production.

The average locations of the pellets at three times are shown in Figure 5. The legend in this figure and others to follow refers back to Figure 2. The three times are the points of maximum lower lip speed during lip opening (CV) and lip closing (VC), and the point of maximum lower lip displacement during the vowel (V). The four pellets on the tongue will be referred to (from front to back) as tongue tip, tongue body 1, tongue body 2, and tongue dorsum. Figure 5 shows that, averaged across vowels and speakers, the lower lip showed a displacement of about 9 mm from CV to V, and about 6 mm from V to VC. The jaw showed similar directions of movement with much smaller magnitudes and the tongue pellets reflected the same pattern with decreasing magnitudes further back in the mouth. The average tongue dorsum pellet location showed almost no movement averaged over the vowels. The carrier phrase was "say /bVb/ between", with the vowel [eᴵ] preceding and [i] following the test word. Figure 5 suggests that vowel-to-vowel coarticulation did not produce a unique movement pattern of the tongue in this context. The tongue movements indicated in the figure appear to be due solely to the movement of the jaw which presumably was coordinated with the lower lip in producing the bilabial closures. It is not clear whether the asymmetry in the lip positions was due to the context in which the words occurred or whether this pattern would be observed in any context.

Figure 6 shows the average pellet positions for the ten vowels. As in other studies of vowel articulation (Stevens & House, 1955), the vowels were separated from each other by differences in the location and degree of vocal tract constriction at the center of the vowel. The two back tongue pellets were higher during /æ/ than during /ɛ/, while the lip opening was greater for /æ/. As mentioned earlier, /æ/ for these speakers was impressionistically transcribed as [ɛᵊ].

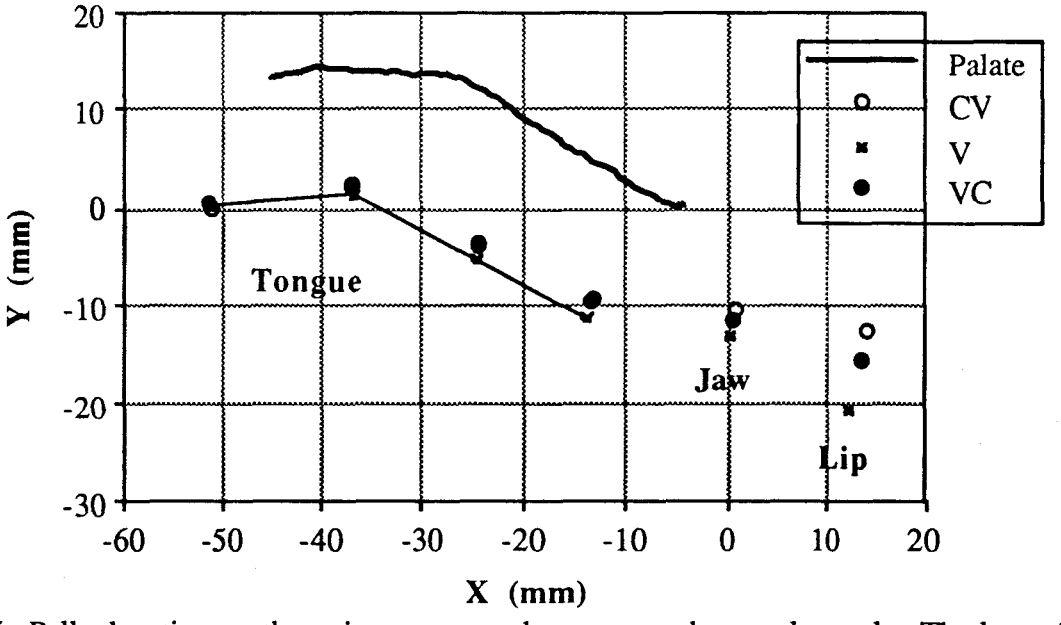


Figure 5. Pellet locations at three times averaged across speakers and vowels. The legend refers to the articulatory events illustrated in Figure 2.

Figure 6 also indicates that there were significant movement patterns associated with several of the vowels. /eᴵ/ and /ɔ/ had tongue raising and some tongue fronting during the vowel. /æ/ had tongue retraction and lowering, /oᵁ/ and /u/ had tongue retraction and raising, and /ɛ/ and /ɪ/ had some tongue lowering.

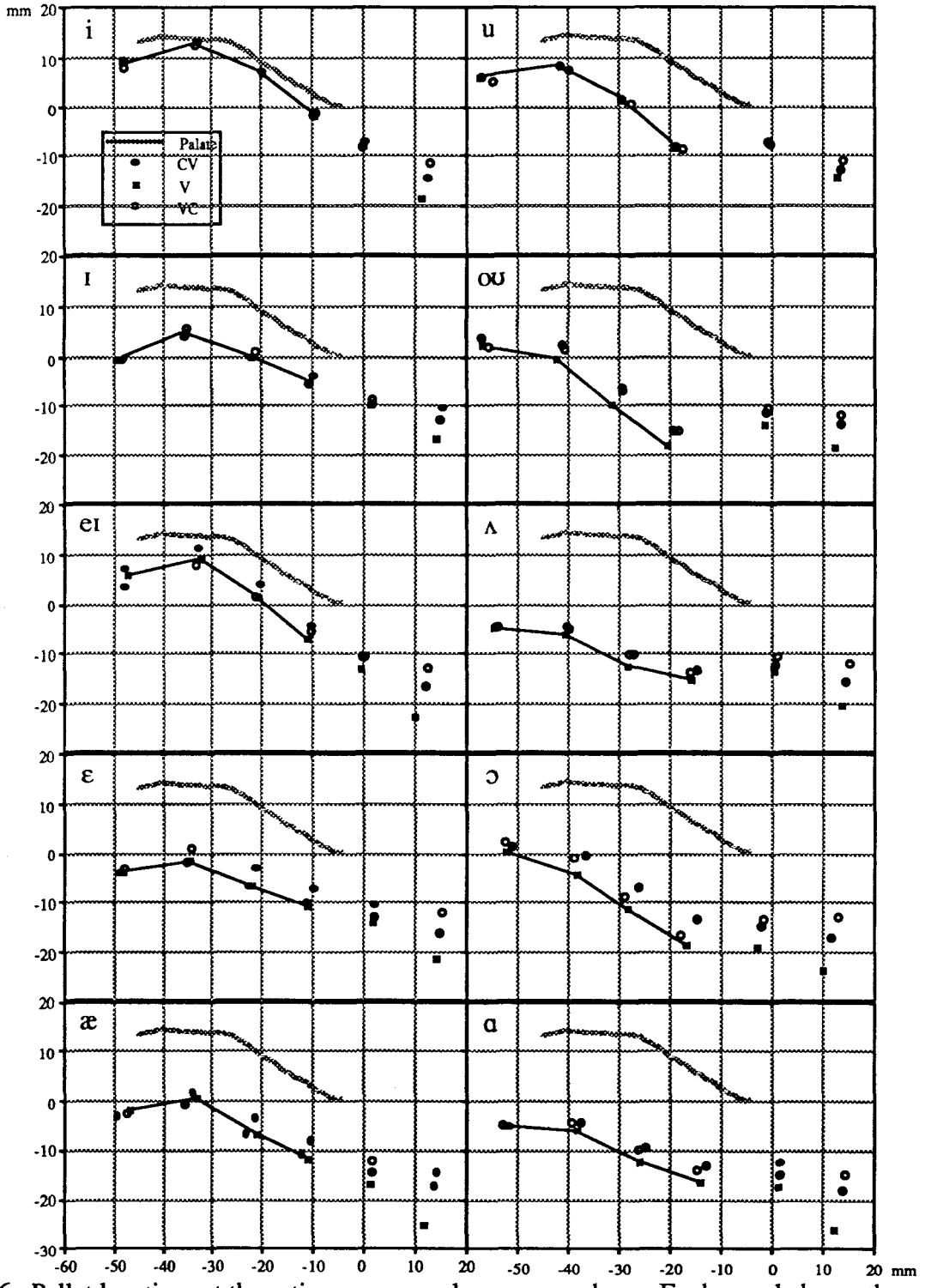


Figure 6. Pellet locations at three times averaged across speakers. Each panel shows the average pellet locations for a particular vowel.

Rather than rely on subjective impressions about the data in Figure 6, canonical discriminant analysis was used to explore general patterns of vocal tract posture and movement. Canonical discriminant analysis finds principal components along which categories can be best discriminated. "Given a classification variable and several quantitative variables, canonical discriminant analysis derives canonical variables (linear combinations of the quantitative variables) that summarize between-class variation in much the same way that principal components summarize total variation" (SAS Institute, 1982). In the analysis reported here, the classification variable was vowel identity and the quantitative variables were the pellet locations at three times. Thus for each of the 202 observations there were (6 pellets X 2 dimensions X 3 times =) 36 quantitative variables and one classification variable. The analysis found coefficients for a set of equations of the form:

$$(1) \qquad v_j = x_1 a_{1j} + x_2 a_{2j} + \ldots + x_{36} a_{36j}$$

where v_j is the value of canonical variable j, the x_n are the 36 quantitative variables, and the a_{nj} are the canonical coefficients for canonical variable j. For each canonical variable the coefficients are optimized to produce maximal separation between categories (in this case vowels). After finding the first set of coefficients (the set which accounts for the greatest amount of between category variability), coefficients for another canonical variable, uncorrelated with the first, are found. This procedure is repeated N-1 times where N is the number of categories. Because the input quantitative variables are transformed to make the pooled within-class covariance matrix an identity matrix, the canonical variables do not represent perpendicular components through the space of the original variables (a common complaint about principal component analysis). The average scores for a particular vowel on the canonical variables (the v_js) define that vowel's location in the derived vowel space.

The relationship between the canonical variables and the original quantitative variables can be calculated in the following way. First, a scale factor for canonical variable j (c_j) is calculated by (2). Where r_{ij} is the correlation between canonical variable j and quantitative variable i, and b_{ij} (derived from a_{ij}) is the standardized canonical coefficient on canonical variable j for quantitative variable i. The patterns of variation (x_{ij}) in the original quantitative variables which are encoded by the canonical variables are then given by (3). Where sd_i is the standard deviation of quantitative variable i and ave_i is the average value of quantitative variable i. If in formula (3) we set v_j equal to a large positive value (within the range of observed values for v_j), the x_{ij} give the pattern of values on the original quantitative variables associated with positive values of canonical variable j. This procedure was used to calculate the articulator loadings shown in Figures 9 through 12. We derive predicted values of the quantitative variables for a particular vowel by setting the v_j in (3) equal to the observed v_j for that vowel and summing the z-score component of (3) over j before multiplying by the standard deviation and adding the mean (4). For the sake of continuity with previous research, the canonical variables will be called "factors".

$$(2) \qquad c_j = \sum_{i=1}^{36} r_{ij} b_{ij}$$

$$(3) \qquad x_{ij} = (r_{ij} v_j / c_j) sd_i + ave_i$$

$$(4) \qquad predicted_i = \sum_{j=1}^{n} (r_{ij} v_j / c_j) sd_i + ave_i$$

Factor analysis has been used previously to identify basic tongue shapes in vowels. Harshman, Ladefoged & Goldstein (1977) found that tongue shapes of the vowels of English could be described with just two factors. Jackson (1988) found that similar tongue shape factors underlie vowel production in English and Icelandic. The main difference between these earlier analyses and the one reported here is that lip and jaw position data and data from 3 times during each vowel were included in the analysis. Consequently, the articulator loadings in this analysis represent underlying patterns of articulator movement which distinguished the vowels in these particular utterances.

The first two factors together accounted for about 80% of the variance, and with the addition of two more factors 90% of the variance was accounted for. The normal method of determining the number of factors to include in a model is to look for an elbow in the variance accounted for curve. By this metric only the first two factors in the present analysis would be chosen. However, since a canonical discriminant analysis gives a unique solution (i.e. the first factor is the same regardless of how many other factors one cares to look at) and the amount of variance accounted for by the third and fourth factors was fairly large (6.9% and 6% respectively, both $p < 0.001$), I will discuss the vowel space and articulatory loadings for the first four factors.

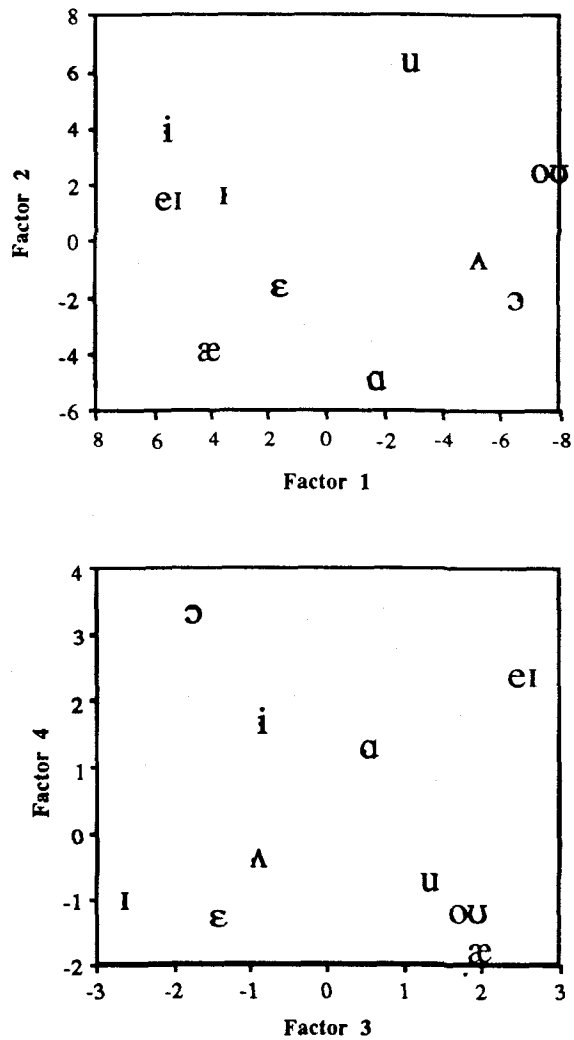


Figure 7. Derived vowel space from the cannonical discriminant analysis. Top: Factor 1 versus factor 2. Bottom: Factor 3 versus factor 4.

Figure 7 shows the four-factor vowel space. The vowel space formed by the first two factors (top panel) is strikingly similar to the traditional impressionistic vowel space. The first factor (the horizontal axis of the top panel) separated the front vowels /i, ɪ, eᴵ, ɛ/ and /æ/ from the back vowels /u, oᵁ, ʌ, ɔ/ and /ɑ/, and the second factor corresponded to the traditional high/low distinction, with /u/ and /ɑ/ having the most extreme values. Note that the short-nucleus vowels were less peripheral in this space than were the long-nucleus vowels, indicating that they were produced with less extreme versions of the first two factors.

The articulatory factor loadings for the first factor (Figure 8) confirmed that vowels with positive values on the factor (top panel) had fronter tongue positions than vowels with negative values on the factor (bottom panel). For instance, the tongue dorsum pellet was about 48 mm behind the upper incisor pellet (the origin of the coordinate space) when factor 1 had a large positive value (top panel), and was about 56 mm back when factor 1 was negative. The tongue shape encoded by the first factor is similar to Harshman et al.'s (1977) "front raising" factor for tongue shapes. Both in Harshman et al.'s analysis and in the present analysis, the first factor distinguished between vowels which had a point of maximum constriction in the front of the mouth (alveolar or post-alveolar) with vowels which had a point of maximum constriction in the pharynx (this must be infered for the x-ray microbeam data, see Lindau & Ladefoged, 1989 concerning this inference). Jackson (1988) also found a front raising component in the production of Icelandic vowels. There are some differences in the detailed tongue shapes found by Harshman et al. (1977), Jackson (1988) and the present analysis, but the general characteristics of the solutions are in agreement.

There are two types of data represented in the present analysis which were not included in Harshman et al. (1977) or Jackson (1988). These are (1) lip and jaw positions and (2) movement over time. Comparison of the jaw pellet in the upper and lower panels of Figure 8 reveals that there was very little difference in jaw position associated with factor 1 (especially as compared with the differences in jaw position associated with the articulatory pattern for factor 2, Figure 9). So, the tongue positions associated with factor 1 differed relative to an essentially fixed jaw. This observation implies that the vowels /eᴵ/ and /oᵁ/ (which differed on factor 1 but not on factor 2) had essentially the same degree of jaw opening while having differing tongue positions. Figure 6 verifies this prediction of the analysis. The lower lip position associated with positive values of factor 1 is relatively low (as compared to the position of the jaw) while the lower lip position associated with negative values of factor 1 was higher and more protruded from the jaw. This indicates negative values of factor 1 were associated with rounded lips.

Figure 8 also shows movement patterns associated with the first factor. The movement patterns found in the factor analysis include vowel and consonant components. Therefore, in evaluating the movement patterns derived in the factor analysis it is necessary to keep in mind the average pattern of movement (Figure 5) for vowels in the /bVb/ context. Some of the tongue movements, particularly of the tongue tip pellet, appear to be jaw related, however, because the tongue dorsum pellet showed no movement in the average pattern, any movement of this pellet in the factor loadings indicate vowel related movements unambiguously. When we compare the movement patterns associated with the first factor to the average pattern of movement, it is apparent that front tongue positions (top panel) were associated with a small degree of tongue backing and tongue body rotation, whereas back tongue positions (bottom panel) were associated with some tongue body and front tongue raising during the closing movement. Generally, however, the degree of within-vowel tongue movement associated with factor 1 was quite small.

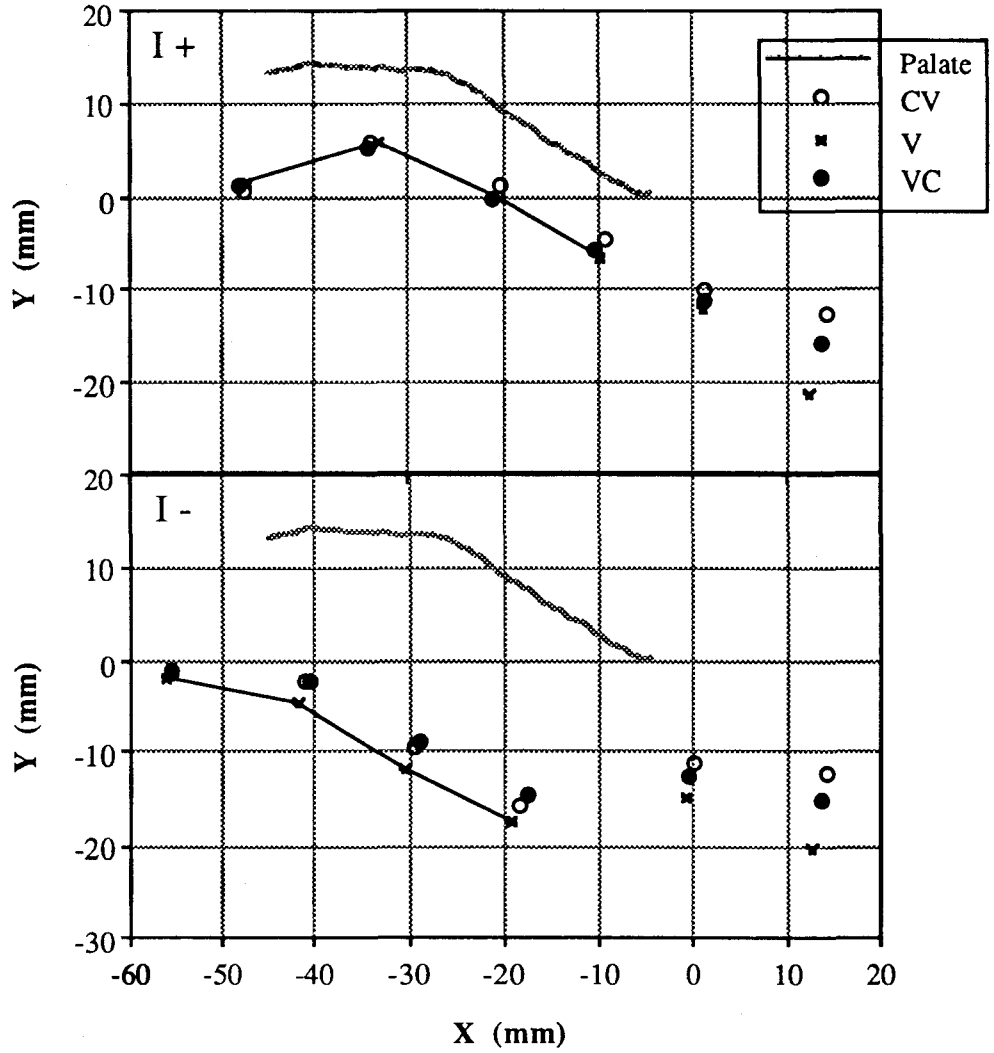


Figure 8. Factor loadings for the first factor. Top: positive loading. Bottom: negative loading.

The articulatory factor loadings for the second factor (Figure 9) had high tongue positions for positive factor values (top panel) and low tongue positions for negative factor values (bottom panel). The tongue loadings on this factor were similar to Harshman et al.'s (1977) "back raising" factor for tongue shapes. Jackson (1988) also found a similar factor for tongue shapes in Icelandic vowels. As was mentioned above, factor 2 was associated with a large difference in the position of the jaw, positive values of the factor were associated with close jaw positions (throughout the vowel) and negative values were associated with open jaw positions. The similarity between factor 2 and Harshman et al.'s (1977) back raising suggests that very open jaw positions (found in the present study) are associated with constriction low in the pharynx (found by Harshman et al.).

As with factor 1, there appears to be a difference in lip rounding associated with factor 2. Positive values of the factor (top panel, Figure 9) were associated with relatively higher and more protruded lip positions at the center of the vowel. Note however that negative values of factor 2 were associated with relatively high lip positions at the vowel edges (CV and VC, in the bottom panel of Figure 9). This may reflect a strategy for attaining a low jaw position in /bVb/ context. If the jaw movement begins before the lip movement (i.e. the lips are help closed while the jaw begins its decent) the amount of time available for jaw movement is greater than it would be otherwise. An adjustment in the relative timing of jaw and lip movement (phase angle change)

seems to be indicated by the pattern of movement found for factor 2. The jaw showed no movement for positive values of factor 2, while the tongue moved up and back (particularly during the opening phase of the lip movement). The articulatory loadings for negative values of the second factor were associated with downward and slightly forward tongue movements during the opening phase.

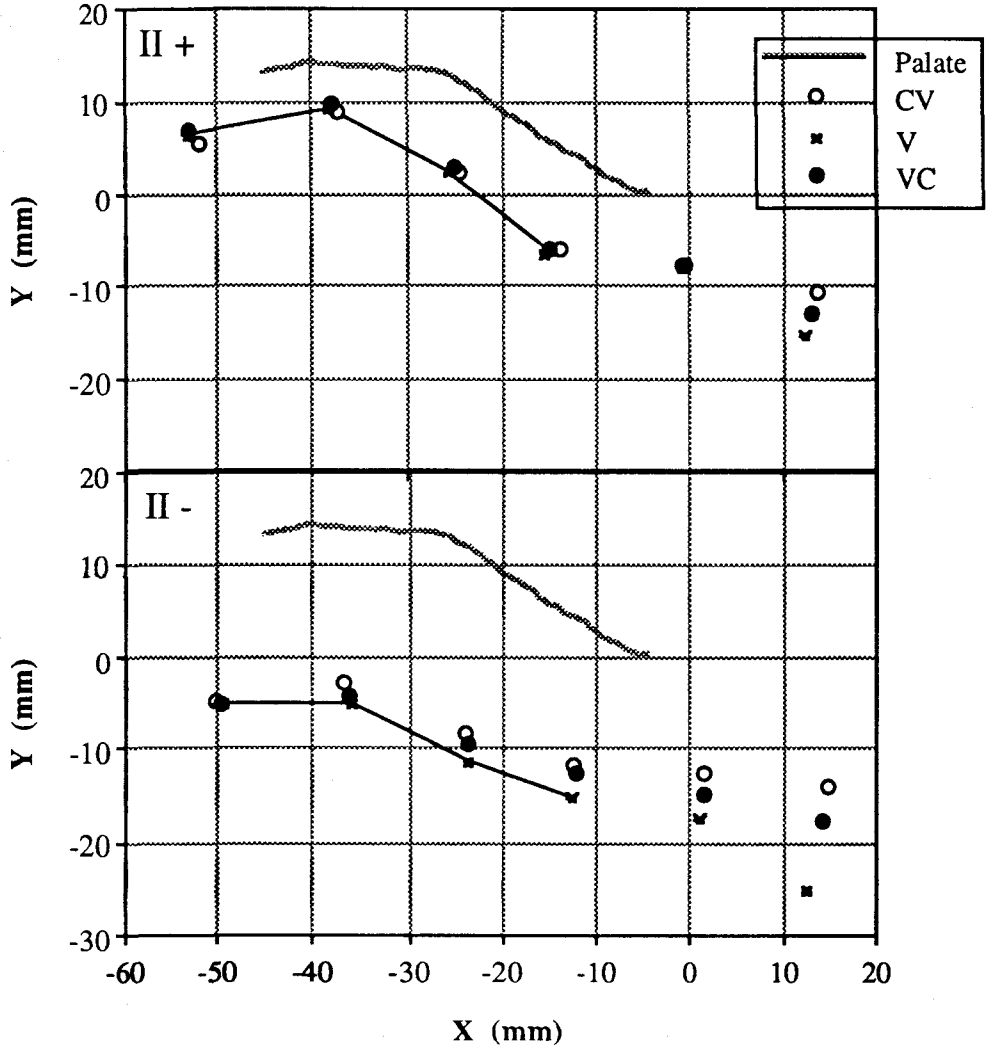


Figure 9. Factor loadings for the second factor. Top: positive loading. Bottom: negative loading.

The bottom panel of Figure 7 shows the average values of the vowels on the third and fourth factors. There are some interesting correspondences between this figure and the lip kinematic data discussed in the previous section. In the lip kinematic data, /ɪ, ɛ/ and /ʌ/ were characterized by stiffer opening and closing lip movements than the other vowels and in Figure 7 they are also separated from the other vowels by having negative values on factors 3 and 4. Both the lip opening and closing movements for /ɔ/ in /bVb/ sequences had reduced stiffness and /ɔ/ was separated from the other vowels by having a large positive value for factor 4 and a negative value for factor 3. /o^{U}, e^{I}/ and /æ/ had less stiff closing movements than the other vowels and also had the largest positive values on factor 3. So, apparently there is some relationship between the kinematic properties of lower lip movement during the vowels and the articulatory patterns found in the factor analysis.

Figure 10 shows the articulatory factor loadings for the third factor. Whereas the first two factors involved large changes in the position of the tongue, the third and fourth factors were associated with more subtle aspects of articulation. The third factor was associated with tongue shape; positive values (top panel) had a more bunched shape than negative values (bottom panel). Also, positive values of the third factor were associated with upward and backward movement of the tongue and negative values were associated with slightly more downward and forward movement than in the average movement pattern (Figure 5). Figure 11 shows the articulatory factor loadings for the fourth factor. As with the third factor, a subtle tongue shape difference was associated with the fourth factor; the tongue dorsum pellet was lower when the fourth factor took a negative value. Positive values on the fourth factor (top panel) were associated with forward and upward movement of the tongue. Negative values of the fourth factor (bottom panel) were associated with a small amount of backward tongue movement and downward movement during the lip opening movement.

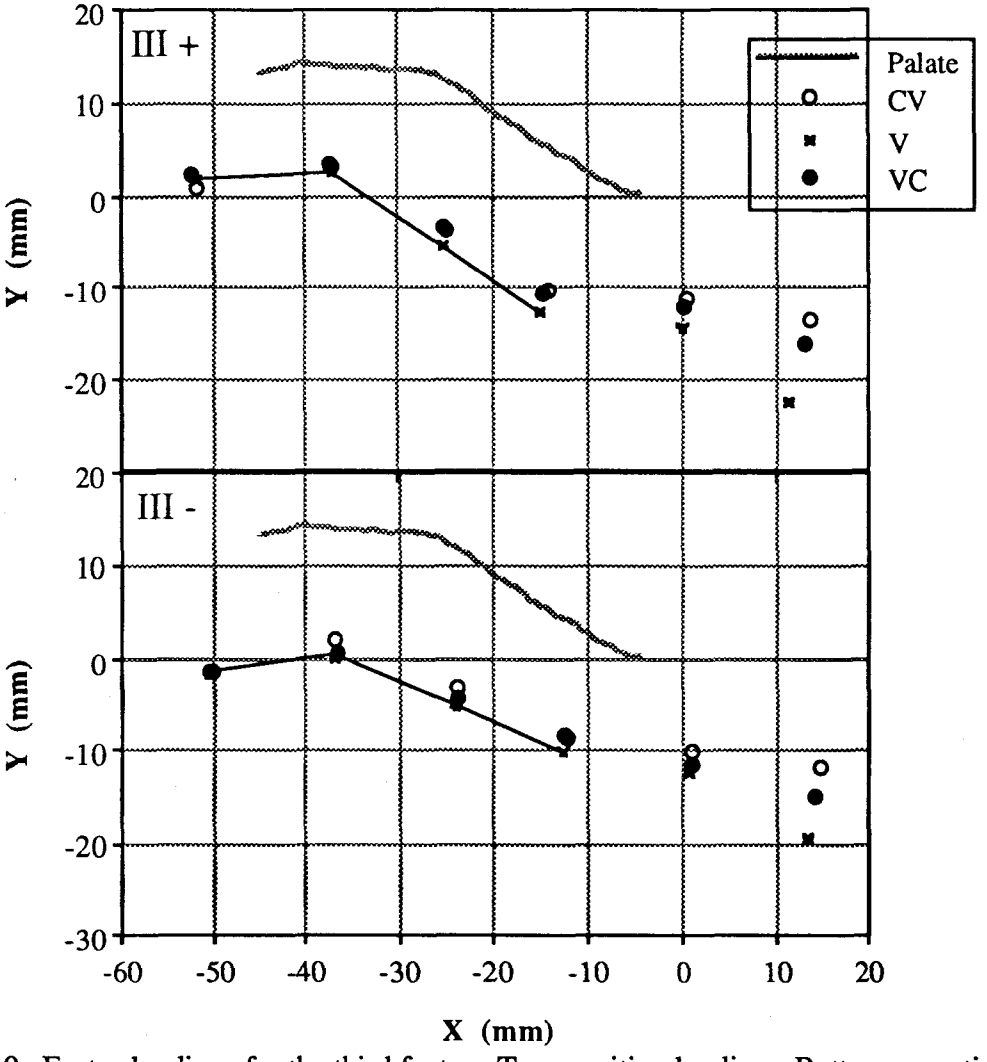


Figure 10. Factor loadings for the third factor. Top: positive loading. Bottom: negative loading.

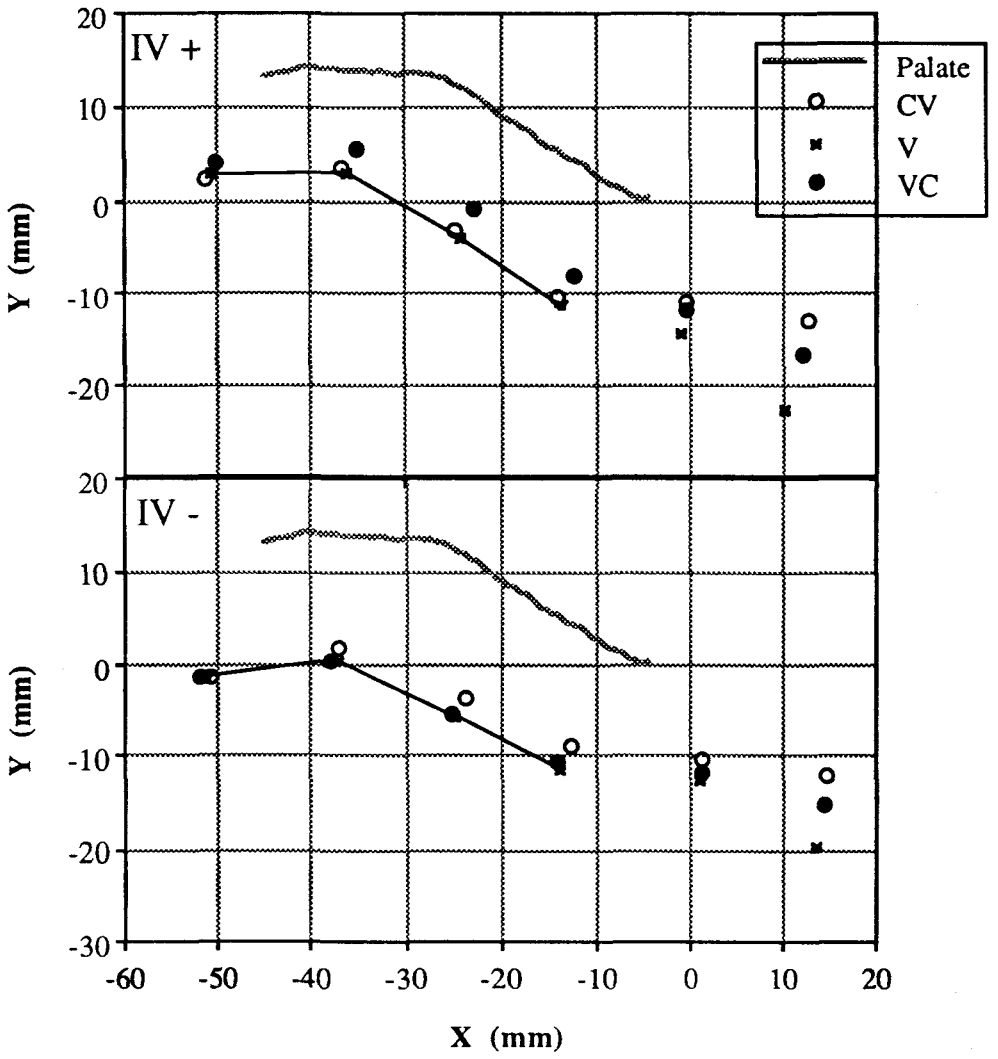


Figure 11. Factor loadings for the fourth factor. Top: positive loading. Bottom: negative loading.

If we focus on the tongue dorsum pellet in Figures 10 and 11, it is apparent that positive values of factors 3 and 4 were associated with greater magnitudes of tongue movement than were negative values. Here a connection between lip kinematics and tongue gestures during vowels can be seen. The vowels which can be characterized as more stiff were produced with less tongue movement than those which would be characterized as less stiff. /ɔ/ had the type of movement captured by factor 4, /æ/ and /o^{U}/ had the type of movement captured by factor 3, and /e^{I}/ had both types of movement. The generalization is this: if there was an appreciable component of tongue movement in the vowel's production, the lip movement was less stiff. The relationship between less stiff opening versus closing movements and this generalization is not obvious. There is no simple relationship like: appreciable movement during the opening phase was associated with a less stiff lip opening movement. This issue must be studied more carefully in another experiment.

Finally note that the within-vowel movement patterns resemble the between-vowel differences in tongue position. When factor 2 was positive (Figure 9, top panel) or when factor 3 was positive (Figure 10, top panel) the tongue dorsum moved up and back during the vowel. This was also the type of difference in tongue position between vowels which had negative values of factor 2 and vowels which had positive values of factor 2. When factor 4 was positive (Figure 11,

top panel), the tongue dorsum moved up and forward. This pattern resembles the between-vowel difference in tongue positions found for factor 1 (Figure 1, top versus bottom panel). These observations suggest that both within and between-vowel movements may be driven by physiological constraints such as muscle lines of action. The extent to which this sort of constraint may limit the set of possible sounds in language is an interesting topic which deserves further research.

5. Conclusion

The act of producing a word involves accessing a memory of how to say the word (a long-term representation) and applying that memory to the vocal organs. The hypothesis implied by saying that a long-term memory representation is 'applied' to the vocal tract is that the representation is composed of a system of constraints (Fowler et al., 1980) for the vocal tract not a sequence of fixed motor commands. For instance, the long-term representation of a word may specify that it must start with the glottis closed. If the word is preceded by another word which ends with a closed glottis, the motor commands at the beginning of the second word is different than it is if the first word ended with the glottis open (see MacNeilage & DeClerk, 1969).

The parameters in a system of articulatory constraints are: (1) the movable part or articulator to which the constraint is addressed, (2) the movement goal or target, (3) kinematic properties or dynamics of the movement, and (4) the patterns of coordination among separate movements. Assuming an articulatory constraints view the description of a speech sound involves specifying the parameters in a system of constraints for the sound. Some aspects of the systems of constraints which may be at work in American English vowel production have been identified in the present study.

5.1 *Vowel Articulators*

In most articulatory synthesizers (Mermelstein, 1973; Rubin, Baer & Mermelstein, 1981; Coker, 1976) the location and shape of the tongue is controlled by specifying the location of the center of the tongue which is represented in the sagittal plane as a circle. This strategy was justified by the observation that "the tongue body moves within the mouth as a rather constant-shaped mass" (Coker, 1976, p. 452). This "single-articulator" approach can be contrasted with an approach to describing vowel articulation which is emerging in linguistic theory (Clements, 1985; Sagey, 1986; McCarthy, 1988; Ladefoged & Halle, 1988). In this approach, three different parts of the tongue are described as active articulators for consonants; coronal [t d s z c ɟ], dorsal [k g x ɣ], and radical [ħ ʕ]. This approach to the description of consonants has been extended recently to the description of vowels (Clements, 1990) giving the vowel classifications; coronal [i], dorsal [u], and radical [a].

Is the tongue during vowel production best described as one articulator or several? Interestingly, factor analyses, both the present analysis and Harshman et al.'s (1977) earlier analysis, support both types of description. We find, as would be expected given its incompressibility, that the shape of the tongue is fairly constant. [u] is not produced with a small flap of the tongue raised toward the velum, rather the mass of the tongue moves. So, Coker's (1976) decision to model the tongue as a single mass is warranted by the x-ray data for vowels. Similarly, however, these data also suggest that for different vowels different parts of the tongue form the primary constriction of the vocal tract. For instance, the configuration associated with positive values of factor 1 (top panel, Figure 8) corresponds very well with Clements' (1990) definition of coronal vowels ("produced with a constriction of the tip, blade or front of the tongue" p. 4). Similarly, the pattern associated with positive values of factor 2 corresponds to the definition of dorsal.

Still, a phonetic desciption of vowel articulation in terms of several functional articulators is to be prefered over the single-articulatory approach for several reasons. First, the geometry of the vocal tract dictates that at least three distinct parts of the tongue are naturally inclined to produce

closures. The fixed walls of the vocal tract for many speakers have two prominent bends separating three relatively accessible regions (accessible in the sense that the tongue can easily approach the passive wall of the vocal tract). The bends occur between the alveolar ridge and the palate and between the uvula and the pharynx wall. The three accessible regions of the vocal tract then are the alveolar ridge, the hard and soft palates, and the pharynx wall. Obviously, the degree to which this is true for any particular speaker depends on the speaker's palate shape (see Hiki & Itoh, 1986). It is easier to produce a constriction along one of these straight portions than it is to produce a constriction in one of the bends. In fact, for many speakers it may not be possible to produce a constriction in one of the bends (with the tongue body mass) without also producing a constriction at the surrounding areas of the vocal tract. Consequently, even though the tongue can be modelled as a circle moving within the mouth, some parts of the tongue are more capable of producing a constriction than are others. Second, the physiology of the vocal tract dictates that some vowel articulations will be more natural than others. Wood (1979) argued that the cross-linguistic preference for the vowels [i], [a], and [u] can be linked to the physiological/kinematic effects of the the genioglossus (to move the tongue forward and up), styloglossus (to move the tongue back and up) and the hyoglossus (to move the tongue back and down). This model disregards the effects of jaw movement on tongue location which, as was seen above, does seem to play an important role in positioning the tongue. Still, Wood's hypothesis fits nicely with the observation noted above that differences in tongue position between vowels were correlated with typical patterns of movement within vowels. Third, nomograms published by Stevens & House (1955) suggest that there are acoustic quantal regions in the vocal tract. The three regions correspond to the three posited articulators coronal, dorsal, and radical. So, vocal tract anatomy, physiology and acoustics seem to conspire to provide three functionally distinct tongue articulators for vowel production.

5.2 *Vowel Targets*

As discussed in the introduction, the vowels /ɪ, ɛ, ʌ, ʊ/ are phonologically distinct from the other vowels of American English. They do not occur word finally or in open upbeat syllables. One way to describe these phonological phenomena is to consider /ɪ, ɛ, ʌ, ʊ/ as having a single vowel target (V) while the other vowels have two targets (VV). For instance the generalization about word final vowels can then be stated in a two-target analysis: words in English must end in heavy syllables, where heavy syllables are defined as having XVC or XVV structure. The articulatory patterns of American English vowels found in this study are consistent with a two target analysis. Single-target vowels are shorter than two-target vowels and have a shorter vowel nucleus. The data show kinematic organization which reflects the impact of vowel-internal dynamics, and this vowel-internal structure can be described in terms of the number and types of targets in the vowel nucleus. Two target vowels in which the targets differ [eᴵ, oᵁ, ɔᵊ, æᵊ] have less stiff lip movements than do two target vowels with identical targets [ii, ɑɑ, uu]. In this analysis the lower lip stiffness differences between vowels are the result of two conflicting demands. First, movement toward the second target must be accomplished before the lips close for the final consonant. Second, English prosody demands that two-target vowel not be twice as long as single-target vowels. These conflicting articulatory demands then lead to a reorganization of the lip movement in which its velocity is reduced to allow extra time during the vowel nucleus for the realization of the second target.

5.3 *Vowel Dynamics*

In Browman & Goldstein's (1986) articulatory phonology the vowels [eᴵ, oᵁ, ɔᵊ, æᵊ] could be implemented by specifying two successive vowel targets during the vowel nucleus. However, it is worth noting that Lehiste & Peterson (1961) found that the vowels in "hide", "how'd", and "hoyd" had two identifiable F2 steady-states, while other long-nucleus vowels did not. This is an interesting observation because the "true" diphthongs involve greater acoustic and articulatory changes during the vowel nucleus than the other long-nucleus vowels, and yet

speakers are more likely to produce two steady-states in the 'true' diphthongs than they are in the other long-nucleus vowels. Lehiste & Peterson's (1961) results suggest that speakers can produce vowels with two F2 steady-states (at normal rates of speech, Gay, 1968), even when the targets are far from each other in articulatory space. This suggests that the kinematic reorganization found here for long-nucleus vowels was not a result of having to produce two targets, but rather the result of simply having to produce a movement.

There are several reasons to suspect that movement itself may become phonologized. First, a two-target model of vowel dynamics requires special statements to account for reduction processes for the second (nonsyllabic) target. Several researchers have noted that the formant values at the ends of complex-nucleus vowels are not the same as the formant values at the center of similarly transcribed short-nucleus vowels (Gay, 1968; Gottfried & Miller, 1991). For instance, Gay (1968) found that [ɪ] at the end of [eᴵ] had more variable and different formant values than the [ɪ] of 'hid'. He concluded 'the targets of /ɔᴵ, aᴵ, aᵁ, eᴵ, oᵁ/ are not necessarily compatible with the vowels used to describe them' (p. 1572). Similarly, in the present data, the offsets of the complex-nucleus vowels did not have the same articulatory positions found for the short-nucleus vowels transcribed with the same symbols. Gay (1968) also found that the formant transitions in complex-nucleus vowels showed much less reduction in duration in fast speech than did vowel steady-states (when steady-states were present at all). Given these observations, then, the phonetics-phonology interface required for the description of vowel targets in complex-nucleus vowels in a two-target model will have to include either separate vowel categories which are only found as non-syllabic vowels in complex-nuclei, or a set of vowel reduction rules which apply only (and obligatorily) to non-syllabic vowels in complex-nuclei. Second, for some voices (women and children primarily) formant movement, or F0 movement is necessary in order for the listener to be able to perceive vowels (Ryalls & Lieberman, 1982). The perceptual problem posed by high pitched voices is that the harmonics of the fundamental are widely spaced in frequency and thus do not specify the vocal tract resonances very well. Ryalls & Leiberman (1982) found that a changing F0 could be used to more accurately specify steady formant values for the listener, but it is obvious that the perceptual problem posed by widely spaced harmonics may also be solved by having the formant values change over time, regardless of the F0 trajectory. Third, Mrayati, Carre & Guerin (1988) discuss the acoustic implications of the incompressibility of the tongue. They suggest that incompressibility results in natural patterns of formant movement because increased constriction in one part of the vocal tract is necessarily accompanied by increased openness in another. So, just as the acoustic resonant properties of the vocal tract cause certain vowels to be more acoustically stable than others (Stevens, 1972, 1989), so also, the physiological properties of the vocal tract appear to cause some formant movement patterns to be more easily produced than others.

Static targets are modeled in a dynamic system as point attractors (Abraham & Shaw, 1989). Current spring/mass models of articulator dynamics (Browman & Goldstein, 1986) fall into this class. The data presented here suggest that vowel dynamics may be more complex than this. One way to model the types of movements which we found here is to string together across time a series of point attractors. Thus, for each vowel we would specify more than one target position. Such a model would definitely give a better fit to the data than would a single-target model, but how many targets is enough and what psychological status do we want to claim for the separate targets? The limit theorem states that as the increment decreases the function becomes better defined. So, as we increase the number of point attractors in our model of a vowel, the series of targets tends to define a function. We may think of the function defined by a series of point attractors as a periodic attractor. The dynamic system defined by a periodic attractor is simpler than the system defined by a series of point attractors because a single function specifies the dynamic properties rather than a system of independent functions. Thus, the system defined by a periodic attractor is more constrained than the system defined by a series of point attractors. Just as a task dynamic model accounts for complex interactions between adjacent gestures using simple

control parameters, so a system of periodic attractors may account for the complex movement patterns present in vowel articulation using simple control parameters.

5.4 *Summary*

This study has found that vowel production in northern midwestern English involves several interesting dynamic aspects. The vowels [ɪ, ɛ, ʌ] had shorter vowel durations, shorter vowel nucleus durations, longer closing deceleration durations, greater articulator stiffness, and smaller tongue movement magnitudes during the vowel than did the other vowels. The vowels [ɔ^{ə}, e^{ɪ}, o^{ʊ}, æ^{ə}] had the opposite pattern. This leaves just [i, ɑ, u] which had extreme values on the two primary vowel gestures. The results support the view suggested by several acoustic and perceptual studies that dynamic information is important in maintaining distinctions between vowels, because all of the factors of vowel production derived in the canonical discriminant analysis contained aspects of movement as well as general tongue shape, and two of the four factors were associated with movement patterns which divided the vowels into the same classes that resulted from the analysis of lip kinematics. This relationship is particularly interesting because it suggests that the stiffness of consonant releases and closures may be adjusted for the sake of completing a vowel-specific tongue gesture.

Acknowledgements

This paper is dedicated to Ilse Lehiste. Many thanks to Peter Ladefoged and Mona Lindau for sharing their data with me and for many thoughtful discussions throughout the project. Thanks also to Ken DeJong, Bruce Hayes, Pat Keating, Peter Ladefoged and Mona Lindau for their comments on an earlier version of this paper. This research was supported by grants DC00029 and DC00330 from the National Institutes of Health.

References

Abbs, J.H., Nadler, R.D. & Fujimura, O. (1988) X-ray microbeams track the shape of speech, *Soma*, Jan. 1988, 29-34.

Abraham, R.H. & Shaw, C.D. (1989) *Dynamics - The geometry of behavior. Part 1: Periodic behavior.* Santa Cruz, CA: Aerial Press.

Beckman, M.E., Edwards, J. & Fletcher, J. (1991) Prosodic structure and tempo in a sonority model of articulatory dynamics, In *Papers from lab phon II: Segment, gesture, and tone* (G. Docherty & D.R. Ladd, editors). Cambridge: Cambridge Univ. Press.

Bloomfield, L. (1933, 1984) *Language*. Chicago: Univ. of Chicago Press.

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

Chomsky, N. & Halle, M. (1968, 1991) *The sound pattern of English*. Cambridge: MIT Press.

Clements, G.N. (1990) Place of articulation in consonants and vowels: A unified theory. In *L'architecture et la géométrie des représentations phonologiques* (B. Laks & A. Rialland, editors). Paris: Editions du C.N.R.S.

Coker, C.H. (1976) A model of articulatory dynamics and control, *Proc. IEEE*, 64, 452-460.

Di Benedetto, M-G. (1989a) Vowel representation: Some observations on temporal and spectral properties of the first formant frequency, *Journal of the Acoustical Society of America,* 86, 55-66.

Di Benedetto, M-G. (1989b) Frequency and time variations of the first formant frequency: Properties relevant to the perception of vowel height, *Journal of the Acoustical Society of America,* 86, 67-77.

Engstrand, O. (1988) Articulatory correlates of stress and speaking rate in Swedish VCV utterances, *Journal of the Acoustical Society of America,* 83, 1863-1875.

Fowler, C.A., Rubin, P., Remez, R.E., & Turvey, M.T. (1980) Implications for speech production of a general theory of action. In *Language production.* (B. Butterworth, Editor), pp. 373-420. New York: Academic Press.

Fujimura, O., Kiritani, S. & Ishida, H. (1973) Computer controlled radiography for observation of movements of articulatory and other human organs, *Computer Biology Medicine,* 3, 371-384.

Gay, T. (1968) Effect of speaking rate on diphthong formant movements, *Journal of the Acoustical Society of America,* 44, 1570-1575.

Gay, T. (1974) A cinefluorographic study of vowel production, *Journal of Phonetics,* 2, 255-266.

Harshman, R., Ladefoged, P. & Goldstein, L. (1977) Factor analysis of tongue shapes, *Journal of the Acoustical Society of America,* 62, 693-707.

Hiki, S., & Itoh, H. (1986) Influence of palate shape on lingual articulation, *Speech Communication,* 5, 141-158.

Huang, C.B. (1986) The effect of formant trajectory and spectral shape on the tense/lax distinction in American vowels, *IEEE-ICASSP* 86, 893-896, Tokyo, Japan.

Jackson, M.T.T. (1988) Analysis of tongue positions: Language-specific and cross-linguistic models, *Journal of the Acoustical Society of America,* 84, 124-143.

Johnson, K., Ladefoged, P. & Lindau, M. (submitted) Individual differences and universal phonetics.

Kent, R.D. & Moll, K.L. (1972a) Cinefluorographic analyses of selected lingual consonants, *Journal of Speech and Hearing Research,* 15, 453-473.

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

Kent, R.D. & Netsell, R. (1971) Effects of stress contrasts on certain articulatory parameters, *Phonetica,* 24, 23-44.

Kiritani, S., Itoh, K. & Fujimura, O. (1975) Tongue-pellet tracking by a computer-controlled x-ray microbeam system, *Journal of the Acoustical Society of America,* 57 1516-1520.

Kshirsagar, A.M. (1972) *Multivariate analysis.* New York: Marcel Dekker.

Ladefoged, P., DeClerk, J., Lindau, M. & Papçun, G. (1972) An auditory-motor theory of speech production, *UCLA Working Papers in Phonetics,* 22, 48-75.

Ladefoged, P. & Halle, M. (1988) Some major features of the International Phonetic Alphabet, *Language,* 64, 577-582.

Lehiste, I. & Peterson, G.E. (1961) Transitions, glides, and diphthongs, *Journal of the Acoustical Society of America,* 33, 268-277.

Lindau, M. & Ladefoged, P. (1989) Methodological studies using an x-ray microbeam system, *UCLA Working Papers in Phonetics,* 72, 83-90.

Lindau, M. & Ladefoged, P. (1990) Interarticulatory relationships in vowel production, *UCLA Working Papers in Phonetics,* 74, 115-123.

Lindblom, B., Lubker, J. & Gay, T. (1979) Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation, *Journal of Phonetics,* 7, 147-161.

MacNeilage, P.F. & DeClerk, J.L. (1969) On the motor control of coarticulation in CVC monosyllables, *Journal of the Acoustical Society of America,* 45, 1217-1233.

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

Mermelstein, P. (1973) Articulatory model for the study of speech production, *Journal of the Acoustical Society of America,* 53, 1070-1082.

Mrayati, M, Carre, R., & Guerin, B. (1988) Distinctive regions and modes: a new theory of speech production, *Speech Communication,* 7, 257-286.

Nadler, R.D., Abbs, J.H., & 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.

Nearey, T.M. & Assman, P.F. (1986) Modeling the role of inherent spectral change in vowel identification, *Journal of the Acoustical Society of America,* 80, 1297-1308.

Nord, L. (1975) Vowel reduction: Centralization or contextual assimilation? In *Proceedings of the speech and communications seminar, 1974* (G. Fant, Editor). Stockholm: Almqvist & Wiksell.
Parker, E.M. & Diehl, R.L. (1984) Identifying vowels in CVC syllables: Effects of inserting silence and noise, *Perception & Psychophysics*, 36, 369-380.
Perkell, J.S. & Nelson, W.L. (1985) Variability in the production of the vowels /i/ and /a/, *Journal of the Acoustical Society of America*, 77, 123-133.
Peterson, G.E. & Barney, H. (1952) Control methods used in a study of the identification of vowels, *Journal of the Acoustical Society of America,* 24, 175-184.
Rubin, P., Baer, T., & Mermelstein, P. (1981) An articulatory synthesizer for perceptual research, *Journal of the Acoustical Society of America,* 70, 321-328.
Ryalls, J., & Lieberman, P. (1982) Fundamental frequency and vowel perception, *Journal of the Acoustical Society of America*, 72, 1631-1634.
Sagey, E.C. (1986) The representation of features and relations in non-linear phonology. Unpublished MIT Dissertation.
Saltzman, E.L. & Munhall, K.G. (1989) A dynamical approach to gestural patterning in speech production, *Ecological Psychology*, 1, 333-382.
SAS Institute (1982) *SAS User's Guide: Statistics, 1982 Edition.* SAS Institute, Inc., Cary, North Carolina.
Stevens, K.N. (1972) The quantal nature of speech: Evidence from articulatory-acoustic data, In *Human communication: A unified view* (E.EE. David and P.B. Denes, Editors), pp. 51-66. New York: McGraw-Hill.
Stevens, K.N. (1989) On the quantal nature of speech, *Journal of Phonetics*, 17, 3-45.
Stevens, K.N. & House, A.S. (1955) Development of a quantitative description of vowel articulation, *Journal of the Acoustical Society of America*, 27, 484-493.
Stevens, K.N., House, A.S., & Paul, A.P. (1966) Acoustical description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation, *Journal of the Acoustical Society of America,* 40, 123-132.
Strange, W. (1989) Evolving theories of vowel perception, *Journal of the Acoustical Society of America*, 85, 2081-2087.
Strange, W., Jenkins, J.J. & Johnson, T.L. (1983) Dynamic specification of coarticulated vowels, *Journal of the Acoustical Society of America*, 74, 697-705.
Trager, G.L. & Smith, H.L. (1951) *An Outline of English Structure.* Norman, Oklahoma: Battenburg press.
Verbrugge, R.R. & Rakerd, B. (1986) Evidence for talker-independent information for vowels, *Language & Speech*, 29, 39-57.
Wood, S. (1979) A radiographic analysis of constriction location for vowels, *Journal of Phonetics*, 7, 25-43.