Reprinted from

SPEECH COMMUNICATION

Speech Communication 14 (1994) 131–142

Phonetic analyses of word and segment variation using the TIMIT corpus of American English

Patricia A. Keating *, Dani Byrd, Edward Flemming, Yuichi Todaka

Phonetics Laboratory, Department of Linguistics, University of California, 405 Hilgard Avenue, Los Angeles, CA 90024–1543, USA

(Received 30 April 1993; revised 23 August 1993)

ELSEVIER

ELSEVIER

Speech Communication 14 (1994) 131–142

SPEECH COMMUNICATION

Phonetic analyses of word and segment variation using the TIMIT corpus of American English

Patricia A. Keating *, Dani Byrd, Edward Flemming, Yuichi Todaka

Phonetics Laboratory, Department of Linguistics, University of California, 405 Hilgard Avenue, Los Angeles, CA 90024–1543, USA

(Received 30 April 1993; revised 23 August 1993)

Abstract

This paper reports a set of studies of some phonetic characteristics of the American English represented in the TIMIT speech database. First we describe some relevant characteristics of TIMIT, and how we use the non-speech files on the TIMIT CD with a commercial database program. Two studies are then described: one using only the non-audio parts of TIMIT (segmental transcriptions and durations, and speaker information), and one using the audio signal for acoustic analysis. Results of such studies should be useful not only to linguistic phoneticians but also for speech recognition lexicons and text-to-speech systems.

Zusammenfassung

Dieser Artikel berichtet über eine Reihe von Untersuchungen zu einigen phonetischen Merkmalen des Amerikanischen Englisch, die in der TIMIT-Datenbank dargestellt werden. Zuerst werden die spezifischen Merkmale der TIMIT-Datenbank beschrieben und wie die nicht-audio Files von der TIMIT CD mit einem kommerziellen Programm für Datenbänke benützt werden. Dann werden zwei Untersuchungen beschrieben: eine benützt die nicht-audio Informationen der TIMIT-Datenbank (segmentale Transkriptionen, Lautdauer and Information über die Sprecher) und eine andere verwendet die Audiosignale für die akustische Analyse. Die Ergebnisse dieser Untersuchungen sollen nicht nur für Phonetiker brauchbar sein, sondern auch für Spracherkennungslexika und Vollsynthesesysteme.

Résumé

Cet article rend compte d'une série d'études sur quelques traits phonétiques de l'anglais américain représentés dans la base de données TIMIT. D'abord nous décrivons certaines caractéristiques pertinentes de TIMIT, et comment nous utilisons les fichiers d'étiquetage du CD TIMIT avec une base de données commerciale. Puis, nous décrivons deux études: l'une n'utilise que les informations non-audio de TIMIT (les transcriptions et durées segmentales et l'information sur les locuteurs), l'autre utilise le signal audio pour une analyse acoustique. Les résultats de ce type d'étude sont utiles non seulement pour la phonétique linguistique mais aussi pour l'élaboration de lexiques pour la reconnaissance de la parole et la synthèse à partir du texte des systèmes de traduction texte-parole.

* Corresponding author.

SSDI 0167-6393(93)E0081-6

Key words: American English; TIMIT; Speech databases; Pronounciation variation; Velar fronting

1. Introduction

1.1. The TIMIT corpus

The TIMIT speech database is a corpus of recorded readings of a large set of English sentences by a large number of native speakers of American English. Although the acquisition of acoustic phonetic knowledge was a design goal of the TIMIT project, very little such work has been reported so far. The phonetic transcriptions that are part of TIMIT have been used in statistical studies of allophonic variation of phonemes in (Cohen, 1989; Randolph, 1989; Riley and Ljolje, 1992). In the UCLA Phonetics Laboratory, we have been using TIMIT to evaluate a number of claims about pronunciation variation in the phonetic, phonological and TESL (Teaching English as a Second Language) literature. Though the TIMIT corpus has various limitations, some of which will be discussed below, it is a valuable tool for some kinds of phonetic analyses. Because TIMIT contains speech samples from many speakers, it is well-suited to uncovering acoustic phonetic patterns that hold generally across speakers, as well as aspects of the variation between speakers. TIMIT can tell us the different ways in which speakers pronounce something, and also which of these different pronunciations are most common.

The TIMIT speech database was developed at Texas Instruments and MIT, and is distributed by the U.S. National Institute of Standards and Technology (and now the Linguistic Data Consortium) on a CD-ROM. It consists of 2342 different sentences read by 630 native speakers of American English and is described in (Lamel et al., 1986; Pallett, 1990; Zue et al., 1990). Each speaker read ten sentences (about 30 sec of speech) for a total of 6300 utterances in the database. Three types of sentences are included. Two "calibration sentences", designed to allow cross-speaker comparisons, were read by all 630 speakers. An example of a calibration sentence is:

"She had your dark suit in greasy wash water all year." 450 "phonetically compact" sentences were designed to provide examples of phonemes in all possible left and right contexts. Each of these sentences was read by seven speakers. An example of a phonetically compact sentence is:

"Grandmother outgrew her upbringing in petticoats." The remaining 1890 sentences are "diverse sentences" selected mostly from the Brown corpus. Each of these was read by only one speaker. An example of a diverse sentence is:

"Turbulent tides rose as much as fifty feet."

The 6300 sentences of TIMIT comprise over five hours of speech. All of the sentences have been segmented and labeled. The transcriptions were done at MIT and were rechecked between the Prototype and final release of TIMIT. They are based on a combination of acoustic and auditory criteria (Seneff and Zue, 1988; Zue and Seneff, 1988) and record several kinds of subphonemic detail. The original transcriptions used a set of non-ASCII phonetic symbols; later, these were converted into Arpabet-style ASCII symbols. A symbol correspondence chart is distributed with the database. In this paper we use standard IPA symbols, following the 1989 Kiel conventions, along with the ASCII symbols as seen in TIMIT ("TIMITBET").

No single recorded corpus can meet every researcher's needs. Some factors about TIMIT that should be kept in mind in choosing how to use it are the following. First, the speech is read, not spontaneous. No sentences are very long, most are declarative, and some are quite odd. The speakers read them under formal recording conditions, and not all were fluent readers. That is, though this is connected speech, the speech style is far from casual. Second, the speaker sample is not as diverse or balanced as desirable. The 630 speakers of TIMIT are divided among eight "dialect regions" as shown in Fig. 1(a). They are mostly white and male, as shown in Fig. 1(b) and they are mostly in their 20s and 30s, as shown in Fig. 1(c). These demographics and the small

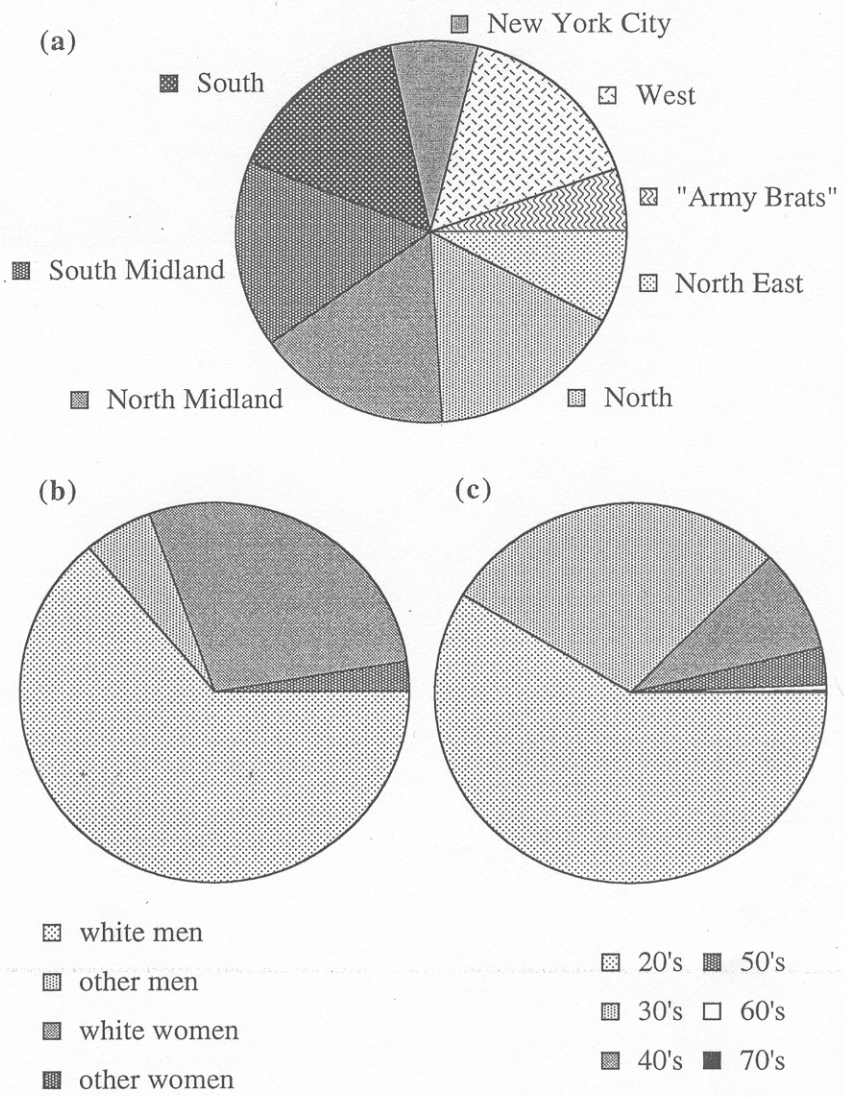

Fig. 1. Personal characteristics of the 630 TIMIT speakers. (a) Dialect region. (b) Sex and ethnicity. (c) Age.

amount of speech per speaker preclude thorough study of interspeaker variation. The coding of speakers for regional dialects, for example, suffices to demonstrate some diversity among speakers, but does not generally permit detailed dialect studies. Third, though the speech materials have been designed to contain all phoneme combinations, the number of tokens of any one combination may be very small. If specific sequences are desired, there may not be enough tokens to offset all the variation in other aspects of the tokens, such as prosodic environments, or to test detailed phonological claims. Fourth, no speaker-specific phonemic or prosodic transcriptions are included on the CD, limiting the kinds of phonologically-interesting questions that can be asked in an automated fashion without further transcription of the speech.

As should be clear from the foregoing discussion, a large speech database is a different kind of source of data from two more common kinds: single-purpose recordings made in the phonetics

laboratory and modern sociolinguistic fieldwork. TIMIT provides much less speech from each of many more speakers than either of these. Rather than telling us something in detail about any one speaker or group of speakers, TIMIT gives us a broad, overall picture of native American English pronunciation, at least for read speech.

1.2. UCLA database of non-speech parts of TIMIT

The TIMIT database, consisting of speech, transcriptions, some information about the speakers and a dictionary, is distributed on a CD-ROM. The CD contains no database tools or structure beyond subdirectories and mnemonic names for files. To work with the information that accompanies the speech recordings, we use a commercial relational database, Borland's Reflex Plus, on a Macintosh Quadra computer. All the non-speech files on the TIMIT CD, except for the phonemic dictionary, were imported into four database files: sentences, phones, words, speakers. The organization of the database – the fields used in each database file and the links between the files – is shown in Fig. 2.

Each record in the "sentences" database contains the orthographic form of a sentence together with its TIMIT ID code and is linked to the speakers who spoke it. Each record in the "phones" database contains one ASCII phonetic symbol (a phone), its start and end times (in msec) in the speech signal, and a coding (provided by us) of its position within its word. Each phone record is linked to the sentence containing it and the speaker who read it. "Next_phone" provides a link to the record of the following phone. Each record in the "words" database contains a word in an utterance and is linked to the speaker and to the phones it contains. Each record in the "speakers" database contains information about one speaker, such as their TIMIT ID code, sex, dialect region and age, and is linked to the sentences uttered by that speaker.

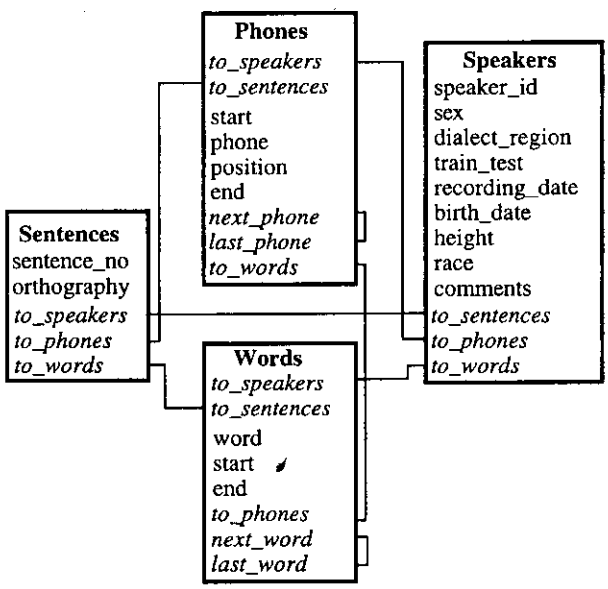

Fig. 2. Organization of the non-speech parts of TIMIT in a Reflex database.

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

In this paper we report on two studies using TIMIT. One is a study of pronunciation variation among speakers which uses the segmental transcriptions and speaker information in the database. The other is a study of general patterns across the speaker group which uses the audio signals from the CD-ROM. Together these two studies demonstrate how TIMIT can be exploited for explicit phonetic knowledge.

2. A transcription study

Because the phonetic transcriptions in TIMIT are fairly narrow and are largely acoustically-defined, many research questions can be answered using the Reflex database and these transcriptions alone, without an independent acoustic

analysis of the speech files. (Use of the audio portion of TIMIT will be discussed in Section 3.) For example, using the segmental phonetic transcriptions we can look at actual pronunciations of words or phonemes over the corpus. Particularly for words occurring with high frequency in the corpus, TIMIT is suited for tracking variation in (reading) pronunciation across the American English population.

In undertaking such studies of the TIMIT transcriptions, we are claiming that these transcriptions are meaningful and reliable. Therefore some comments on the principle and practice of transcription are in order. The TIMIT labels represent a necessarily-arbitrary segmentation and categorization of utterances into phonetic segments. Obviously, any study of these labels will be acceptable only to researchers who accept segmental phonetic transcriptions as a useful record of speech-events. It is possible to admit that segmental transcriptions have no theoretical basis as formal representations of speech, and yet still find them useful: they are a shorthand for, a pointer to, key articulatory and acoustic events. This is our perspective. On this basis, then, we can consider the TIMIT labels themselves. The hand-labeling of TIMIT was carefully done, and no other existing corpus of transcribed speech can be expected to be as reliable or useful. The labels have been implicitly accepted and found useful by all those investigators who have used TIMIT to train statistical models of phones for automatic speech recognition.

The TIMIT labelers used context-sensitive criteria, both auditory and acoustic, in segmentation and labeling. Since many decisions in phonetic categorization are difficult and arbitrary, one looks for reasonable conventions and consistent application of them by labelers. The labels of TIMIT cannot be taken at face value without an understanding of these conventions. In general, we find the TIMIT labeling conventions sensible, and the resulting transcriptions reliable; the difficulties we noticed are not relevant to the studies reported here. Because the vowel transcriptions are used in our first study below, it is worth commenting on them in particular. TIMIT distinguishes between full and reduced vowel qualities, with reduced [ə] ("ax"), voiceless [ə̥] ("ax-h"), and [i] ("ix") used only for very short vowels of unclear quality. A very short vowel may be labeled with one of the full-vowel-quality symbols if that quality is clear. The two voiced reduced vowels are in turn distinguished according to the relative value of $F2$. Distinctions between the full vowel qualities such as [i] ("iy") and [ɪ] ("ih") are made on auditory grounds, but with a strong preference for the expected, phonemic transcription.

Similarly, the usefulness of the durations associated with segments in the database depends on the usefulness of the criteria used for segmentation and alignment, and the consistency of their use, by the labelers. This is true for any study of acoustic segment durations. In TIMIT as in traditional phonetic studies, the boundaries of stops, fricatives and nasals, and of vowels when between these consonants, are relatively reliable and meaningful, but boundaries of other segments can be more arbitrary. Two conventions of segmentation in TIMIT which affect vowel durations should be mentioned here. One concerns sequences of vowel plus phonemic nasal consonant. If in the signal there is no nasal consonant segment, but instead only a nasalized vowel, then the last one or two pitch periods of the vowel are arbitrarily segmented as the nasal consonant. Therefore some vowels before phonemic nasal consonants are measured as shorter in TIMIT than they would be in other studies. The other convention concerns laryngealization associated with a word-initial vowel. Such laryngealization was segmented separately as glottal stop [ʔ] ("q") and not as part of the vowel or vowels around it. Therefore any vowel before or after "q" is most likely measured as shorter in TIMIT than it would be in other studies.

With understanding of the transcription system of TIMIT, topics for phonetic study can be selected which exploit the strengths of the system and avoid its difficult areas.

Consider the pronunciation of the word "the", which, with 2202 tokens, is the most common word in TIMIT (Lamel et al., 1986). Ninety percent of the tokens of "the" in TIMIT consist of the consonant [ð] (TIMITBET "dh") followed by a vowel, and another 6.7% of the tokens consist

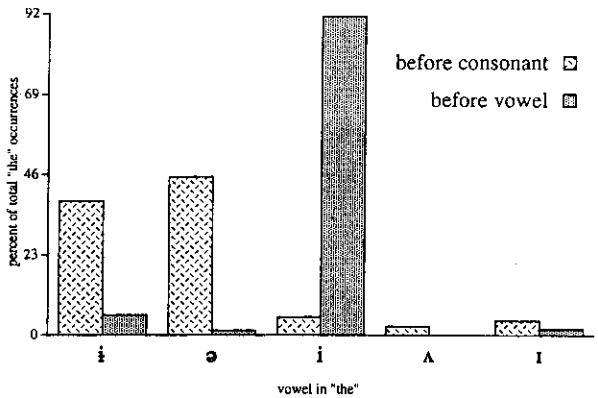

Fig. 3. Most frequent vowels in "the" before words beginning with vowels and with consonants. Percent before vowels and percent before consonants each add to 100.

of only a single vowel. (The remaining 69 tokens provide 32 assorted other pronunciations.) The vowels found in "the" cover a wide array of vowel qualities, including several which appear in only a very few tokens (for example, [æ] ("ae")). Fig. 3 shows the 5 most frequent vowel qualities in "the" in TIMIT; the reduced vowels [ə] ("ax") and [i] ("ix") are most common overall, followed by [i] ("iy"). The vowels [ə] and [i] are the most common in "the" because many more tokens of "the" occur before consonants than before vowels in TIMIT. As Fig. 3 shows, the choice of vowel in "the" is also highly dependent on the first segment in the following word. Before consonants, the reduced vowels [ə] ("ax") and [i] ("ix") dominate, while before vowels, [i] ("iy") is by far the most common; this difference is statistically significant ($\chi^2 = 1173.336$, $p = 0.0001$).

We also examined the durations of vowels in "the" by tabulating the durations of the final phone in all of the tokens of "the" in the database, using the start and end times associated with each labeled segment, and submitting these to ANOVA. Individual post hoc comparisons show that the reduced vowels [i] and [ə] are reliably shorter than the non-reduced vowels [i], [ʌ] and [ɪ]. Thus, in general, in "the" shorter vowels occur before consonants and longer vowels occur before vowels.

In addition to these durational differences, the occurrence of [i] before vowels most likely reflects true vowel quality differences. A small subset of tokens of "the" with [i] and with [i] before both consonants and vowels was selected, and the first two formant frequencies were measured (see next section for details of method). For this subset, $F1$ did not differ between [i] versus [i] or before consonant versus vowel. As would be expected, $F2$ was higher for [i] versus [i] ($p = 0.0002$). $F2$ was also higher before vowels than before consonants ($p < 0.04$), even within each vowel, so that [i] before vowels has a higher $F2$ than [i] before consonants. Thus the vowels in "the" before consonants are more centralized in quality. If this pattern holds of the whole corpus, then it can be said that vowels in "the" before consonants are reduced in both duration and quality.

Standard and normative descriptions of the pronunciation of "the" follow the pattern seen in Fig. 3: they suggest that it is pronounced [ði] or [ðɪ] before a word beginning with a vowel, possibly with an intervening palatal glide, or with an intervening glottal stop before stressed vowels; and as [ðə] before a word beginning with a consonant. This difference is painstakingly taught to non-native learners of English as appropriate for both read and spontaneous speech. However, we have observed that among UCLA undergraduates taking introductory phonetics courses, the norm seems to be [ðə] before a consonant and [ðəʔ] before a vowel. (See also (Henton and Bladon, 1987) for the same observation about a similar population.) How can [i] be by far the most common vowel in "the" before vowels in TIMIT, yet seem rare among UCLA undergraduates? A striking finding is that in TIMIT the choice of vowel in "the" before a phonemic vowel is age-dependent. Fig. 4 shows the use of [i] versus all other vowels, by age. *No one* over 50 years old uses any vowel but [i] in "the" before a vowel, and while [i] remains the most common vowel even for younger speakers, other vowels occur in more than a third of the tokens for the youngest speakers. This difference is highly significant ($\chi^2 = 13.365$, $p < 0.01$). Since some TIMIT speakers were recorded several years ago, the UCLA undergraduates would probably be in the next age bin down and have advanced further along in what appears to be an ongoing change in a pro-

nunciation norm. It is also possible that the change is more common in the spontaneous speech of the students than in their read speech, which would account for some of the difference between the students and TIMIT, but it would not account for the age differences seen within TIMIT itself.

Another important variable in the pronunciation of "the" is the presence or absence of a glottal stop [ʔ] or laryngealization (both TIMIT-BET "q") at the end of the word. Glottal stop or laryngealization is expected to occur after "the" when the following word begins with a vowel. In TIMIT, there are only 336 tokens of "the" before a vowel-initial word. Glottal stop or laryngealization is not especially common among them; only 27% of a random sample of 242 tokens of "the" before a phonemic vowel have a glottal stop or laryngealization between the two words. (We have checked a subset of these and confirmed that the labeling is reliable on this point.) Almost all of these occur when the vowel after "the" has primary stress. The quality of the following vowel does not seem to matter.

Because [i] is common in "the" before vowels, "q", when it occurs, frequently follows [i]. However, it is less common there than would be expected. Fig. 5 shows that across the whole

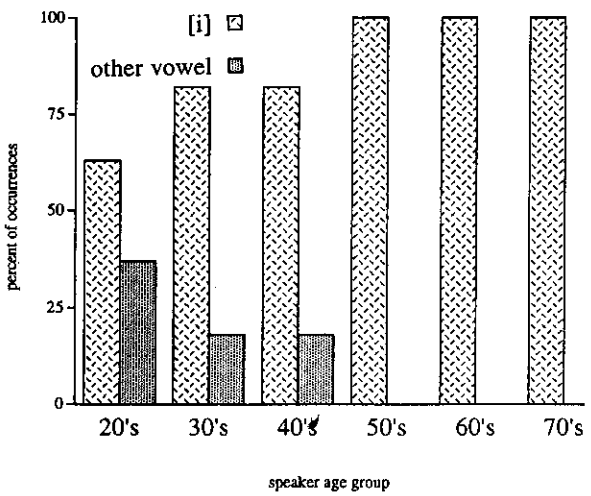

Fig. 4. Use of [i] in "the" before words beginning with vowels according to age of speaker. Percents in each age group add to 100.

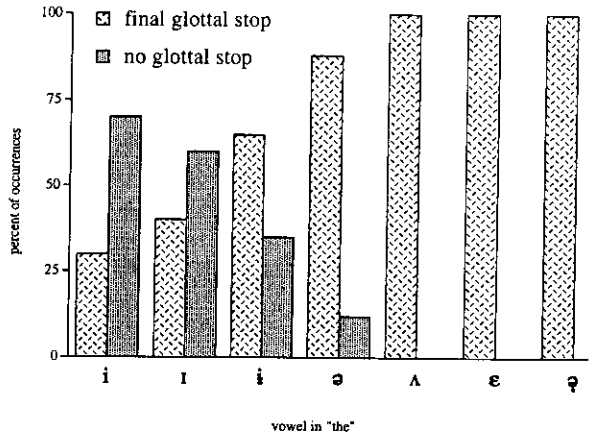

Fig. 5. Glottal stop versus no glottal stop between "the" and words beginning with vowels, according to the vowel found in "the". Percents in each vowel group add to 100.

sample the occurrence of "q" between "the" and a phonemic vowel is proportionately lowest after [i]. Tokens of prevocalic "the" with [ə] are usually followed by "q" and those with the other non-high vowels always are, whereas a smaller proportion of tokens of prevocalic "the" with [i] are. Glottal stop follows [i] and [ə] proportionately more frequently than it follows [i] and [ɪ] ($\chi^2 = 20.6$, $p < 0.0005$).

A random subset of 30 tokens was drawn from the 336 tokens of "the" before vowel-initial words for formant frequency analysis of the vowels in "the". The labels for the vowels in "the" correspond well to formant differences, with the two full vowels [i] and [ɪ] acoustically distinct. The only vowel whose formants are not distinct from those of other vowels is reduced [i] ("ix"), which covers the same values as both [i] ("iy") and [ɪ] ("ih"). As noted earlier, [i] is used largely for very short vowels. Within the sample, the vowels transcribed as being followed by [ʔ] or laryngealization ("q") are those having a low $F2–F1$, and those labeled [i] ("ix"). That is, "q" is not observed when the vowel in "the" has formant values appropriate for full vowels "iy" and "ih". This analysis suggests that the above observations based on the larger sample of TIMIT labels are reliable.

As mentioned above, since all phones in TIMIT are listed with their start and end times (to allow

alignment with the waveforms), the "phones" database can also be used to study segment durations. For example, the durations of the vowels in all tokens of "the" can be compared across different speaker variables. Speakers' dialect region shows a slight tendency to affect these durations: a Scheffe post hoc test indicates a trend for the Southern speakers to have longer durations than the New England speakers ($p = 0.06$). This result is consistent with the finding that overall, the Southern speakers in TIMIT show the slowest speaking rate (Byrd, 1992).

The study of "the" before words beginning with a phonemic vowel shows that several vowel qualities occur in "the", with [i] the most frequent, but that younger speakers show an emerging trend away from [i]. The implication of these results for English language teaching and for text-to-speech synthesis with "the" is that there is probably no point in forcing the distinction between [ði] versus [ðə] unless an older style of speech is being modeled. The implication for speech recognition, however, is that both variants still need to be listed in a recognition lexicon if a system is to be prepared for a variety of users. Both of these conclusions concern segmental representations, and as such they abstract away from details of acoustic structure. TIMIT, with its segmentation and labeling, makes such studies easy to do.

3. An acoustic study

Acoustic analysis can also be performed on speech tokens identified by database searches. Individual audio signal files are read from the CD-ROM and analyzed in the Kay Elemetrics Computer Speech Lab (CSL) environment for PCs. CSL reads in the TIMIT ASCII phonetic transcription, converts it to standard IPA symbols, and displays the IPA version time-aligned with the waveform (though without the start and end times of each phone).

We have used the TIMIT corpus to study the effect of vowel context on the acoustics of velar stop consonants. It is well-known that in English (and many other languages), velar stops are fronted in their place of articulation on the palate when they are followed by a front vowel (Ladefoged, 1982); see also references and discussion in (Keating and Lahiri, 1993). This fronting is reflected acoustically in the frequency of the strongest spectral peak in the release burst, and in formant transitions (Zue, 1976). A study was designed to determine whether this fronting is as strong after front vowels as before. In this study, in contrast to the previous one, we do not rely on the transcribed vowel qualities in comparing the effects of vowels on velars, but instead use measured values for formant frequencies. We do however rely on the labeling of velar stop closures and releases for the consonants; these were verified auditorily by the experimenter.

3.1. Method

Speech materials

All velar stops coded as having releases and with a pause or schwa on one side and a vowel on the other were located in the database. With a pause or schwa on one side, the vowel on the other side of the velar should be the primary determinant of the velar's place of articulation. (Schwas are very short and have a contextually-determined quality.) All of the tokens including a pause, and some of those with a schwa, were used for the analysis. Each token generally came from a different speaker, though this was not necessarily the case.

The TIMIT corpus includes very few tokens of some velar/vowel combinations. It was therefore necessary to collapse together certain vowel categories into larger groups or to eliminate tokens with certain rare vowels from the dataset. This collapsing was done on the basis of measured values for $F2$, rather than the TIMIT labels.

Procedure

Each token was loaded into CSL (at the TIMIT sampling rate of 16 kHz), and a 100 point FFT spectrogram was displayed with the waveform. From the display, three timepoints were located in a token: the release burst, the midpoint of the adjacent vowel, and either (for prevocalic stops) the onset of formant transitions after the burst

(whether voiced or aspirated) or (for postvocalic stops) the offset of formant transitions before the closure. At each of these timepoints, a 20 ms Blackman window was centered and an autocorrelation LPC spectrum (14 coefficients, 512 points) was computed. The LPC spectrum, along with an FFT spectrum if necessary, was displayed, and spectral peak frequencies were measured. For the burst, the measured value was the frequency of the highest-amplitude peak below 4 kHz (see also (Sereno and Lieberman, 1987)); at vowel midpoint and transition edge, the measured values were the frequencies of $F2$, $F3$ and $F4$. For any one token, then, seven measurements were made: a single burst frequency, $F2/F3/F4$ at vowel midpoint, and $F2/F3/F4$ at transition edge. Subsequent Analysis of Variance designs were factorial non-repeated-measures.

Tokens were divided into vowel groups post hoc, on the basis of measured formant frequencies of vowels. Vowels were categorized into four groups according to their steady-state $F2$ values, called high $F2$, mid-high $F2$, mid-low $F2$ and low $F2$. These corresponded to (most instances of) labeled [i], other front vowels, back or central unrounded vowels and back rounded vowels. For CV tokens this categorization yields sufficient numbers of tokens in each group for statistical comparison. For VC tokens, even this collapsing is not enough, and only the two middle groups (mid-high $F2$ and mid-low $F2$) can be compared.

3.2. Results

First, a starting assumption of the study, that the following vowel affects velar release acoustics, was verified for 244 velar + vowel (CV) tokens from TIMIT. ANOVA with factors vowel-group (4 levels), speaker-sex (2 levels) and consonant-voicing (2 levels) showed that burst frequency does vary significantly with the $F2$ of the following vowel ($F(3,228) = 61.394$, $p = 0.0001$), and with speaker sex ($F(1,228) = 16.052$, $p < 0.0001$), but not with consonant voicing ($F(1,228) = 1.512$, $p > 0.22$) or with any interaction of these factors. The significant effect of following vowel, shown in Fig. 6, indicates very strong coarticulatory effects in CV sequences: the frequency of the burst peak is higher when the $F2$ of the following vowel is higher. Post hoc comparisons indicate that almost every vowel-group differs significantly from every other, the one exception being the two groups of front vowels. Lack of interaction effects involving speaker-sex shows that data from men and women pattern alike with respect to the other factors, though absolute values differ. Be-

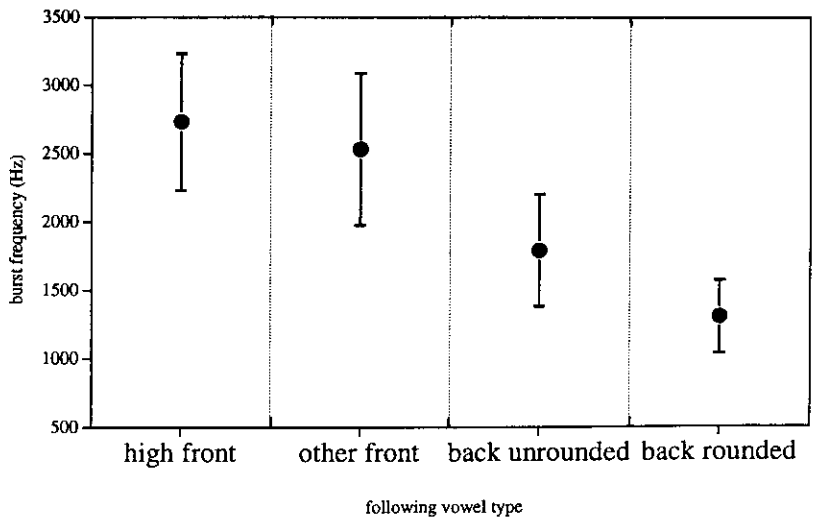

Fig. 6. Mean values for burst frequencies of velars before 4 vowel groups.

cause consonant voicing had no effects on burst frequency, /k/ and /g/ are combined in the following analyses.

With this assumption of coarticulation verified, analysis then addressed the comparison of burst frequencies for velars in CV versus VC sequences. Because there are few VC sequences followed by pause or schwa, statistical comparison of CV versus VC was possible only for the mid-high $F2$ and mid-low $F2$ vowel contexts. Recall that this means a comparison between a group generally consisting of front vowels other than [i] and a group generally consisting of unrounded back and central vowels. 56 tokens of velars before mid-high-$F2$ vowels and 54 tokens of velars before mid-low-$F2$ vowels, a total of 110 CV tokens, were compared with 50 tokens of velars after mid-high-$F2$ vowels and 28 tokens of velars after mid-low-$F2$ vowels, a total of 78 VC tokens. ANOVA again used the factors speaker-sex and vowel-group, along with segment-order (CV versus VC). Again both the speaker-sex and vowel-group effects were highly significant, with the frequency of the burst being higher when the vowel has a higher $F2$. The effect of segment-order was significant overall, but its direction depends on the vowel group. Fig. 7 shows the interaction of vowel-group and segment-order. The key comparisons are the vowel-groups within each segment-order. A higher burst frequency adjacent to a mid-high-$F2$ vowel means that the consonants are coarticulated with the vowels.

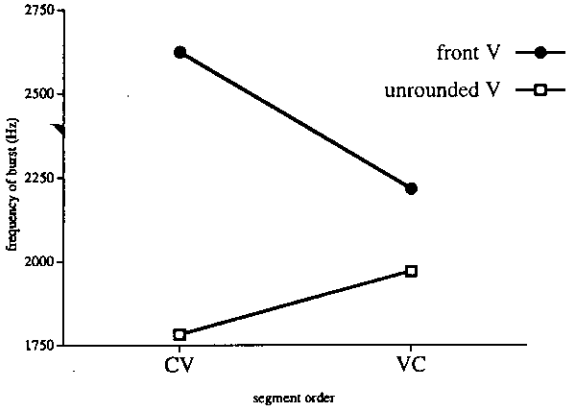

Fig. 7. Mean values for burst frequencies of velars before and after 2 vowel groups.

Both CVs and VCs show such coarticulation. However, the vowel effect – viewed as the difference between the values in the two vowel contexts – is much weaker in VCs. Prevocalic velars show a large effect of the following vowel, a difference of over 800 Hz between the two vowel groups, while postvocalic velars show a much smaller effect of the preceding vowel, a difference of under 300 Hz. Thus we can conclude that as measured in the release burst, velars are indeed more fronted (acoustically speaking, at least) before than after the fronter vowels.

Various regression analyses were also performed to clarify the relation between bursts, transitions and vowel midpoints. All speakers and vowels are combined for these analyses: 244 CV tokens and 126 VC tokens. Burst frequency was regressed against each of the three formant values ($F2/F3/F4$) at vowel midpoint, somewhat in the style of Sussman et al. (1991). In both CV and VC positions, the burst value was better predicted by the lower formants ($F2$ better than $F3$ better than $F4$, as expected on the basis of the results in (Keating and Lahiri, 1993)), and for each formant, the fit was better for CV than VC samples, though it was significant in every case. Fig. 8 shows the relation of burst to $F2$ midpoint in CV versus VC samples. This analysis, like the ANOVA, shows that the burst value is more dependent (higher r^2 value) on the vowel quality in CVs than in VCs, but that in both cases the relation is strong ($p = 0.0001$).

These analyses, then, show that the release of a velar is more influenced by the adjacent vowel in CV than in VC tokens. If, however, we look at formant transitions rather than bursts, this difference disappears or is even reversed. For CV tokens, the value of the onset of each formant transition (whether voiced or aspirated) was regressed against the value of that formant at vowel midpoint, and for VC tokens, the value of the offset of each formant transition was regressed against the value of that formant at vowel midpoint. (This analysis is much like that in (Sussman et al., 1991), the difference being that here formant edges are not determined by voicing onset/offset.) Fig. 9 shows the relation of $F2$ edge (onset or offset) to $F2$ midpoint in CV

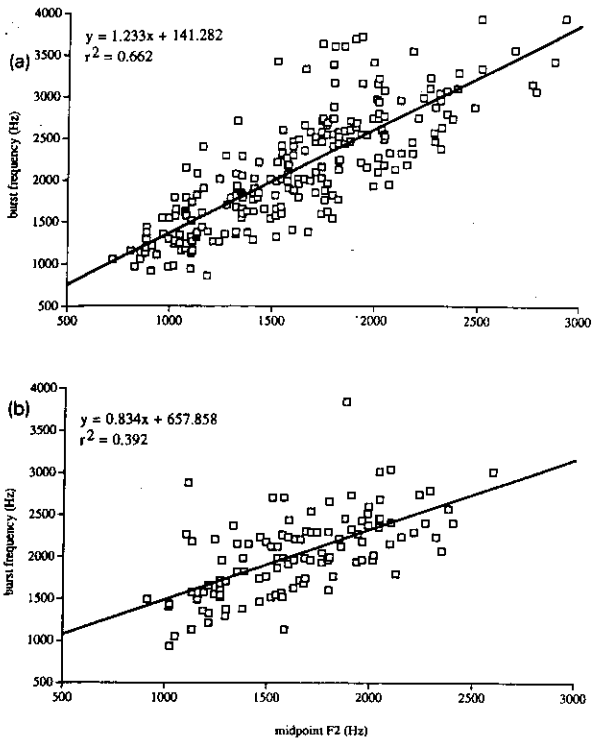

Fig. 8. Regressions of burst against $F2$ midpoint. (a) CV; (b) VC.

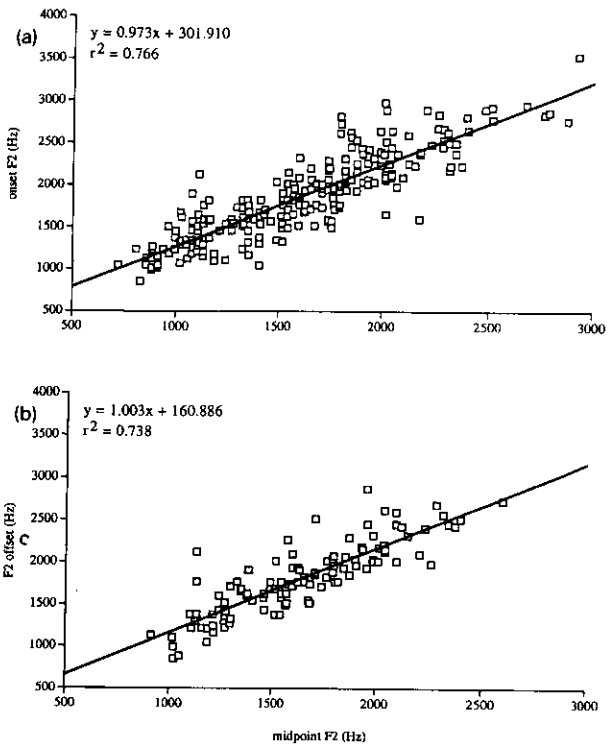

Fig. 9. Regressions of $F2$ edge against midpoint. (a) CV; (b) VC.

versus VC samples. As would be expected, these fits of formant edges to midpoints were better than those of bursts to midpoints presented above. Also, in both CV and VC orders, again the fits were better the lower the formant. What is interesting is that for these formant transition data the VC fits are about as good as the CV fits. Thus we can conclude that in terms of formant transitions alone, velars are about equally affected by vowels on either side, while in terms of release burst, velars are more affected by a following vowel.

This result may seem contradictory at first, but in fact makes sense. Transition onset and offset are equi-distant from the vowel midpoint, so show equal dependence on the midpoint. In contrast, the bursts in CV and VC differ in distance from the vowel midpoint: the burst in VC is much further from the midpoint of V than is the burst in CV. Burst measures are thus inherently asymmetric for CV versus VC. In the VC case, the burst is further away from V, and we would expect it to be relatively free of influence of V.

What this suggests is that for a listener the perceived effect of a preceding vowel on a velar will depend on whether the listener attends more to the formant transition into the closure or to the release. If the listener attends to the transition, the velar will sound affected by the vowel; if the listener attends to the release, it will not. This interpretation accords with that in a recent small laboratory study (Nolan, 1993), in which velars were paired with preceding versus following vowels [ɪ] (front) and [ʌ] (back). Using electropalatography, Nolan recorded the locations of tongue-to-palate contact for the velars in the different contexts. Nolan's intuition prior to the experiment was that his velars are more fronted before front vowels than after them. However, the electropalatography data showed the reverse to be true: the preceding vowel had a greater effect on contact location. Nolan attributed his "intuition" to the burst properties of the consonants. That is, the following vowel, by affecting the release burst, sounded as if it had a greater importance.

4. Conclusion

The TIMIT CD database can be a useful tool for linguistic phoneticians interested in describing the phonetic characteristics of American English. The studies reported in this paper concern two different kinds of phonetic questions. The first study, on pronunciation variation for the word "the", looked at the range of variants across the corpus and asked what factors enter into determining this variation. It was found that phonetic context and speaker age are two such important factors. The second study, on acoustic variation in velar stop consonants as a function of an adjacent vowel, tested whether preceding and following vowels are equally important in determining consonant acoustics. It was found that the effects on adjacent formant transitions are about equal, but that following vowels have a much stronger effect on stop release bursts.

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

5. Acknowledgments

Thanks to Michael Shalev for data analysis; to the Committee on Research of the UCLA Academic Senate for financial support to P. Keating and P. Ladefoged for work with TIMIT, and to John Choi and Barbara Blankenship for active participation in the seminar where this work was begun. Parts of this paper appear also in (Keating et al., 1992a,b).

6. References

D. Byrd (1992), "Sex, dialects, and reduction", *ICSLP 92 Proc.*, ed. by J. Ohala et al. (University of Alberta), Vol. 1, pp. 827–830.

M. Cohen (1989), Phonological structures for speech recognition, U.C. Berkeley Ph.D. dissertation.

C. Henton and A. Bladon (1987), "Developing computerized transcription exercises for American English", *J. IPA*, Vol. 17, pp. 72–82.

P. Keating and A. Lahiri (1993), "Fronted velars, palatalized velars, and palatals", *Phonetica*, Vol. 50, pp. 73–101.

P. Keating, B. Blankenship, D. Byrd, E. Flemming and Y. Todaka (1992a), "Phonetic analyses of the TIMIT corpus of American English", *ICSLP 92 Proc.*, ed. by J. Ohala et al. (University of Alberta), Vol. 1, pp. 823–826.

P. Keating, B. Blankenship, D. Byrd, E. Flemming and Y. Todaka (1992b), "Phonetic analyses of the TIMIT corpus of American English at UCLA", *UCLA Working Papers in Phonetics*, No. 81, pp. 1–16.

P. Ladefoged (1982), *A Course in Phonetics* (Harcourt Brace Jovanovich, New York), 2nd Edition, p. 58.

L. Lamel, R. Kassel and S. Seneff (1986), "Speech database development: Design and analysis of the acoustic-phonetic corpus", *Proc. DARPA Speech Recognition Workshop*, February 1986, pp. 100–109.

F. Nolan (1993), "Phonetic correlates of syllable affiliation", in *Phonological structure and phonetic form: Papers in Laboratory Phonology III*, ed. by P. Keating (Cambridge Univ. Press, Cambridge), in press.

D. Pallett (1990), "Speech corpora and performance assessment in the DARPA SLS program", *ICSLP 90 Proc.*, pp. 24.3.1–24.3.4.

M. Randolph (1989), Syllable-based constraints on properties of English sounds, MIT Ph.D. dissertation.

M. Riley and A. Ljolje (1992), "Recognizing phonemes vs. recognizing phones", *ICSLP 92 Proc.*, pp. 285–288.

S. Seneff and V. Zue (1988), "Transcription and alignment of the TIMIT database", unpublished manuscript distributed with the NIST TIMIT CD-ROM database.

J. Sereno and P. Lieberman (1987), "Developmental aspects of lingual coarticulation", *J. Phonetics*, Vol. 15, pp. 247–257.

H. Sussman, H. McCaffrey and S. Matthews (1991), "An investigation of locus equations as a source of relational invariance for stop place of articulation", *J. Acoust. Soc. Amer.*, Vol. 90, pp. 1309–1325.

V. Zue (1976), Acoustic characteristics of stop consonants: A controlled study, MIT Dissertation, Distributed by Indiana University Linguistics Club 1980.

V. Zue and S. Seneff (1988), "Transcription and alignment of the TIMIT database", *Proc. Second Meeting on Advanced Man–Machine Interface through Spoken Language*, pp. 11.1–11.10.

V. Zue, S. Seneff and J. Glass (1990), "Speech database development at MIT: TIMIT and beyond", *Speech Communication*, Vol. 9, No. 4, pp. 351–356.