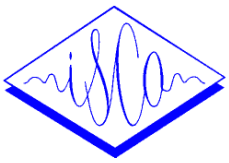

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

PHONETIC ANALYSES OF THE TIMIT CORPUS OF AMERICAN ENGLISH

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

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

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 generally how we use the non-speech files on the TIMIT CD with a commercial database program. Some studies of pronunciation variation using only the segmental transcriptions and durations of TIMIT are then described. Results of such studies should be useful not only for linguistic phonetics but also for speech recognition lexicons and text-to-speech systems.

I. INTRODUCTION

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

The TIMIT speech database, developed at Texas Instruments and MIT and distributed by the National Institute of Standards and Technology on a CD, consists of 2342 sentences read by 630 native speakers of American English. It is described in [8] and [6]. Three types of sentences are included. Two "calibration sentences", designed to allow dialect comparison, were read by all 630 speakers. The speakers are coded as belonging to one of 8 "dialect regions" (New England, New York City, North Midland, South Midland, Southern, Northern, Western, and "Army brat"). 450 "phonetically compact" sentences were designed to provide examples of phonemes in all possible left and right contexts. Each of these sentences was read by seven speakers. The remaining 1890 sentences are "diverse sentences" selected mostly from the Brown corpus. Each of these was read by only one speaker. Each speaker read ten sentences altogether (about 30 sec of speech) and there is a total of 6300 utterances in the database. (There was also a partial, Prototype version released earlier, and one of the studies in this paper used that version.) All sentences in TIMIT are segmented and labeled. The transcriptions are based on a combination of acoustic and auditory criteria [7].

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

sentences
phones
words
speakers.

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

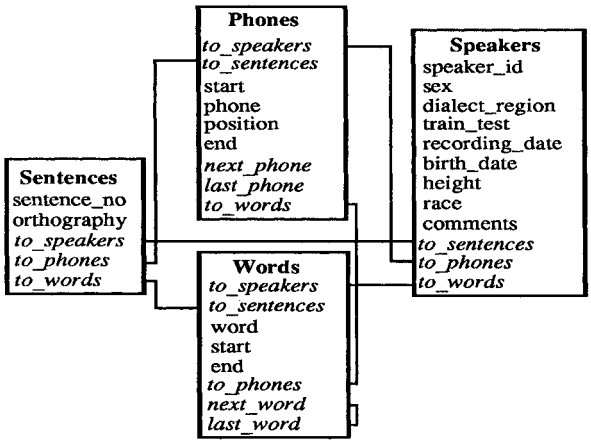

Figure 1 - Organization of TIMIT in a Reflex database.

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

10.21437/ICSLP.1992-270

searches can be exported into commercial spreadsheet, graphing, and statistical programs for analysis.

II. TRANSCRIPTION STUDIES

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

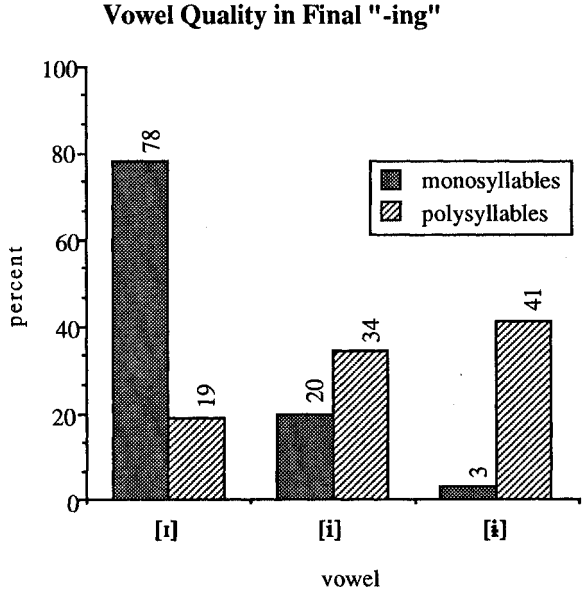

Figure 2 - Vowels in word-final "ing" in monosyllables and polysyllables.

Consider the pronunciation of the word "the", the most common word in TIMIT [5]. It occurs with a wide array of vowel qualities, including several which appear in only a very few tokens (for example, [æ]). Figure 3 shows the five most frequent vowel qualities in "the" in TIMIT; [ə] and [ɨ] are most common overall, followed by [i]. This variation is significantly influenced by the speaker's sex: women use [ɨ] and [ɪ] more than men, and [ə] and [i] less. As Figure 3 shows, the choice of vowel in "the" is also highly dependent on the first segment in the following word. Before consonants, [ə] and [ɨ] dominate, while before vowels [i] is by far the most common; this difference is significant.

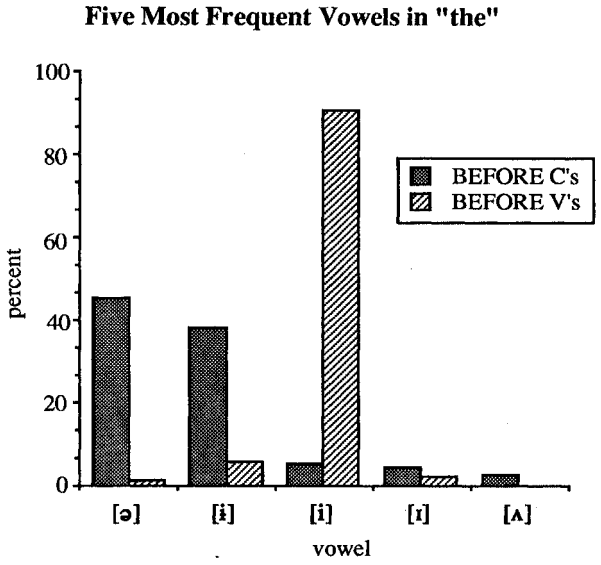

Figure 3 - Most frequent vowels in "the" before words beginning with vowels and with consonants.

Standard and normative descriptions of the pronunciation of "the" follow this pattern: they suggest that it is pronounced [ðiˈ] or [ðɪˈ] before a word beginning with a vowel, possibly with an intervening palatal glide, or with an intervening glottal stop before stressed vowels; and as [ðə] before a word beginning with a consonant. This difference is painstakingly taught to ESL students. However, among UCLA undergraduates the norm seems to be [ðə] before a consonant and [ðəʔ] before a vowel (see also [3]). How can [i] be by far the most common vowel in "the" before vowels in TIMIT, yet be rare among UCLA undergraduates? A striking finding is that in TIMIT the choice of vowel in "the" before a phonemic vowel is age-dependent. Figure 4 shows the use of [i] vs. all other vowels, by speaker age. No one over 50 years old uses any vowel but [i] in "the" before a vowel, and while [i] remains the most common vowel even for younger speakers, other vowels occur in more than a third of the tokens for the youngest speakers. This difference is highly significant. Since some TIMIT speakers were recorded several years ago, the UCLA undergraduates are probably the next age bin down and have advanced further along in what appears to be a current change in a pronunciation norm.

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

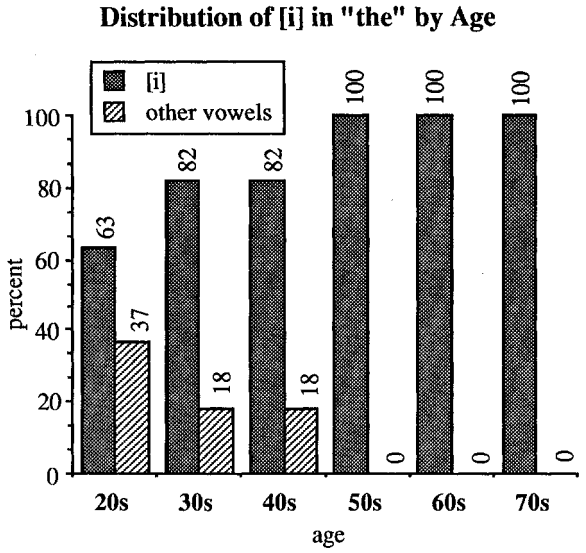

Figure 4 - Use of [i] in "the" before words beginning with vowels according to age of speaker.

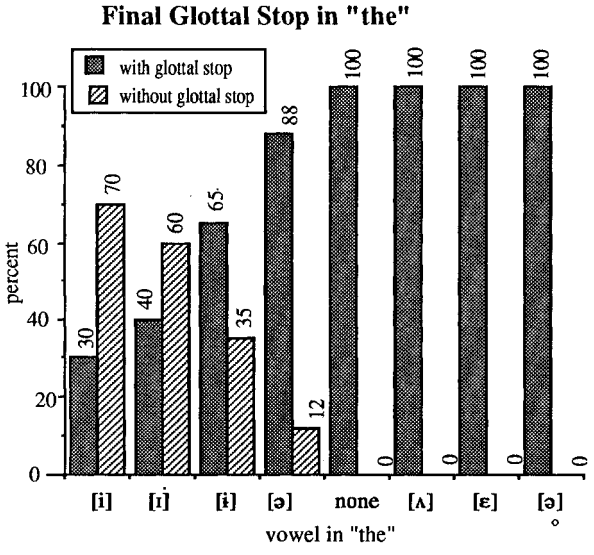

Figure 5 - Glottal stop vs. no glottal stop between "the" and words beginning with vowels, according to the vowel found in "the".

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

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

A second study using the TIMIT segment durations concerns the occurrence of so-called "epenthetic [t]" in certain contexts. The most common context in TIMIT is between underlying /n/ and /s/. With [t] epenthesized between these, "sense" can be pronounced like "cents" and "prince" can be pronounced like "prints", etc. Many phonologists and phoneticians have discussed this phenomenon, which is seemingly pervasive among speakers of American English. A particularly interesting claim that has evoked much attention is that epenthetic [t] is shorter in duration than [t] from underlying /t/ [2]. This claim was based on 60 tokens of each type only some of which showed the difference. TIMIT provides many more tokens: 187 (26%) of the 712 tokens with underlying /ns/ were transcribed with an intervening [t] (closure or release or both), and there are 129 tokens of /nts/. These tokens were subdivided according to the stresses of the preceding and following vowels and according to utterance-final vs. nonfinal position. In every comparison the epenthetic [t]s were a few msec shorter than the phonemic [t]s, but never significantly so. Thus TIMIT does not provide clear evidence for or against a durational difference, though it does make clear that if such differences exist they are not robust enough for word recognition in a speech recognition system. Small differences that may be significant in controlled laboratory speech samples or with homogeneous groups of speakers may be washed out in a speech database like TIMIT (see also [1]).

III. ACOUSTIC STUDIES

Acoustic analysis can also be performed on speech tokens identified by database searches. We are currently using Kay Elemetrics's CSL, which can read TIMIT files and display the time-aligned phonetic transcription with the waveform, though each token must be individually analyzed. We are analyzing the velar stops in TIMIT to test the claim that velar stops in English are fronted before front vowels, but not after them ([4]). We will present results of this study orally at this conference, and in a forthcoming issue of UCLA Working Papers in Phonetics devoted to studies of TIMIT. More detailed accounts of the studies presented here will also be found there.

IV. CONCLUSION

The TIMIT CD-ROM speech database can be a useful tool for linguistic phoneticians interested in describing the phonetic characteristics of American English. It does have some limitations, however, even beyond the obvious limitation to read speech. First, it requires additional software for any kind of analysis; however, commercial database programs and acoustic analysis systems are viable options for at least some research questions. Second, though the speech sample has been designed to contain all phoneme combinations, the number of tokens of any one combination is typically small. If comparisons of particular sequences are desired, there will usually not be enough tokens to offset all the variation in other aspects of the tokens, such as prosodic environment. Thus TIMIT does not comprise enough speech for reliable study of many phonetic and phonological hypotheses, which typically refer to highly specific contexts; our study of epenthesis is an example.

Third, the regional dialect coding of speakers is only marginally useful, since the dialect regions are broad areas of the country of very different sizes, which tend to wash out all but the most pronounced phonetic differences and underrepresent large population groups such as California. Nonetheless, we hope to have shown that if such limitations are kept in mind, TIMIT provides a powerful new kind of resource for phonetics.

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

REFERENCES

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

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

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

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

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

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

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

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

ACKNOWLEDGMENT

Some of this work was supported by the Committee on Research of the UCLA Academic Senate.