Relations of sex and dialect to reduction

Dani Byrd

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

Received 8 April 1993; revised 18 May 1994

Abstract

A set of phonetic studies based on analysis of the TIMIT speech database is presented which addresses topics relevant to the linguistic and speech recognition communities. First, the advantages and shortcomings of using TIMIT for linguistic research are considered, and a database methodological approach is outlined. Next, several small studies are presented which detail new results on the effect of speakers’ sex and dialect region on pronunciation. The goal of this paper is to use the database to explore sex and dialect related variation thereby ascertaining differences which may merit further experimental study. This report concerns speaker-dependent effects on certain phonetic characteristics often involved in reduction such as speech rate, stop releases, flapping, central vowels, laryngeal state, syllabic consonants, and palatalization processes. Specifically, it is suggested that the phonetic characteristics found more commonly with male speakers are also those typical of reduction in speech.

Zusammenfassung


1 A shorter and more preliminary version of this work was presented as a talk (Byrd, 1992a). A preliminary report on a subset of the topics presented here uses the earlier release of the training portion (two-thirds) of the TIMIT database and appears in (Byrd, 1992b).
2 Now at Haskins Laboratories, 270 Crown St., New Haven, CT 06511-6695 USA, where correspondence should be addressed.

0167-6393/94/$07.00 © 1994 Elsevier Science B.V. All rights reserved
SSDI 0167-6393(94)00025-5
Résumé

Cet article présente une série d'études phonétiques qui sont fondées sur une analyse du corpus TIMIT de l'anglais américain parlé, et qui sont susceptibles d'intéresser les spécialistes de linguistique et de reconnaissance de la parole. D'abord, nous discutons les avantages et les défauts de TIMIT pour la recherche linguistique, et nous indiquons l'esquisse d'une méthodologie utilisant la base de données. Ensuite, plusieurs petites enquêtes révèlent de nouveaux résultats sur les effets du sexe et du dialecte des locuteurs sur la prononciation. Cet article essaie d'utiliser la base de données pour examiner la variation attribuée aux différences de sexe et de dialectes afin d'établir les différences qui pourraient mériter une étude expérimentale plus approfondie. Ce rapport porte sur la variation entre locuteurs qui se manifeste dans certaines caractéristiques phonétiques qui sont souvent associées à la réduction, telles que le débit de la parole, les relâches d'occlusives, les sons battus, les voyelles centrales, les laryngales, les consonnes syllabiques, et le processus de palatalisation. Plus précisément, nous suggérons que les traits phonétiques qui sont le plus fréquent chez les locuteurs masculins sont ceux qui caractérisent la réduction dans le langage parlé.

Keywords: TIMIT; Gender; Sex; Dialects; Reduction; Speaker-specific; Speaker-dependent; American English; Database; Allophonic variation

1. Introduction

The majority of acoustic-phonetic studies to date have shared, in the most general terms, a common methodology. Each speaker reads an entire set of carefully controlled experimental speech materials designed to answer the specific questions motivating a given experiment. Much valuable linguistic knowledge has been gathered from experimentation of this sort; however, this general method has limitations. It considers a small number of homogeneous speakers which may be unrepresentative of the diversity found in a larger population of the language’s speakers. The limitation to carefully controlled test items may focus the speaker’s attention on contrasts, thereby exaggerating them. Finally, a new experiment must be designed and executed for each new question which arises. Recently, however, more interest has emerged in linguistic analysis of spontaneous speech materials; see for example the special issue of Speech Communication (Vol. 11, Nos. 4–5, October 1992) on speaking styles.

Another recent trend in speech investigation is the development of general-purpose speech databases. Such databases are well-suited to acoustic-phonetic analysis where small variations of specific sentential context are irrelevant due to the size and diversity of the data set. General factors in pronunciation variability, such as speaker sex, can also be investigated in large databases. In 1966, Labov described the advantages of having a large sample of speakers for investigating pronunciation variants, specifically in studying the effects of ethnicity, age and other independent variables (Labov, 1966a, pp. 105, 107). In practice, a large speech database will include either a lot of speech from relatively few speakers, or fewer samples from many speakers. Examples of the first type include sociolinguistic studies, and the New York University database (Umeda, 1991). An example of the second type is the TIMIT database for American English.

The TIMIT database, which is described below, was designed jointly by the Massachusetts Institute of Technology, Texas Instruments, and SRI International under sponsorship from the Defense Advanced Research Projects Agency–Information Science and Technology Office (DARPA–ISTO) for the development and evaluation of automatic speech recognition systems, as well as the acquisition of acoustic phonetic knowledge (Lamel et al., 1986). It is distributed on CD-ROM by the U.S. National Institute of Standards and Technology and the Linguistic Data Consortium. Much of the published research using TIMIT addresses database design or incorporates TIMIT as a corpus for training and testing speech recognizers. However, little linguistic work has been reported to date. Allophonic variation has been studied by Cohen (1989), Randolph (1989), Riley and Ljolje (1992),
and Withgott and Chen (1993). It was desired that the database incorporate sufficient variability to examine details of the acoustic realization of phonetic segments as affected by canonical characteristics of the phoneme, contextual dependencies, syntactic effects, and such speaker-specific factors as dialect, sex, age, and education (Lamel et al., 1986). Because it is such a large and meticulously transcribed database, TIMIT also provides an interesting testing ground for linguists to assess the accuracy of generalizations regarding allophony and regularity in English that have previously been based on laboratory experiments or naturalistic observation. Additionally, the linguist's perspective may highlight fertile areas in which to gather acoustic-phonetic data relevant to the phonetic classification and speech recognition goals which TIMIT serves.

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

It is distressing how little is known about the effects of speaker sex and dialect on the choice of allophonic variants. Even though acoustic studies of sex and dialect variation are more numerous, their coverage is also not complete. For example, Henton's (1992) survey of phonetic studies having “representative” data on vowels showed that only 4% contained data from female speakers. Widely used phonetic textbooks lack sufficient and appropriate information about female speech, often leaving their students without having seen a spectrogram of a female voice. The studies that have considered sex differences in pronunciation have often been of two varieties. First, such studies may compare acoustic differences within a particular category, for example, the vowel [ʌ], or due to physiology, for example, voice quality. Alternatively, some studies have concentrated on stylistic differences such as choice of lexical items. Research along both of these lines has often addressed sound change in progress. The following approach differs from these two in that we are considering distributional differences in the occurrence of common allophones of American English. We interpret certain of these differences to be indicative of greater or lesser reduction. Additionally, the use of the TIMIT database offers a much larger pool of speakers than is often available in either sociolinguistic or acoustic phonetic studies.

Speaker-specific variation is of interest to the linguist attempting to describe the universal, language-specific, and speaker-dependent characteristics of speech. For example, Labov states that “sexual differentiation of speech often plays a major role in the mechanism of linguistic evolution” (1972, p. 303). For example, Herold (1990; in Labov, 1991) in studying the merger of don and dawn in Philadelphia found that girls led substantially in this merger. Exploring how physiological and social factors interact in the speech of both men and women is important in understanding the linguistic principles governing language variability. An improved understanding of the influences of speaker-dependent variables on both acoustics and allophonic variation may also be valuable in improving the performance of recognition systems which will presumably be employed by a wide variety of users. Complementary research in the field of speech recognition will be reviewed in Section 4.

The general goal of this study is to highlight areas of phonetic variation between men and women and between broad geographical dialect divisions in order to determine where linguists'
understanding might profit from further experimental investigation. The data presented below is not meant to be a comprehensive or in-depth examination of sex and dialect differences; it is, however, intended to illuminate potential areas of interest.

2. Method

Because of the quantity of speech and the segmentation and labeling, TIMIT provides a unusual corpus for phonetic research. The TIMIT database includes 630 talkers and 2342 different sentences, comprising over five hours of speech. The sentences are of three types. Two calibration sentences are spoken by every talker. These sentences were designed to “incorporate phonemes in contexts where significant dialectical differences are anticipated” (Zue et al., 1988). They were:

She had your dark suit in greasy wash water all year.

Don’t ask me to carry an oily rag like that.

Additionally, 450 phonetically-compact sentences were designed to incorporate as complete a coverage of phonetic pairs as practical (Lamel et al., 1986), i.e., phonetically-compact denotes that examples of phonemes in all possible left and right contexts are included. Each one of these sentences is spoken by seven talkers. Finally, 1890 sentences were chosen to provide a variety of contexts and multiple occurrences of the same phonetic sequence in different word sequences (Fisher et al., 1986). These sentences are drawn from the “Brown Corpus” (Kuchera and Francis, 1967) and a playwright’s dialog corpus (Hultzen et al., 1964). Each talker reads two calibration sentences, five phonetically compact sentences, and three from the varied group.

The speech was digitally recorded at 20 kHz with a peak signal to noise ratio of 29 dB. The signal was high-pass filtered at 70 Hz and down-sampled to 16 kHz (Zue et al., 1990; Fisher et al., 1986). The version of TIMIT which is distributed was recorded with a close-talking microphone. Speakers were seated in a recording booth and were presented with prompts on a monitor. Documentation distributed with the corpus describes the instructions to the speakers as minimal, noting that they were asked to read in a “natural” voice (Garofolo et al., 1993). The digitized recordings are segmented and labeled with a time aligned phonetic transcription. The transcriptions were done at MIT and were rechecked between the first and final release of the database. A description of the inventory of transcribed elements and criteria for segmentation can be found in (Zue and Seneff, 1988). Transcribed phones include stop closures and releases, syllabic and non-syllabic nasals and laterals, flaps (nasal and non-nasal), pauses, epenthetic and glottal stops, and a wide variety of vowels including [a], [y], [æ], [ə], and [i]. An orthographic transcription and the waveform are also provided. The validity of the results reported in this work depends entirely on the correctness and consistency of the phonetic transcriptions. It has been our experience at UCLA that the TIMIT transcriptions are very reliable. See (Keating et al., 1994) for a description of UCLA’s database implementation of TIMIT which was employed in the following study. Because the transcriptions are narrow and acoustically defined, they were chosen as the basis for the studies presented below. These transcriptions are especially helpful in obtaining an overall picture of the distribution of the high frequency phones across the database.

The use of the TIMIT transcriptions in linguistic study is described in detail in (Keating et al., 1994). Recapitulating the remarks there, let me emphasize that studies of these labels are acceptable only to the extent that the segmentation and labeling is itself accepted as an accurate record of linguistic aspects of the acoustic signal. As Keating et al. note, the hand-labeling of TIMIT was carefully done and checked and is implicitly accepted and found useful by all those who have used the database to train statistical models for automatic speech recognition. No other large corpus of transcribed speech can be expected to be as reliable.

Eight dialect regions were established for clas-
Like any experimental corpus, TIMIT has its limitations and shortcomings. Some of these shortcomings are noted below and are also detailed in (Keating et al., 1994). The speech is, of course not spontaneous, but read. The sentences are relatively short and usually declarative. The recording setting, as is usually the case, does not inspire a casual speech style. The speakers are mostly white, mostly in their 20's and 30's, and, as noted above, mostly male. The dialect divisions are too general for any detailed dialect studies. Also, as no phonemic transcription is provided, automating an investigation into many phonological processes is difficult or impossible. Finally, the number of tokens of any particular combination or sequence of phones is likely to be too small to offset variation in other aspects of the tokens. TIMIT is best suited for giving a broad overview of American speech which can be used to pinpoint areas for further detailed linguistic study.

3. Results

3.1. Speech rate

The first and perhaps most important quality examined for the TIMIT speakers was their speaking rate. The duration of both calibration sentences was used to examine speech rate as all 630 speakers read these same two sentences. A three-factor analysis of variance (ANOVA) with the factors of sex, dialect region, and calibration sentence number (#1 or #2) with all interactions was used to test the effect of sex and dialect region on speaking rate. There is a significant effect of sex on rate ($F(1, 1228) = 37.301$, $p = 0.0001$) with men speaking 6.2% faster than women. Recall that calibration sentence one has 13 syllables, and calibration sentence two has 12 syllables. Men had an average rate of 4.69 syllables/second and women a rate of 4.42 syllables/second (based on an average of 12.5 syllables in the calibration sentences).

There is also a significant effect of dialect region ($F(7, 1228) = 5.424$, $p = 0.0001$). The dialects range from slowest to fastest in the follow-
Fig. 1. Average duration of calibration sentences in ms for male and female speakers.

Fig. 2. Average duration of calibration sentences in ms for speakers in each TIMIT region; standard error is shown by y-error bars.

In order to determine whether the frequency or duration of pauses could have contributed to the above effects, these were examined in the calibration sentences. A chi-square test \(^3\) determines that pauses are randomly distributed between men and women but are not randomly distributed between dialects \((\chi^2 = 19.325, p = 0.0072)\). Speakers from the South Midland and the South paused more often than expected while speakers from the North Midland, West, and the "Army Brat" paused less often than expected given a random distribution. This result explains, at least in part, the effect of dialect region on rate described above, as pauses contributed to sentence duration. A three-factor ANOVA shows no effects (or interactions) of sex, dialect region, and sentence number on the duration of pauses.

In summary, both sex and dialect have significant influences on speaking rate. Differences in speaking rate are important to bear in mind in the overall examination of reduction, as rate has a substantial influence on the production of reduced word forms.

A recent report on French found no overall effect of sex on speaking rate (Ryalis et al., 1994); however, they considered 40 speakers reading sentence of only five to seven syllables—about half the length of the TIMIT sentences. However when the under-age-60 subset of their data is considered, the women have a longer mean final syllable duration (288 ms) in statements than men (264 ms). Applegate (1984) found women to have significantly longer phrase durations than the men, and found a trend \((p = 0.052)\) for this to be the case for sentence duration.

\(^3\) This and the following chi-square (contingency table) tests assume that each of the observations is independent. Because a single speaker may produce more than one token of the observed phone, this assumption is not strictly accurate. However, it is felt that the independence of each observation is generally preserved as a single speaker did not produce more than a small fraction of the extremely large number of tokens in most of the tests detailed below.
We will see in several of the results below that not every phone in TIMIT displays durational differences. In particularly, many of the short or reduced phones would not be expected to, and do not, display longer durations depending on speaker sex. The rate effect is probably distributed to some degree across many different articulatory units, but is also likely to be concentrated in the full vowels; see (Gay, 1981).

3.2. Sentence-final stop releases

All sentence-final oral stops (n = 1130) were evaluated as to whether their closures have a release or not. A contingency table analysis was conducted where the expected number of releases is assumed to be randomly distributed with respect to sex or dialect within the group of speakers who produced stops in this position. The contingency table analysis determines that sex has a significant effect on the distribution of final released and unreleased stops ($\chi^2 = 11.651$, $p = 0.0006$). Women released their sentence-final stops more often than men: 67% versus 56% of the time. There is no significant effect of dialect region on the frequency of releases in sentence-final stops. As place of articulation has a significant effect on the frequency of release of sentence-final stops (see (Byrd, 1993)), the effect of sex was tested separately at each of these places. This contingency table analysis shows significant effects of sex on the frequency of release at the alveolar and velar place of articulation but not at the bilabial place of articulation.

The final word “that” of one of the calibration sentences read by all 630 speakers also ends in an oral stop. This stop is realized as released 23% of the time, as unreleased 67% of the time and as a glottal stop 9% of the time. (Five cases were excluded due to collection error in searching the database.) For the speakers who produced a stop in this position, a contingency table analysis determines there to be a significant effect of sex ($\chi^2 = 5.57$, $p = 0.0183$) but no effect of dialect on whether a release was produced. Women released this stop 32.5% of the time and men 23.1% of the time. In summary, the sex of the speaker exerts a significant effect on the frequency of a sentence-final stop release. Such releases are characteristic of carefully articulated speech and are less often found in reduced pronunciations.

3.3. Flaps

Another process found in continuous speech is alveolar flapping. This rule as stated by Oshika et al. (1975) describes a process whereby an intervocalic stop, optionally preceded by [r] or [n], is realized as a flap when it occurs in a falling stress pattern (as in “winter”) or between reduced vowels (as in “ability”). Across word boundaries, there are no stress conditions (as in “what#is” or “not#equal”) (Oshika et al. 1975). Two analyses of flaps in TIMIT were conducted. In the first, the frequency distribution of all oral and nasal flaps in the database was considered where the expected distribution given the null hypothesis is assumed to be random over the database as a whole, i.e. that men will produce 69.5% of the flaps and women 30.5%. A chi-square test indicates a significant effect of sex on the frequency of both nasal (n = 1331) and oral (n = 3649) flaps ($\chi^2 = 55.341$, $p = 0.0001$ and $\chi^2 = 11.41$, $p = 0.0007$, respectively).

The women produce significantly fewer flaps than the men. The result is in accordance with Zue and Laferriere’s (1979) finding that in an

![Fig. 3. Number of occurrences of nasal and oral flaps (combined) for men and women (value for women has been multiplied by the ratio of men to women in the database to normalize for differences in sample size).](image)
intervocalic [nt] sequence female speakers had a marked tendency to avoid deletion of the [n] and flapping; these variants occurring in only 4% of productions. Male speakers in contrast produced 96% of their tokens in one of these two forms. No effect of dialect region is found on the frequency of oral or nasal flaps.

A second analysis was conducted on the sequence "suit in" and the word "water" which are included in one of the calibration sentences read by all speakers. We would expect most speakers of American English to have a medial flap in "water". Word final flaps ($n = 121$) were produced by 19% of the speakers and word medial flaps ($n = 624$) by 99%. While there is, not surprisingly, no effect of sex or dialect on the frequency of a word-medial flap in "water", a chi-square test indicates a significant effect of sex on the frequency of flaps in the "suit in" sequence ($\chi^2 = 13.934, p = 0.0002$). Only 9% of the women flapped the alveolar consonant in this sequence while 19% of the men did. Dialect region also had an effect on the frequency distribution of the word-final flap ($\chi^2 = 20.03, p = 0.0055$). The North and North East speakers flapped less often than expected, and the North Midland speakers more often than expected. (Note: The effect of sex on the distribution of oral flaps across the whole database remains significant ($\chi^2 = 8.829, p = 0.003$) when the flaps in the calibration sentence are excluded from analysis).

3.4. Central vowels

Recall that vowels may be more peripheral in less reduced speech. Given the preceding findings, it might be expected that the men would produce more instances of the central unstressed vowel [ə] than the women. The second unstressed reduced vowel transcribed in TIMIT is [i]. According to Zue and Seneff (1988), [i] and [ə] are distinguished in the TIMIT transcription in the front-back dimension with [i] being transcribed when the second formant is closer to the third than the first, and [ə] being transcribed otherwise. It is unclear from Zue and Seneff (1988) whether a height distinction also exists, but one might infer based on standard phonetic convention that [i] is the higher vowel. Thus, while still central and unstressed it is more peripheral in both dimensions. A third central, but not stressless, vowel, [ʌ], is also relevant. Although no description of its phonetic quality is given, it is presumably somewhat lower and backer than [ə], that is, also more peripheral.

An analysis of the distribution of the three central vowel transcribed in TIMIT—[ə], [i], [ʌ]—was conducted. A total of 17,858 central vowels were transcribed of which 55% were [i], 18% were [ʌ], and 27% were [ə]. (Note: the calibration sentences were excluded from this analysis so as not to over represent any particular central vowel tending to occur there.) A chi-square test determines there to be no effect of sex or dialect region on the frequency distribution of the total number of central vowels. However, when each vowel is considered separately, differences in their use across sex and dialect emerge as seen in Fig. 4.

A chi-square test for the effect of sex and dialect region on the distribution of [ə] show significant effects of sex ($\chi^2 = 21.591, p = 0.0001$) and dialect region ($\chi^2 = 30.15, p = 0.0001$). Women used this central vowel less frequently than the men. Speakers from the North, NY City, and the West use this vowel less frequently while speakers from the N. Midland, South, and, especially, the South Midland use this vowel more frequently. Additionally, both sex and dialect have
a significant effect on the distribution of [i] (χ² = 7.161, p = 0.0074 and χ² = 14.203, p = 0.0477 respectively). Men use [i] less frequently than would be predicted by a random distribution. Speakers from the North East, NY City, and the West use [i] more frequently than expected and speakers from the North and N. Midland less frequently than expected. Finally, a chi-square test for the effect of sex and dialect region on the distribution of [a] shows a significant effect of sex (χ² = 5.79, p = 0.0161) but no effect of dialect region. Women used this central vowel more frequently than expected given a random distribution. In summary, we have found that women use the two more peripheral of the central vowels, [i] and [a], more frequently than men and the more central of the vowels, [a], less frequently than the man.

\[ \begin{align*}
\text{i} & \quad M-, \ W+ \\
\text{a} & \quad M+, \ W- \\
\lambda & \quad M-, \ W+
\end{align*} \]

This difference again suggests that the women's speech may be less reduced in certain respects than the men's. Though more difficult to interpret, there does appear to be some complementarity between [i] and [a] for the North Midland, New York City, and Western regions. Recall that there was no effect of dialect on the frequency of [a]. Speakers classified in the New York City and the West regions appear to have a greater preference for [i] than [a] relative to the other regions. The North Midland region behaves in the reverse manner having [a] more frequently. The North seems to use both these vowels less frequently than do the other dialects.

An ANOVA was conducted on the durations of these vowels as duration often corresponds to the degree of acoustic reduction as in cases of articulatory undershoot (Lindblom, 1963, 1964). Of course, reduction is related to both duration and formant frequencies which is why we considered both the vowel quality labels and duration. A four-factor ANOVA with all two-level interactions was conducted to test effects on vowel duration with the factors of vowel, sex, dialect region, and position in the word (medial, initial, final, or unaffiliated). Not surprisingly, vowel and position had significant effects, however, so did dialect region (F(7, 17796) = 6.668, p = 0.0001). The South and South Midland have the longest central vowels and differ significantly from most other dialect regions as determined by a post-hoc Scheffe's S test. These, recall, were also the slowest-speaking regions. Sex did not have a significant main effect. There were also significant interactions of vowel and dialect region; sex and position; and vowel and position. No other interactions were significant. Clearly, as the central vowel durations were not affected, the differences between the sexes in speaking rate lie in some other portions of the signal.

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

The distribution of glottal stop [ʔ], the voiceless vowel [ɻ], [h], and the voiced /ɦ/ [ɦ] were evaluated. Use of the glottal stop can occur in English between vowels, before a vowel, in place of an alveolar stop, and in many other positions. This corpus provides a data set for determining general distributional patterns of glottal stop across a variety of prosodic and phonological contexts. The frequency distribution of glottal stops (n = 4834) is significantly affected by both sex and dialect region (χ² = 30.906, p = 0.0001 and χ² = 148.154, p = 0.0001, respectively) as shown by a chi-square test.

Women have significantly more glottal stops than the men. Speakers from the North and

![Fig. 5. Number of occurrences of glottal stops for men and women (value for women has been multiplied by the ratio of men to women in the database to normalize for differences in sample size).](image)
South use more glottal stops than expected while speakers from the North Midland and the "Army Brats" use less. When the position of the glottal stop in the word is considered (initial (49%/total), medial (6%/total), final (16%/total), or unaffiliated (not part of a word) (29%/total), the effect of sex on the frequency of glottal stops is significant at all positions, with the effect always in the direction indicated above. The effect of dialect region is significant in initial position and final position only. It is somewhat unexpected to find that the speaker-dependent characteristic of sex was related to the frequency of glottal stop in this way. In fact, women's voices are often characterized as more breathy (see for example, (Henton and Bladon, 1985, and Klatt and Klatt, 1990)), and men's as more often employing creak (see for example, (Henton and Bladon, 1988)). Glottal closure is often related to creakiness in the voice quality of the signal, and, in TIMIT, laryngealization associated with a word-initial vowel was segmented as a glottal stop. It may be that the glottal stop is used as a devoicing mechanism more often by women or that it participates in allophonic patterns which are less productive for the men.

When we consider the small sample of 57 glottal stops produced in place of the sentence final [t] of the word "that" in one calibration sentence, we find that the production of a glottal stop in this position is not significantly influenced by sex ($\chi^2 = 1.763$, $p = 0.1843$), although the distribution favors the direction demonstrated above.

An ANOVA testing the effect of sex and word position on the duration of all glottal stops in the database, found there to be no significant main effect of sex. There was, however, a significant interaction of sex and position ($F(3, 4826) = 4.554$, $p = 0.0034$). For men, the word internal glottal stops are longer than initial glottal stops; for women, the reverse is the case. Word-internal position was the only position in which the mean duration of the glottal stop was shorter for females than for males, a difference of approximately 7 ms.

The devoiced schwa is a very short vowel typically occurring between voiceless consonants in TIMIT and having no "vocalic" portion or only one or two visible pitch periods (Garofolo et al., 1993). All ($n = 478$) instances of the (only) voiceless vowel [ʊ] transcribed in TIMIT were evaluated with respect to sex and dialect region of the speaker. The frequency distribution of the voiceless vowel shows a significant effect of sex ($\chi^2 = 36.471$, $p = 0.0001$) but no effect of dialect region as determined by a chi-square test.

Women have significantly fewer voiceless vowels than the men. Again, this result is surprising given the commonly accepted conception of women's voice quality as being more breathy than men's. It is, however, not unexpected in light of the findings reported here which suggest that women produce less reduction in their speech than do men. Many voiceless vowels would presumably be created by overlapping a neighboring laryngeal opening movement(s) with the syllable nucleus. Zue and Seneff (1988) offer the word "secure" as a sample context. The generally less reduced forms and slower speech apparently produced by women would be likely to have less overlap.

In comparison with the voiceless schwa, all ($n = 1313$) [h]'s in TIMIT were evaluated. The frequency distribution of [h] shows a trend due to sex ($\chi^2 = 3.815$, $p = 0.0508$) but no effect of dialect region as shown by a chi-square test. Similar to [ʊ], women have fewer [h]'s than the men. Here, we again see what appears to be an odd
result given assumptions about breathiness in female voices. This result then merits further investigation.

Finally, we can compare the findings regarding voiceless schwa, which we suggested might be caused by coproduction with adjacent voiceless consonants, with the voiced [f]. Voiced [f] generally occurs in TIMIT between vowels (Zue and Seneff, 1988), thus can be considered to be the result of coproduction with adjacent voiced sounds. We might, based on the findings above, expect it to be more frequently produced by men than by women. However, when the frequency distribution of all [f]'s in TIMIT ($n = 1523$) is examined, we find no effect of sex or dialect region as determined by a chi-square test. This differs from the findings regarding voiceless schwa. While we suggested that coproduction of schwa with neighboring voiceless sounds might be greater for men than women, it seems that a potentially parallel mechanism for voicing in [h] is not exhibited.

3.6. Syllabic consonants

Syllabic consonants are syllable peaks and are longer than their non-syllabic counterparts (Roach and Sergeant, 1992). Some lexical items, like "bottle", necessarily have syllabic consonants. We expect no differences between men and women in such productions, much as we saw that all speakers produced a medial flap in "water". However, many syllabic consonants arise as the result of complete reduction of the vocalic syllable nucleus. For example, frequently, endings like [−n] may reduce to [−n]. Given the findings above with respect to reduction, we might expect the men overall to produce more instances of syllabic consonants than the women. Accordingly, all the syllabic consonants transcribed in the database were evaluated for speaker-specific effects on their distribution. These consonants include [l] ($n = 1291$, 52% of total syllabic consonants), [m] ($n = 171$, 7% of total), [n] ($n = 974$, 39% of total), and [ŋ] ($n = 43$, 2% of total). No effect of sex and dialect region is found in a chi-square test on the distribution of [l], [m], and

![Fig. 7. Number of occurrences of syllabic [n] for men and women (value for women has been multiplied by the ratio of men to women in the database to normalize for differences in sample size).](image)

[ŋ]. (The chi-square test is not valid for effect of dialect region on [ŋ] distribution due to its infrequent occurrence.) Additionally, no effect of dialect region was found for [ŋ]. However, the sex of the speaker did have a significant effect on the frequency of [n] ($\chi^2 = 12.632$, $p = 0.0004$), such as might occur in the word "hidden" or "button".

Women use significantly fewer syllabic [n]'s than the men. The findings did not support the general hypothesis that men have more syllabic consonants. It is important to note that men and women appear to produce this, and only this, syllabic consonant with different frequencies. However, it might have been easier to find the differences in [n] distribution because there were substantially more tokens of the syllabic [n] than syllabic [m] or [ŋ] (but not more than the syllabic [l], however). Additionally, instances of the allophone [n] can occur for canonical [ŋ]. The other syllabic consonants might be more limited to lexical items in which their production is not optional. That is, production of [n], in particular, may be more reflective of reduction in casual speech than production of the other syllabic consonants.

3.7. Palatalization

Because palatalization is a well-known and frequently reported casual speech process in
American English, potential environments for palatalization in TIMIT were examined to determine if speaker sex or dialect region affected whether palatalization occurred. In palatalization alveolar obstruents become palatals or palatalized before palatals. All sequences of "s_sh" [ʃʃ], "z_sh" [ʒʃ], "sh_s" [ʃʃ], and "sh_z" [ʃʒ] occurring across a word boundary in the canonical forms of the TIMIT sentences were evaluated to determine whether both consonants were produced by the speaker or whether assimilation occurred. (Sequences of palatal-alveolar were also included to allow for the possibility of carryover coarticulation as well as anticipatory. See (Byrd, 1992c). The presence of a pause (as determined by the TIMIT transcription) between words was also noted. The null hypothesis is that the number of assimilations and pauses are randomly distributed with respect to dialect region and sex within the group of speakers reading these sequences. However, we have seen what we suggest are differences in amount of coarticulation as a function of speaker sex in Sections 3.5 and 3.6.

A contingency table analysis of these potential palatalization sites determines there to be no significant effect of sex on whether both consonants were produced or whether there was a pause between them. This is the case both when C1 is the post-alveolar consonants and when it is an alveolar consonant. Assimilation to a single consonant occurred 70% of the time. (See (Byrd, 1992c) for a general report on palatalization in TIMIT.)

In a second analysis, one of the calibration sentences including the phrase "had your" was investigated. Three types of productions were included in the contingency table analysis. In 44% of the cases the intervocalic sequence [dj] was produced (where in the TIMIT transcription [j] denotes the release portion of an affricate); in 20% of the cases [dʃj] was produced, and 36% of the time [dj] was produced where the last two productions differ in the absence or presence of an alveolar release before the glide. Thirty speakers who produced unusual sequences occurring in less than 1.5% of the cases were not included in the analysis. The null hypothesis is that each of the three sequences described above are randomly distributed with respect to dialect region and sex within the remaining group of 600 speakers. A contingency table analysis determines there to be no significant effect of sex or dialect region on which sequence was produced; nor is there any effect on whether the stop was released in the cases where it occurred before a glide. Lastly, there is no effect of sex or dialect region on whether an affricate or a glide was produced. Thus it seems that speakers do not differ by sex or dialect in this pervasive casual speech process. Palatalization seems to occur relatively consistently regardless of other tendencies which may exist in terms of reduction.

3.8. Summary of speaker sex effects

To summarize the significant findings regarding the effect of speaker sex in the TIMIT database, consider Table 3.

However, no differences were found in palatalization across a word boundary, nor in the frequency of [ʃ], [l], or [m].

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Summary of significant effects of speaker sex on frequency of phone occurrence in TIMIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech rate</td>
<td>Men speak faster than women</td>
</tr>
<tr>
<td>Sentence-final stop releases</td>
<td>Men release stops less often than women</td>
</tr>
<tr>
<td>Central vowels</td>
<td>Men use [a] more often, [l] and [a] less often than women</td>
</tr>
<tr>
<td>Glottal stops</td>
<td>Men produce glottal stop/laryngealization more often than women</td>
</tr>
<tr>
<td>Syllabic consonants</td>
<td>Men produce syllabic [n] more often than women</td>
</tr>
<tr>
<td>Voiceless schwa</td>
<td>Men have more voiceless schwa's than women</td>
</tr>
<tr>
<td>[h]</td>
<td>Men tended to have more [h] than women</td>
</tr>
</tbody>
</table>
4. Discussion

This report has shown that in the TIMIT database speaker-specific characteristics of sex and dialect region influence speaking rate, the choice of central vowels, and the frequency distribution of stop releases, flaps, glottal stops, breathy vowels, and the syllabic alveolar nasal. Additionally, it is suggested that many of these phonetic characteristics produced more frequently by men in this read corpus are typical of reduction in speech. Various researchers such as Kramer (1978), Henton (1992; using data in Bladon et al., 1983; and Henton, 1983, 1985), and Labov (1966b, 1970) have noted greater distinctiveness in female speech than in male speech. Henton finds for British English a larger, more peripheral vowel space for females than for males, particularly in the F1 dimension. In studying allophonic variants in the TIMIT calibration sentences, Cohen (1989) notes that females tend to use less reduced (i.e. “more careful” or “slower” speech) forms than males. Smith (1979) and Thorne (1983) cite several studies finding more “standard” pronunciation by girls and women than by boys and men. He notes that Fischer (1958), Anshen (1969), Wolfram (1969), Fasold (1968), Shuy et al. (1967), and Trudgill (1974, 1975) find more “-ing” versus “-in” endings for females. Shuy et al. (1967) also find men more likely to use nasalized vowels, with words like “man” realized as [m æ], and Wolfram (1969) concludes that women are more likely to produce final consonant clusters and post-vocalic [ɹ]’s. Similarly, several studies have noted that /ə/ is more often reduced or altered by men than women (Anshen, 1969; Labov, 1966b; Milroy, 1976; Wolfram 1969). Many writers have found “prestige” forms to be used more often by women than men, particularly in formal settings. See (Thorne, 1983) for an overview. It does not seem unreasonable to suppose that the “standard” or “prestige” forms are the less reduced forms.

It is not suggested that the results described here illuminate causes of these correlations; these effects may be task-specific or not. We have no obvious, systematic, way to determine whether different groups of speakers used similar speech styles in reading for the TIMIT recordings, despite the efforts of the database designers to maintain uniformity. For example, Labov found at least five stylistic levels differentially affecting certain phonetic studies: casual speech, careful speech, reading, word lists, and minimal pairs (Labov, 1970, p. 49) However, the results above do suggest that speech analysis for both synthesis and recognition goals will provide a more comprehensive picture of variation if similar numbers of men and women are included in speech databases. Furthermore, because there are many aspects of pronunciation which seem to differ between men and women, it may be profitable to incorporate within a recognition lexicon different probabilities leading to particular pronunciations for a male as compared to a female speaker. Or, if a single most likely pronunciation is sought, certain differences in the lexicon for male and female speakers might improve accuracy. It is interesting to consider whether many of the effects described are highly enough correlated with speaking rate to make rate rather than sex a driving force for a recognition system. For example, Cohen (1989) found that many of the speaker-specific effects he noted were also dependent on speech rate.

In Section 1 we claimed that an understanding of sex and dialect variation might, in addition to being of linguistic interest, be relevant in the pursuit of machine speech recognition. Cohen and his colleagues at SRI have explored ways of improving the lexical representations in an automatic speech recognition system to better reflect natural speech, as well as methods for faster and more accurate adaptation to a speaker (Cohen et al., 1987a). Of specific relevance here is their research on grouping speakers into a small number of “pronunciation clusters” predictable by demographic and other speaker-specific characteristics such as speech rate. Examining the co-occurrence of certain pronunciation variants for words in the TIMIT calibration sentences, they found that if the appropriate pronunciation cluster is known, predictive ability can be increased by 30–50%. They were interested in determining whether speaker-specific predictors of cluster membership are available and whether cluster
membership is consistent within a speaker. A significant non-independence between cluster and all tested demographic factors including age, race, region, sex, and education was found (Cohen et al., 1987a). Additionally, cluster membership was somewhat dependent on speech rate (Cohen et al., 1987a). They also found it inadequate to choose a single cluster for a speaker and suggest that a weighting function over all the clusters given some experience with a speaker may be reasonable (Cohen et al., 1987b).

Cohen (1989) found a significant increase in predictive power for several allophonic forms occurring in the calibration sentences when speaker-specific characteristics, including rate, are considered. Effects of dialect region included rlessness in “your”, “dark” and “water”, voicing of the fricative in “greasy” (see also (Keating et al., 1992)), pronunciation of the vowel in “like”, and word-initial glottalization in “an” (Cohen, 1989, p. 91). Cohen found sex to significantly affect the presence of word-final stops in “don’t ask” and “suit in”, bursts for the final stops in “ask” and “that” and glottalization at the beginning of “all” and “oily” (Cohen, 1989, p. 91). Unfortunately for our purposes, he does not detail the nature of these effects. Cohen (1989) showed that automatic clustering of utterances into similar pronunciation groups based on certain allophonic choices resulted in a model with lower entropy (better predictive power) than explicit monitoring of speaker-specific characteristics. Cohen et al. (1987b) note that “some co-occurring groups [of allophonic choices] may correspond to dialectal phenomena, which remain fairly consistent within a speaker. Other[s]... may reflect phenomena that change within a single speaker, due to things such as speech rate, and these clusters may need to be rechosen from time to time for an individual speaker.” (p. 3).

The findings presented in this report suggest that some consistent relationship between speaker sex and reduction in read speech might be useful in automatic speech recognition. Reduction is a concept which unifies effects that Cohen and others have preliminarily treated as co-occurring by accident. That is, for at least some phenomena like reduction, one need not separately model all possible cases; rather, the sex of the speaker may provide a reliable prediction that a variety of reduction processes will occur with increased or decreased frequency.

5. Conclusions

We have seen here a number of indications that sex, and to a lesser extent, dialect region may influence reduction processes even in this relatively formal scripted speech. The geographical regions used in TIMIT appear to be too broad for many linguistic purposes, yielding inconclusive results here. In fact, many of the dialect region effects reported here are possibly due to differences in the representation of men and women across the various dialects. Unfortunately, the limited number of women makes it difficult to adequately address sex and dialect interactions here. Also, given the fact that there are eight dialect regions, the statistical tests used are less powerful in discerning differences between the dialect regions than between the sexes.

The report presented here and similar database studies offer a promising new methodology for approaching speech analysis. Examination of speech databases can help inform phoneticians as to the extent to which we can generalize from experimental studies to corpora including a great deal of variation in reading material and speakers. Of course, an even broader scope of study would be provided by comparison of large, transcribed, read corpora with non-read corporuses, something not now possible. TIMIT has proven to be fertile ground for gathering acoustic-phonetic knowledge which is of interest to a wide spectrum of speech scientists.

Acknowledgments

This research was supported by a National Science Foundation Graduate Fellowship and the Department of Linguistics at UCLA. The author is grateful to Oliver Foellmer, Patricia Keating, Peter Ladefoged, Ian Maddieson, Caroline Smith, Richard Wright, Lisa Zsiga and two anonymous
Speech Communication reviewers for their valuable assistance. Many thanks are due to Edward Flemming for designing and implementing the structure of a relational Macintosh database with which TIMIT is used at UCLA.

References

J. Applegate (1984), Selected measures of speech sounds as a function of speaker age and sex, Ph.D. dissertation, Kent State University Graduate College.


W. Labov (1966b), The social stratification of English in New York City, Center for Applied Linguistics, Washington, DC.


W. Labov (1972), Sociolinguistic Patterns (Univ. of Pennsylvania Press, Philadelphia).


