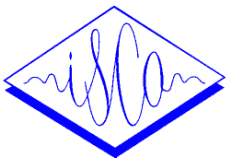

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

SEX, DIALECTS, AND REDUCTION

Dani Byrd

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

ABSTRACT

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

INTRODUCTION

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

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

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

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

METHOD

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

The digitized recordings are accompanied by a time aligned phonetic transcription. Transcribed phones include stop closures and releases, syllabic and non-

10.21437/ICSLP.1992-271

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

Central Vowels

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

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

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

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

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

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

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

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

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

Table 1 - Sex and Dialect Distribution in the TIMIT database

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

RESULTS AND DISCUSSION

Speech rate

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

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

Sentence-final stop releases

All sentence-final oral stops (n=1130) were evaluated as to whether their closures have a release or not. A contingency table analysis was conducted where the expected number of releases is assumed to be randomly distributed with respect to sex and dialect within the group of speakers who produced stops in this position. The contingency table analysis determines that sex has a significant effect on the distribution of final released and unreleased stops ($\chi^2=11.651$, $p=.0006$). Women released their sentence-final stops more often than men: 67% versus 56% of the time. There is no significant effect of dialect region on the frequency of releases in sentence-final stops. As place of articulation has a significant effect on the frequency of release of sentence-final stops ($\chi^2=40.829$, $p=.0001$) with the probability of a release increasing from the bilabial to the alveolar to the velar place, the effect of sex was tested separately at each of these places. This contingency table analysis shows significant effects of sex on the frequency of release at the alveolar and velar place of articulation but not at the bilabial place of articulation.

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

Flaps

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

Syllabic consonants

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

Palatalization

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

In a second analysis, one of the calibration sentences was investigated where the sentence included the phrase *'had_your'*. Three types of productions were included in the contingency table analysis. In 44% of the cases the intervocalic sequence [dj̈] was produced; in 20% of the cases [d˗y] was produced, and 36% of the time [dy] was produced where the last two productions differ in the presence or absence of an alveolar release before the glide. Thirty speakers who produced unusual sequences which occurred in less than 1.5% of the cases were not included in the analysis. The null hypothesis is that each of the three sequences described above are randomly distributed with respect to dialect and sex within the remaining group of 600 speakers. A contingency table analysis determines there to be no significant effect of sex or dialect region on which sequence was produced. Nor is there any effect on whether the stop was released in the cases where it occurred before a glide. Lastly, there is no effect of sex or dialect on whether an affricate or a glide was produced.

CONCLUSION

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

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

Endnotes

[1]A preliminary report by the author on a subset of this research is to appear in *JASA* .

Acknowledgment

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

References

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

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

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

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

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

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

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