Preliminary results on speaker-dependent variation in the TIMIT database

Dani Byrd

Citation: The Journal of the Acoustical Society of America **92**, 593 (1992); doi: 10.1121/1.404271
View online: http://dx.doi.org/10.1121/1.404271
View Table of Contents: http://asa.scitation.org/toc/jas/92/1
Published by the Acoustical Society of America

Preliminary results on speaker-dependent variation in the TIMIT database

Dani Byrd
University of California, Department of Linguistics, Los Angeles, California 90024-1543

(Received 4 September 1991; accepted for publication 1 April 1992)

A set of phonetic studies based on analysis of the TIMIT speech database is presented. Using a database methodological approach, these studies detail new results in speaker-dependent variation due to sex and dialect region of the talker including effects on stop release frequency, speaking rate, vowel reduction, flapping, and the use of glottal stop. TIMIT was found to be fertile ground for gathering acoustic–phonetic knowledge having relevance to the phonetic classification and recognition goals for which TIMIT was designed, as well as to the linguist attempting to describe regularity and variability in the pronunciation of read English speech.

PACS numbers: 43.70.Gr, 43.70.Hs, 43.70.Fq

INTRODUCTION

The majority of acoustic–phonetic studies to data have shared, in the most general terms, a common methodology. Each speaker reads the 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 that 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 producing exaggerated differentiation of minimal differences. 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 in which small variations of 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.

The TIMIT database of American English, which is described below, was designed jointly by the Massachusetts Institute of Technology, Texas Instruments, and Speech Research Institute under the DARPA speech recognition effort as a training and testing ground for speech recognition systems (Lamel *et al.*, 1986). 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 that TIMIT serves.

Just as allophonic variation is dependent on phonological context, phonological or phonetic variation might be influenced by speaker-dependent factors such as age, race, socio-economic class, sex, education, and assorted other details ranging from health to mood. These effects tend to be more elusive for the researcher than variation caused by physiology, motor control, or acoustics. Consequently, the researcher has less intuition and experience regarding what pronunciation variability to look for and what speaker qualities to correlate it with. In what follows, I will describe a series of small studies which exploit TIMIT for the purpose of investigating the influence of the speaker-specific factors of sex and dialect on speaking.

I. 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" (Zue *et al.*, 1990, p. 352). Additionally, 450 phonetically compact sentences were designed to incorporate as complete a coverage of phonetic pairs as practical (Lamel *et al.*, 1986). 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 (Fisher *et al.*, 1986; in Zue *et al.*, 1990). 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. Broadly speaking, these include the seven geographical regions of New York City, the Western United States, New England, the northern Midwest, the southern Midwest,

 0001-4966/92/070593-04$00.80

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 digitized recordings are accompanied by a time aligned phonetic transcription. An orthographic transcription and the waveform are also provided. The validity of the results reported in this work depend entirely on the correctness and consistency of the phonetic transcriptions. A description of the inventory of transcribed elements and criteria for segmentation can be found in Zue and Seneff, 1988.

The first two-thirds or "training" portion of the database was an early release of the entire TIMIT database. (The final official release is now completed and available on CD.) The results reported here are based on this training data set. It includes 420 talkers saying 10 sentences each: two calibration sentences, five compact sentences, and three random sentences. This data set includes 31% female speakers and roughly even distribution across the eight dialect regions (although New England, New York City, and the unaffiliated group are slightly underrepresented). For each of the analyses described below a subset of the database was analyzed that satisfied criteria of size and context, as outlined in each section. The selected portion of the database and only this portion were considered for each analysis.

II. RESULTS AND DISCUSSION

A. Stop releases

One way in which the pronunciation of stop consonants varies is in the probability of a plosive release. A search of the random (1260 sentence readings) and compact (2100 sentence readings) sentences of the database was conducted to determine the frequency of stop releases in sentence-final position following each of the six oral stops closures [b], [p], [d], [t], [g], and [k]. (The calibration sentences were not included here as 420 repetitions of the word "*that*" would bias the results.) The search of sentence-final position found 726 sentence final stops of which 299 were released. Speaker-specific characteristics of (sentence-final) stop releases in the TIMIT corpus were investigated. Table I shows the number of released and unreleased stops for each dialect region and sex.

An ANOVA was conducted to determine if any effects of sex or dialect region on the presence of sentence-final stop releases existed. In this analysis, all sentence-final stops were coded as released or unreleased based on the TIMIT phonetic transcription, as being spoken by a talker from one of the eight TIMIT dialect regions, and as being spoken by a male or a female. In a two-factor ANOVA a marginal effect of dialect region on stop releases was found [$F(7,710) = 1.979$ $p = 0.0555$]. New England speakers released their stops the least often, 47% of the time, and the "Army Brat" group the most often, 69% of the time. The ANOVA also showed a significant effect of sex on whether the sentence-final stop was released [$F(1,710) = 7.111$ $p = 0.0078$]. Female speakers in the TIMIT database released sentence-final stops reliably more often than male speakers. There was no interaction of sex and dialect region in influencing sentence-final stop releases [$F(7,710) = 0.564, p = 0.7856$].[1]

B. Speaking rate

While the relationship of sex and speaking rate has been noted anecdotally, there are evidently few investigations of the comparative speaking rates of men and women (Henton, 1988; but see Zue and Laferriere, 1979). Data regarding the effect of dialect on speaking rate are equally scarce. The analysis below investigates whether sex or dialect influenced the rate at which sentences in the TIMIT corpus were read.

Because all speakers read the same two calibration sentences, the durations of these sentences were used to evaluate speaking rate. The two sentences are:

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

(2) Don't ask me to carry an oily rag like that.

Thus a total of 840 sentence readings were measured, two sentences for each of 420 speakers. (Note: silent periods digitized and encoded in the transcription were not included in the measurement.) The speaker-dependent variables of sex and dialect region were evaluated. Mean sentence durations and standard deviations for each dialect region and sex are shown in Table II.

A two-factor ANOVA was conducted to determine if sex and dialect region had any effect on speaking rate. A significant effect of dialect region on speaking rate was found [$F(7,824) = 2.082, p = 0.0431$]. This effect is more reliable for the first calibration sentence, [$F(7,404) = 2.395$, $p = 0.0207$] than for the second, [$F(7,404) = 1.273$, $p = 0.2622$], as shown by a two-factor ANOVA. This is not unexpected in light of the fact that sentence one, the longer sentence, had a greater mean duration, 3032 ms (s.d. = 420.74) than did the shorter sentence two, 2427 ms (s.d. = 327.67). Sex was also shown to have a very significant effect on speaking rate, [$F(1,824) = 15.169$, $p = 0.0001$], as shown by the two-factor ANOVA. Under these recording conditions, women speak reliably more slowly than do men. There was a marginal interaction of sex and dialect region in influencing speaking rate [$F(7,824) = 1.801, p = 0.0838$], with speakers from New York City and the "Army Brat" group going against the

TABLE I. Number of sentence-final oral stops released and not released in compact and random sentences (a total of 3360 sentence readings).

Category	Number released	Number unreleased
New England	28	31
Northern	64	45
North Midland	61	49
South Midland	63	41
Southern	60	54
New York City	40	24
Western United States	80	41
"Army Brat" (unaffiliated)	31	14
Female	147	74
Male	280	225

general trend of women speaking more slowly than men.

In order to determine whether education, and by extension reading fluency, could have created this asymmetry, the respective educational levels of male and female speakers in the database was considered. Here, 79% of the women have college degrees as compared to 88% of the men. While there is a slight difference, this probably was not the main factor causing the asymmetry in speaking rate; although effects of education on the speech in this corpus deserve further investigation.

It was also possible that the frequency or duration of pauses in the calibration sentences might be different for men and women thereby contributing to the rate effect. A chi-square test was conducted comparing the observed and expected number of pauses where the expected number of pauses for each sex was calculated by multiplying the total number of occurrences by the proportional representation of that sex in the database. This test showed no influence of sex on the frequency of occurrence of a pause [$\chi^2 = 0.457$, $p = 0.499$]. A two-factor ANOVA testing for effects of sex and dialect region on pause duration in the calibration sentences also found no effect nor interaction.

As the speaking rate results presented above are based on a fairly large corpus of data and show robust differences between men and women for this task, this serves as a contribution to attempts to specify sex differences in, at least, read American English. An account of why and under what circumstances this and other sex-dependent differences exist awaits further research considering such factors as the speaking rate of the interlocutor's instructions (Goldman-Eisler, 1968; C. Henton, personal communication) and sociological expectations in different settings (Labov, 1972).

C. Vowel reduction and sex

As speaking rate was determined to be affected by the sex of the speaker, and vowel reduction is known to be affected by the rate of speech, an analysis was undertaken to determine if vowel reduction differed between men and women speakers. The "random-type" sentences in the TIMIT database were examined as a corpus to determine if the frequency of occurrence of the reduced vowel [ə] would show a distributional pattern affected by the sex of the talker. This was motivated by the hypothesis that slower speakers would show less vowel reduction. The random sentence set included 1260 sentences each spoken by a single speaker. This corpus of data was chosen for analysis because it yielded a large number of schwas; 1410 instances were found. The sample produced in this search was considered large enough, and no additional searches were conducted. Although reduced vowels other than [ə] are transcribed in the database, this study proceeded under the working hypothesis that the frequency distribution of schwa would be representative of the tendency of speakers to reduce vowels. This approach additionally assumes that neither group (women or men) received a greater proportion of sentential contexts for vowel reduction. Considering the large sample size and the fact that sentences were created and distributed randomly, this appears to be a safe assumption.

TABLE II. Mean sentence durations and standard deviations of calibration sentences for each TIMIT dialect region and for sex.

Category	Mean sentence duration (ms)	s.d.
"Army Brat" (unaffiliated)	2589	416
North Midland	2682	397
Western	2689	467
New England	2714	530
Northern	2735	484
New York City	2736	420
South Midland	2767	498
Southern	2835	570
Female	2849	513
Male	2676	460

A chi-square test was conducted to compare the observed and expected frequencies of schwa. The expected number of schwas for each sex was calculated by multiplying the total number of occurrences by the proportional representation of that sex in the database. 401 schwas were pronounced by female speakers as compared to an expected 437.1, and 1009 by male speakers as compared to the expected 972.9. These results were significant [$\chi^2 = 4.321$, $p = 0.0376$] and support the hypothesis that the men, shown to be faster speakers than the women, tend to reduce their vowels (to [ə]) more often than the women in this corpus of speech. The possibility exists, however, that the two groups may employ different sets of reduced vowels. While outside the scope of this Letter, a more detailed study of the type and frequency of vowel reduction as a function of sex is necessary to clarify the above result.

We can note also that a chi-square test using these data was very significant [$\chi^2 = 35.003, p = 0.0001$] for the effect of dialect region on the frequency of schwa. However, in this small investigation, there is no way to ascertain whether this is because some dialect regions reduce their vowels more or whether certain dialect regions prefer another reduced vowel such as [ɨ]. While the latter explanation appears more likely, this result deserves further investigation.

D. Flapping

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).

A study on flapping was undertaken using TIMIT to determine if such a ballistic physiological realization of a phonemic [t] would be affected by the speaker-dependent variables of sex and dialect region. The readings of the first calibration sentence were used to examine the occurrences of flaps. This data set was chosen because it was read by all speakers and provides a controlled context for this highly context-sensitive process. This sentence also provides two distinct flapping environments, one word finally and one word medially, for evaluation. The sentence is: "She had your dark suit in greasy wash water all year." The potential flap sites are shown in boldface.

A total of 491 flaps occurred; 148 by females, and 343 by males. However, χ^2 values show that there is no significant difference between the observed distribution of flaps between males and females and the expected random distribution [$\chi^2 = 0.169, p = 0.6812$] (but see Zue and Laferriere, 1979). Additionally, a χ^2 test shows no effect of dialect region on flap frequency $\chi^2 = 1.881$, $p = 0.9661$]. In this TIMIT material, sex (and speaking rate) and dialect region have no significant effect on the frequency of oral flaps.

E. Glottal stop and sex

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, particularly across speakers. However, glottal stop is one of the transcribed phones in the TIMIT database. This corpus can provide a data set for determining general distributional patterns of glottal stop across a variety of prosodic and phonological contexts.

A brief examination of the occurrence of glottal stop was made using the "phonetically compact" sentences of the database. This data set was chosen because it includes a very complete coverage of possible phonetic sequences in English and because it produced a large sample; 1454 glottal stops were found in the 2100 sentence readings examined. This sample was considered large enough for analysis, and no further searches were conducted. Although I am aware of no evidence that the use of glottal stop is related to speaking rate, a chi-square test showed there to be an effect of the sex of the speaker on the frequency distribution of glottal stop. Recalling that under the null hypothesis there is no effect of the sex of the speaker, we would expect the number of glottal stops found for each sex to be a function only of the number of speakers of that sex. Here, 555 glottal stops were produced by women as compared to the expected 450.74, and 899 by men as compared to the expected 1003.26. This yields a highly significant chi-square value of 34.951 with a probability level of 0.0001. This result shows that women used glottal stops with significantly greater relative frequency than did the men in this data set. It is somewhat unexpected to find that this speaker-dependent characteristic was related to the use of glottal stop. 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.

A chi-square test assessing the effect of dialect region on the frequency distribution of glottal stop also showed a highly significant result [$\chi^2 = 328.129, p = 0.001$]. Specifically, New England and New York City speakers appear to use relatively more glottal stops than expected by a random distribution, and South Midland and Western speakers appear to have a relatively smaller number of glottal stops than expected.

III. CONCLUSION

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 effects 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" (1972, p. 303). While women represent only one-third of the speakers in this database, there is still a great deal of work which can be undertaken both in the spirit of the studies presented here and of a type more concerned with the specific detail of allophonic patterns.

In conclusion, TIMIT has proven to be fertile ground for gathering acoustic–phonetic knowledge that is of interest to the linguist attempting to describe English speech. It has been shown that speaker-specific characteristics of sex and dialect region influence stop release frequency, speaking rate, vowel reduction, and the frequency of glottal stop but do not affect the frequency of flaps in the TIMIT material analyzed. These studies and similar database approaches to linguistic investigation offer a promising new methodology for investigating questions of speech analysis.

ACKNOWLEDGMENT

This research was supported by the National Science Foundation and the Department of Linguistics at UCLA. The author wishes to thank Patricia Keating, Keith Johnson, James R. Glass, Ralph N. Ohde, A. W. F. Huggins, and James R. Byrd for insightful comments.

[1]A separate analysis of the sentence final [t] in the word "that" from the calibration sentence was conducted. An ANOVA showed there to be no effect of sex or dialect region on whether this [t] was realized as a glottal stop, an unreleased alveolar stop, or a released alveolar stop. It may be that frequent (i.e., common) words such as "that" will undergo less speaker-dependent variation than more uncommon words or simply that the environment of sentence-final [t]'s is not the environment where such variation occurs.

Fisher, W. M., Doddington, G. R., and Goudie-Marshall, K. M. (**1986**). "The DARPA speech recognition research database: specifications and status," Proc. DARPA Speech Recog. Workshop, 93–99.

Goldman-Eisler, F. (**1968**). *Psycholinguistics: Experiments in Spontaneous Speech* (Academic, New York).

Henton, C. (**1988**). "Vocal discord," UC Davis Magazine, 18–21 (December 1988).

Labov, W. (**1972**). *Sociolinguistic Patterns* (Univ. of Penn., Philadelphia).

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

Oshika, B., Zue, V. W., Weeks, R. V., Neu, H., Aurbach, J. (**1975**). "The role of phonological rules in speech understanding research," IEEE Trans. Acoust. Speech Sig. Proc. **ASSP-23**(1), 104–112.

Zue, V., and Laferriere, M. (**1979**). "Acoustic study of medial /t,d/ in American English," J. Acoust. Soc. Am. **66**, 1039–1050.

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

Zue, V., Seneff, S., and Glass, J. (**1990**). "Speech database development at MIT: TIMIT and beyond," Speech Commun. **9**, 351–356.