S2.18

THE DEMOGRAPHICS OF SPEAKER INDEPENDENT DIGIT RECOGNITION

Joseph Picone

Speech and Image Understanding Laboratory
Texas Instruments, Inc.
P.O. Box 655474, MS 238
Dallas, Texas 75265 USA

ABSTRACT

Though high performance speaker independent continuous digit recognition (SICDR) has recently been demonstrated in laboratory environments on databases consisting of several hundred speakers, robust speech recognition of large populations of speakers over telephone channels is still an elusive goal. In this paper, we introduce the Voice Across America (VAA) database, a database designed to provide a statistically significant model of the demographics of the continental U.S. population. Recognition performance on a simple digit recognition task is analyzed and shown to be most highly correlated with signal to noise ratio, dialect, and age. Several other demographic features, including sex, income level, education level, and market size, were found to have little correlation with recognition error rate.

I. INTRODUCTION

Texas Instruments' Voice Across America (VAA) program is designed to produce a demographically balanced database that exhaustively samples four important dimensions to the speaker independent speech recognition problem: (1) dialect variations; (2) handset and channel characteristics; (3) speaking style; and (4) lexicon. The goal of the database program is to collect a 100,000 speaker database over standard long distance telephone lines. Presently, over 3500 subjects have been recorded, and over 2000 sessions have been orthographically transcribed. In this paper, we describe the VAA database, and analyze the performance of a simple SICDR system [1-3] on a 1088 speaker subset of the database.

The VAA scenario provides an efficient and cost-effective methodology for collecting a large database containing a mixture of read and spontaneous speech material. The cornerstone of VAA is a commercially available database of 225,000 potential participants that has been demographically balanced to model the United States Census data [4]. This database was chosen primarily because of its increased coverage of minorities. The database is segmented into balanced subsets of 5,000 subjects, providing a convenient means to build a large database incrementally. An important feature of the VAA database is the availability of a complete demographic profile for each subject. One important usage of this data is to permit oversampling of the outgoing solicitation database to improve coverage of various dimensions of interest.

Potential subjects in VAA are solicited by mail. Included in this mailing is a unique session sheet describing the phrases a subject will be requested to speak. Subjects participate by calling a toll-free 800 number. The VAA data collection system is completely automated, and is available 24 hours a day. The entire session takes about 3 minutes. The incoming speech data is collected from an analog telephone line using a high quality telephone interface and a 16 bit A/D. The speech data is automatically excised and downsampled during collection, and saved as 8 kHz sampled data with a 0.5 second pad of "silence" on each end of the utterance.

In the first version of VAA, we studied a digit lexicon (including natural numbers). The prompts used to elicit speech data are given in Figure 1. Eleven of the fourteen utterances requested in a session represent numeric information. The read strings were selected to reflect common segmentations of digit strings. Spontaneous speech is also solicited in the form of familiar numeric information, such as telephone numbers. The session is designed to balance the amount of spontaneous and read speech. One read sentence from the TIMIT Acoustic Phonetic Database [5] has been included in each session.

Are you ready?
Please say your 7 digit Panel ID#.
Please say a zipcode.
Please say a familiar phone number.
Please say the highlighted price.
Please say the highlighted phone number.
Please say the highlighted driver's license number.
Please say the highlighted serial number.
Please say the highlighted phone number.
Please say the highlighted catalog number.
Please say another familiar phone number.
What is a typical cost for the listed item?
Please say the highlighted sentence.
Was this session difficult for you?

Figure 1. The Phase I Voice Across America Prompts.

CH2847-2/90/0000-0105 $1.00 © 1990 IEEE

II. PHASE I VAA DATABASE

A 5,000 subject cell was solicited in the first phase of this program. The overall response rate was 36%, with females responding at a ratio of 2 to 1 to males. The mean response time was 3.75 days. 90% of the responses occur within 2 weeks of the mail date. The busiest hour of the day occurs at 3 PM CST, and 85% of the calls are handled between 8 AM and 8 PM. Thus, having dedicated two audio systems to data collection, we are able to solicit at least 1500 subjects every three days and still maintain a negligibly small probability of a caller receiving a busy signal.

A large portion of the effort required in VAA actually occurs after a subject's session is complete. The data from each session is manually validated and orthographically transcribed in as detailed a fashion as possible. In Figure 2, a lexical analysis of the Phase I database is given. In addition to the orthographic transcription, subjective assessments of various speaker and channel attributes, such as dialect, signal and channel qualities, and style of speech, are made for each utterance. Qualitative judgements are generally tree-structured, permitting fine distinctions to be made only when appropriate. For example, a speaker can be classified generally as a Southern dialect, or, if appropriate, a Deep South or South Midland dialect.

A. Responses to Yes/No Questions

210 Unique Responses
81% Responded "YES" or "NO"
97% Contained "YES" or "NO"
281 Word Vocabulary (Entropy = 13.1 Symbols)

B. Acoustic Phonetic Sentences

1239 Unique Sentences (Only 36 Appear 3 Times)
4029 Unique Words (Avg. of 12.2 Words/Sent.)
Word Entropy = 970 Symbols
Word Pair Entropy = 8600 Symbols

C. Prompts Requesting Numeric Information

397 Unique Words (Entropy = 17.6 Symbols)
87% of Words are Basic Digits (One-Nine, Zero, Oh)
5% Natural Numbers (25 words)
Remaining 8% Constitute 350 Words
~10% of Read Strings Include Natural Numbers
~15% of Spont. Strings Include Natural Numbers.

D. Comparison Of Read and Spont. Phone Numbers

Strings Including	Read	Spont.
Digits Only	60%	85%
Phrase "Area Code"	30%	10%
Common Words "Number", "Dash", or "Is"	10%	2%
Natural Numbers	3%	4%
Other	3%	1%

Figure 2. A Lexical Summary of the Phase I Database

A summary of the overall signal to noise (SNR) ratio is shown in Figure 3(a), along with a breakdown of SNR by geographic region in Figure 3(b). Note that the overall SNR of the database is surprisingly good. There is a minor correlation of SNR with the geographic distance of the originating phone call from Dallas. Most likely this is related to long distance call routing patterns.

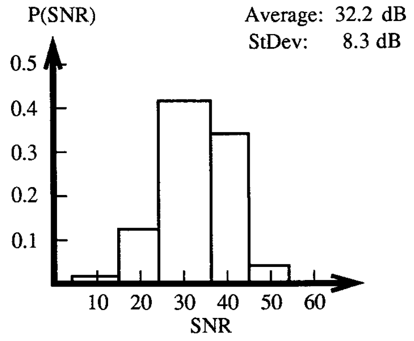

Figure 3(a). VAA SNR Statistics

Census Region	SNR
New England	31.9
Middle Atlantic	33.4
East North Central	33.8
West North Central	34.5
South Atlantic	34.4
East South Central	34.5
West South Central	35.0
Mountain	28.1
Pacific	30.6

Figure 3(b). SNR Versus Census Region

A subset of the database was constructed that consisted of every speaker from whom two read and two spontaneous digit strings could be found that contained only the vocabulary one, two, ..., zero, and oh. In instances where telephone numbers could not be found, other items were substituted as appropriate. This selection procedure resulted in a database of 1188 speakers. For the remainder of this paper, we will consider this subset of the database.

III. TRAINING DATABASE CONSTRUCTION

In order to investigate the importance of demographic factors on such a SMALL database, we must maximize the size of the testing database. Initially the database was split in half, yielding a test database of 594 speakers. Performance on the 594 speaker test set was evaluated versus both the type and style of training data. Two parameters were varied simultaneously, the number of training speakers, and the amount of training data per speaker. The tokens were ordered such that the two read strings were placed first, and the two spontaneous strings last. We see from Figure 4 that performance saturates

Figure 4. On the Importance of Large Training Databases (Performance Quoted as Percent Word Errors.)

No. Speakers	25	50	100	200	400	594
No. Tokens/Spk.						
1	12.2	9.3	7.6	6.9	6.7	6.4
2	9.5	8.0	7.4	6.7	6.6	6.4
3	9.3	8.3	7.5	6.6	6.6	6.3
4	9.1	8.2	7.0	6.6	6.4	6.3

somewhere in the neighborhood of 100 speakers. We also see that for SICDR, performance improves more rapidly by adding speakers rather than by adding more data from each speaker.

Next, we estimate the variance in the error rate as a function of speaker set, as shown in Figure 5. The first four training sets, denoted t1 through t4, were constructed by randomly sampling the 594 speaker training database. The set denoted "General American" was constructed by choosing speakers that had been classified as having a General American dialect. There is the least acoustic variability within this group. In the last set, denoted "Maximize Dialects", speakers were chosen to maximize each dialect's representation in the training database (due to the limited size of the database, all dialects could not be equally represented). Figure 5 demonstrates the importance of achieving a good dialect coverage in the training database. Random Set #4, though not containing equal dialects, contained a better sampling of dialects than any of the other random sets.

Training Set	Performance
Random Set #1	7.1%
Random Set #2	7.2%
Random Set #3	7.3%
Random Set #4	6.7%
General American	7.7%
Maximize Dialects	6.8%

Figure 5. On the Importance of Dialect Coverage

Thus, the 100 speakers (equal numbers of males and females) "Maximize Dialect" training set was fixed as the training database. The remaining 1088 speakers were used in the test database. A summary of the demographic profile of the test database is given in Figure 6. There are actually 48 demographic fields in the database. In Figure 6, we show only the most significant data. Dialect categories are based only on judgements made from the acoustic data, and therefore are not available for the solicitation database.

IV. BASELINE EXPERIMENTS

The "Maximize Dialect" training set constitutes a good starting point for analyzing recognition performance on the test set. We have conducted some preliminary experiments on the VAA database using a simple continuous density HMM continuous digit recognizer. Throughout this analysis, we will consider the average word error rate per speaker as a measure of performance for each speaker. A histogram of word error rates indicated that for a third of the speakers in the database, the error rate was zero, while for 25% of the speakers in the database, the word error rate was over 9% (mean + 0.5*stdev). Below, we investigate the nature of these errors.

DATABASE:	Solicitations	Test DB
SEX:		
Male	50%	36.3%
Female	50%	63.7%
AGE:		
18-24	15.6%	10.3%
25-34	24.4%	21.4%
35-49	25.3%	27.8%
50-64	18.7%	23.0%
65+	16.0%	17.6%
CENSUS REGION:		
New England	4.9%	5.9%
Middle Atlantic	16.2%	17.5%
East North Central	16.3%	15.5%
West North Central	8.3%	7.7%
South Atlantic	15.9%	16.3%
East South Central	7.8%	5.9%
West South Central	10.6%	9.0%
Mountain	6.1%	7.2%
Pacific	14.0%	15.1%
HIGHEST EDUCATIONAL LEVEL:		
Unknown	2.8	1.7%
Grade School	3.9	3.4%
Some High School	17.7	14.2%
Graduated High School	35.9	32.1%
Some College	17.2	18.7%
Graduated College	13.8	17.8%
Post College Graduate	8.6	12.0%
DIALECT:		
Northern		0.6%
Black		0.2%
General American		65.2%
American		17.4%
New England		1.7%
British		0.2%
South Midland		4.4%
Southern		8.5%
Foreign		0.4%
New York		0.9%
Deep South		0.5%

Figure 6. A Comparison of the Demographic Profile of the VAA Database to the U.S. Census Targets.

The most significant influence on performance is signal to noise ratio. An overall analysis of the recognition performance versus SNR is shown in Figure 7.

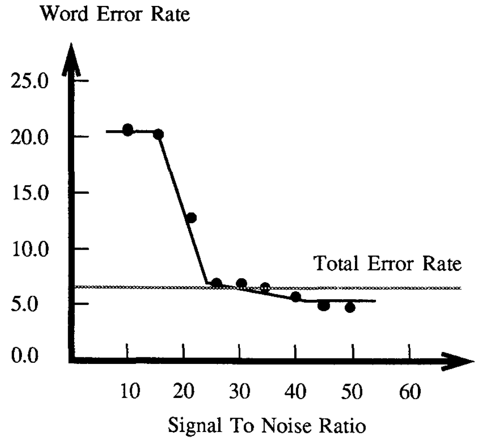

Figure 7. Error Rate As A Function of SNR

The next most significant factor in performance is dialect. To isolate degradations in performance due to dialect, we consider the subset of the database for which SNR is greater than 25 dB. In Figure 8, error rates as a function of dialect are tabulated. The average word error rate is given, along with the number of speakers used in computing the error rate. Note that even a 1,000 speaker database does not seem large when analyzing microscopic characteristics of the speakers. From Figure 8, however, we can conclude that performance is consistently worse for dialects acoustically distinct from General American. For example, we have historically noted poorer performance for Southern dialects [1], and this is verified in Figure 8. It is interesting to observe that the American dialect, which was used to group all speakers who displayed conflicting dialect cues, but were definitely not General American, also displays a somewhat higher error rate than the General American dialect.

Dialect	Average Speaker Word Error Rate	
General American	5.1%	(591)
American	6.5%	(164)
Southern	7.1%	(80)
South Midland	9.2%	(47)
New England	9.3%	(16)
New York City	9.4%	(9)
Northern	4.2%	(6)
Deep South	19.9%	(5)
Foreign	14.6%	(5)
Black	5.9%	(2)

Figure 8. The Sensitivity of Performance to Dialect

The error rates for particular dialect sensitive words did not correlate well with the cumulative results in Figure 8. For Southern dialects, it is interesting to note that the word error rates are highest for the words "four" and "five". In the American dialect, the word "three" has the most errors, followed by "seven".

Next, we attempted to isolate the effects on performance of a speaker's age by examining all speakers whose session SNR was greater than 25.0 dB and whose dialect was General American. This is tabulated in Figure 9. We see that performance is worse for the 18-24 age group and the 65+ age group, and comparable for all other age groups.

Age	Performance	
18-24	6.9%	(66)
25-34	4.8%	(137)
35-49	4.7%	(172)
50-64	4.8%	(134)
65+	5.5%	(88)

Figure 9. The Sensitivity of Performance to Speaker Age.

Finally, we examined performance as a function of four other demographic variables of interest, speaker sex, income, education, and proximity to a major metropolitan area. No significant correlates were observed. The last category represents an attempt to localize "pure" examples of dialects, based on the underlying assumption that the more rural the area, the more likely one is to find uncontaminated examples of dialect. We speculate that the database, at this point, is too small to support analyses of these second order effects.

V. SUMMARY

We have only scratched the surface of the types of analyses that can be supported by the VAA database. We have shown SNR, dialect, and age to be important dimensions of recognition performance. Until the VAA database approaches its goal of 100,000 speakers, measuring statistically significant correlations of recognition performance with other demographic factors will be difficult. We also expect the sensitivity of performance to these factors to be far greater on more difficult recognition tasks.

REFERENCES

[1] J. Picone, G.R. Doddington, and J.J. Godfrey, "A Layered Grammar Approach to Speaker Independent Speech Recognition", presented at the 1988 IEEE Speech Recognition Workshop at Harriman, NY, June 1988.

[2] J. Picone, "On Modeling Duration In Context", in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 421-424, Glasgow, Scotland, May 1989.

[3] J. Picone, "Duration In Context Clustering For Speech Recognition", submitted to *Speech Communication*, July 1989.

[4] "Home Testing Institute Mail Panel Balancing Position Paper", available from Home Testing Institute, 1300 West Higgins Road, Park Ridge, Illinois, November 1985.

[5] W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status", in *Proceedings of DARPA Speech Recognition Workshop*, pp. 93-99, February 1986.