CONTINUOUS RECOGNITION OF LARGE-VOCABULARY TELEPHONE-QUALITY SPEECH

Pedro J. Moreno, Matthew A. Siegler, Uday Jain, Richard M. Stern

Department of Electrical and Computer Engineering
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

The problem of speech recognition over telephone lines is growing in importance, as many near-term applications of spoken-language processing are likely to involve telephone speech. This paper describes recent efforts by the CMU speech group to improve the recognition accuracy of telephone-channel speech, particularly in the context of the 1994 ARPA common Hub 2 evaluation of speech over long-distance telephone lines. The greatest amount of work was directed toward determining a training procedure that provides the greatest recognition accuracy when the incoming speech is known to be collected over the telephone. We compare the effectiveness of three training procedures, finding that training using high-quality speech that is bandlimited to 8 kHz can achieve results that are as good as those obtained by training on speech of a similar bandwidth collected over actual telephone channels.

We also compare the recognition accuracy of the SPHINX-II system using high-quality speech and telephone speech, and we comment on the reasons for differences in system performance.

1. INTRODUCTION

Considerable progress has been made in the field of large vocabulary speech recognition in recent years. However, good recognition accuracy is far more difficult to achieve when the incoming speech has been passed through a telephone channel, compared to the recognition accuracy that is obtained using high-quality speech. For example, a compact implementation of SPHINX-II achieved a an error rate of 8.8% on non-telephone environments in the evaluation set of the 1992 common ARPA 5000-word CSR evaluations using secondary microphones, while the best error rate obtained for speech from the same database that was passed through a local telephone loop was 19.5% [5].

The telephone network is known to introduce a great deal of inherent environmental variability. Every telephone results in different channel and noise conditions. Nonlinear distortion and the use of different telephone microphones further complicate the problem. The development of training and compensation procedures to produce greatest recognition accuracy for telephone speech remains an active field of research.

In this paper we describe and compare the performance of a series of training and compensation procedures that were developed to improve the recognition accuracy of the CMU SPHINX-II speech recognition system in telephone speech environments. All experiments were performed using the development and evaluation set of the 1994 ARPA CSR H2 speech corpus. This corpus consists of unlimited-vocabulary speech read over long-distance telephone lines. The speech was collected directly from a digital T1 line (using 8-bit μ-law compression) with no constraints on the type of handset and microphone used by the subject.

In Section 2 we describe in detail the training procedures that were explored. In Section 3 we briefly describe environmental compensation techniques and their performance. Finally, in Section 4 we compare the performance of our system using telephone speech (from the H2 task) to the performance obtained using high-quality speech (from the H1 task), and we discuss possible reasons for the increase in error rate observed using telephone speech.

2. TRAINING STRATEGIES

Conventional wisdom has suggested that in order to maximize speech recognition performance, the training and testing sets used should be as similar as possible [9]. This implies that the best possible training data for a telephone speech recognizer telephone speech would also be telephone-quality speech. However, training with telephone speech has several problems including (1) a smaller amount of speech available for training, (2) greater effort is required to “clean” and post-process telephone databases, and (3) training procedures more slowly when lower-quality telephone speech is used as the input.

In this section we compare the performance of several different alternatives for training HMMs for use with telephone speech. We evaluate the training procedures by comparing word error rates obtained using 160 utterances from the ARPA CSR H2 development set. The gender of each speaker in this set is presumed to be known *a priori*. A dictionary of 58,000 words was used, and a language model covering 14.1 million bigrams and 17.9 million trigrams. This model is identical to the one used by CMU for the 1994 ARPA H1 CSR test. We compare the use of three procedures: training using simulated telephone speech, training using real telephone speech, and training using filtered high-quality “clean” speech.

2.1. Training using simulated telephone speech

The use of hardware or software telephone-channel simulators is one possible solution to the problem of the limited availability of telephone databases. Simulated databases are generally obtained

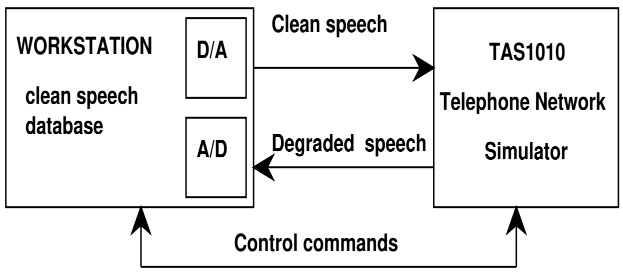

Figure 1. Hardware setup used for the recording of simulated telephone speech databases.

by passing clean speech through the simulator [*e.g.* 6]. Different degradations can be activated in the simulator allowing a wide range of telephone conditions. Figure 1 summarizes our recording apparatus. Simulated "stereo" databases (which contain both the original clean speech and the corresponding simulated telephone speech) can be produced for possible use in environmental compensation techniques.

We passed 7000 sentences from the WSJ0 training database through a hardware simulator, the TAS 1010 [8], Thirty different telephone-channel conditions were simulated with an average of 200 sentences per condition to generate female and male sets of HMMs.

The performance of these models on the development set was quite poor, producing error rates of up to 30.0%. A possible reason for this poor performance is the failure of telephone simulators to accurately capture all of the significant degradations introduced by real telephone channels, including microphone variability.

2.2. Training using real telephone speech

Our second approach was to train the system using real telephone speech collected over the network. We used 10,734 TIMIT and WSJ utterances from the training component of the Macrophone corpus [2]. This subset had no disfluencies or background noises. The database was split on the basis of gender, and the standard SPHINX-II training procedure was used to produce male and female set s of HMMs.

In a separate experiment the HMMs were reformulated to estimate the variance of the power coefficients from the training data. Previous versions of the SPHINX-II system had used a fixed default value.

We also explored the possible relationship between signal-to-noise ratio (SNR) and recognition accuracy. The training set was divided into three parts according to averaged sentence signal to noise ratio (SNR) and gender identity. Figure 2 shows a histogram of SNRs of utterances from the Macrophone training set. Based on this histogram the training set was divided into three subsets based on SNR: utterances with SNRs below 17 dB, utterances with SNRs above 15 dB, and the remaining utterances. As before, separate models were developed for males and females. This was done because it was assumed that low-SNR sentences are normally very noisy and would damage the training of HMMs. This procedure allows for the creation of HMMs for sentences with low SNR.

The performance of these three sets of HMMs on the development set was quite similar. A 21.5%error rate was observed using both male-female models and the SNR-dependent male-female models.

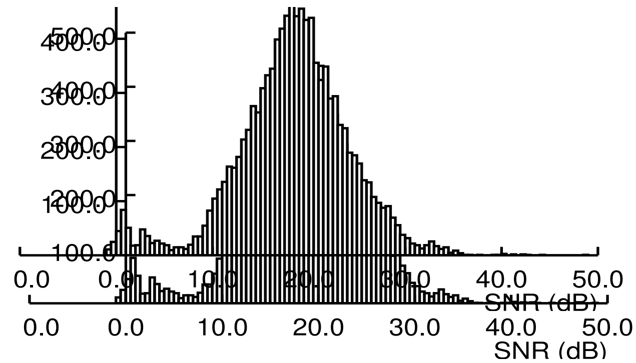

Figure 2. Histograms of SNRs of utterances from the Macrophone training database.

The addition of power variance modelling to the male-female set improved the performance very slightly to 21.2%.

2.3. Training using filtered clean speech

For our third training procedure we used high-quality speech recorded using a close-talking microphone that was downsampled to 8 kHz and passed through a linear filter to approximate the long-term average spectral shape of telephone speech. The magnitude of the equalization filter was determined by computing the ratio of averaged power spectra of downsampled WSJ0 and WSJ1 utterances and averaged power spectra of utterances from the Macrophone corpus. The magnitude of this filter, shown in Fig. 3, corresponds very closely to the average telephone channel response reported in previous studies of the telephone network (*e.g.* [3])

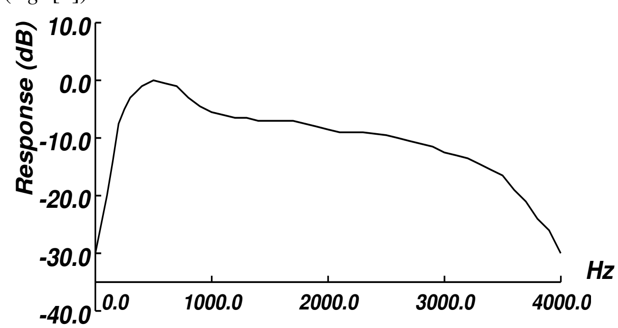

Figure 3. Equalization filter applied to WSJ0 and WSJ1 utterances to approximate the power spectrum of telephone speech.

The WSJ0 and WSJ1 training sets were chosen as the training corpus to produce a gender dependent set of HMMs. Each utterance was downsampled and filtered with the optimal equalization filter. HMMs were produced using the phone-dependent modelling approach, where triphones are tied together to a set of Gaussians based on preset phonetic classes [4]. This is a first-order approximation to continuous HMM modelling. Two sets of male-female models were produced: one that used only one generic phone class (resulting in a standard semi-continuous HMM), and a second set using 27 phone classes.

In an additional experiment, the training corpus was processed with the noise-robustness algorithm N-CDCN (N-CDCN, [7]) after undergoing the linear filtering for telephone-channel equalization. N-CDCN, an extension of the CDCN algorithm introduced by Acero [1], is a noise compensation algorithm based on maximum likelihood techniques that do not require simultaneously-recorded "stereo" data to achieve noise and channel compensation. Compensation is achieved on a sentence-by-sentence basis. Preprocessing the training data by N-CDCN accomplishes the goals of producing HMMs that uncorrupted by noise for the high-SNR utterances, and to get more accurate models for the low-SNR utterances.

One set of gender-dependent models was produced with 27 global phone classes using N-CDCN, using proper the power-variance modelling. In evaluations using this training procedure, the testing utterances were also processed by N-CDCN.

Using these procedures, we obtained an error rate of 21.3% training on utterances from the WSJ0 and WSJ1 training sets using a single phone class. The error rate dropped insignificantly to 21.2% when the number of phone classes was increased to 27, and it dropped again to 20.2% when N-CDCN and explicit modelling of power variances is added.

From the results reported in Sections 2.2 and 2.3 one might conclude that the performance of HMMs trained on telephone speech and HMMs trained on clean speech is similar. However, this comparison is not completely valid because the sizes of the two databases are quite different. Specifically, the Macrophone set has only 10,000 sentences compared to about 30,000 sentences used in WSJ0+WSJ1.

3. COMPENSATION TECHNIQUES

A new algorithm, the Multivariate Gaussian Based Cepstral Normalization (RATZ) technique [7], was used for environmental compensation. RATZ requires stereo data to produce compensation statistics. These stereo sentences represent a simultaneous recording of a clean utterance and its telephone counterpart. Since this is not easy to construct, the simulated stereo telephone database previously described in Sec. 2.1 was used for this purpose.

The RATZ compensation procedure attempts to estimate the unobservable clean speech utterance given a telephone-speech utterance and the empirically-derived relation between clean and telephone speech statistics. The estimated "clean" sentence is recognized using clean speech HMMs.Results achieved with this technique were disappointing as no improvement was observed. We suspect that the use of a simulated telephone speech database to formulate the stereo pairs instead of real telephone speech was the cause of this poor performance.

4. PERFORMANCE USING THE 1994 CSR EVALUATION DATA

We summarize in this section the results of experiments using the 1994 ARPA CSR H2 evaluation test set, including official results, and further refinements. We draw comparisons between H2 and H1-P0 for this year and similar results from H1-P0 and Spoke% from last year.

4.1. H2 Evaluation Results

The H2 evaluation system for H2 first performs *a priori* gender classification using a codebook distortion technique. Then, the utterances are processed using N-CDCN. The system was trained according to the procedures outlined in Sec. 2.3, passing utterances through N-CDCN, using 27 phone-dependent classes. A 58,000-word lexicon was used.

The average word error rate was 23.5% using NIST alignment tools, 20.6% for females and 26.6% for males. The ratio of word error rates observed for telephone speech and clean speech in Spoke 5 of the 1993 CSR evaluation was 2.2 (for a 5000-word task). This year the same ratio is 2.0, showing consistent behavior.

4.2. Effect of long silences

Over 25 percent of the frames in the H2 evaluation set are composed of non-speech events occurring before and after the sentence. Many insertion errors occurred at the end of some of these sentences during background noises such as paper-rustling and breath noises. To combat this problem, we used NIST routines to strip the silence periods at the beginning and end of each utterance.

Recognition was repeated using the same system configuration as in the official evaluation. The average word rate decreased to 22.2% overall, 24.4% for males and 20.3% for females.

4.3. Effect Out-of-Vocabulary Words

Table 1 compares the out-of-vocabulary (OOV) rate and error rates for each speaker in the H2 evaluation.

Speaker (female)	Error (%)	OOV (%)	Speaker (male)	Error (%)	OOV (%)
510	17.8	0.2	512	**31.7**	0.3
511	12.0	0.0	514	21.5	0.0
513	**38.5**	**8.2**	515	**35.7**	**1.1**
516	16.4	0.2	518	**26.8**	**1.8**
517	17.0	0.2	51b	16.0	0.3
519	11.6	0.3	51d	23.0	**1.8**
51a	**26.0**	**2.1**	51g	**37.0**	0.2
51c	20.1	0.0	51h	12.0	0.5
51e	**34.6**	**1.9**	51i	**45.3**	0.7
51f	19.7	0.3	51j	12.9	0.7
Sum Avg.	20.6	1.0	Sum Avg.	26.6	0.8

Table 1: Word error rate and out-of-vocabulary rate for speakers reported in the official CSR 1994 H2 evaluation set. Boldface entries indicate significantly high values. The overall error rate was 23.5% and the OOV rate was 0.9%.

The mean OOV rate for our system in the 1994 CSR H2 evaluation set was 0.9%, while for H1-P0 the OOV rate was 0.5%. It can be seen that the OOV rate has a significant impact on recognition. For example, Speaker 513 has an OOV rate of 8.2% and a word error rate of 38.5% In addition, the average word error rate for all speakers having an OOV rate greater than 1% was 30.1%; for those with an OOV rate less than 1%, the error rate was 20.8%.

4.4. Effect of Coverage by the Language Model

One method of measuring the modeling ability of a trigram language model is the *trigram-hit ratio* [8]. The trigram-hit ratio is a computation of the fraction of trigrams in the test set that occur in the language model. When this value is small, the language model is forced to back off to bigram or even unigram probabilities. As a result, a sentence with a low trigram-hit ratio will tend to be less well recognized.

Table 2 compares the trigram-hit ratio for each speaker in the H2 evaluation set. Although the overall trigram-hit ratio is a reasonable 66%, several speakers have a particularly low value. This can be caused by a high OOV rate, or by an atypical mode of speech, such as the use of the first or second person. The average word error rate for speakers with a trigram hit ratio below the mean is 30.0%, while it is only 17.2% for speakers with a trigram hit ratio above the mean.

Speaker (female)	Err. (%)	Trigram Hit Ratio (%)	Speaker (male)	Err. (%)	Trig Hit Ratio (%)
510	17.8	64	512	**31.7**	**54**
511	12.0	76	514	21.5	**59**
513	**38.5**	**54**	515	**35.7**	63
516	16.4	**56**	518	**26.8**	75
517	17.0	80	51b	16.0	76
519	11.6	73	51d	23.0	71
51a	**26.0**	**57**	51g	**37.0**	62
51c	20.1	72	51h	12.0	74
51e	**34.6**	**59**	51i	**45.3**	**58**
51f	19.7	64	51j	12.9	68

Table 2: Word error rates and trigram-hit ratios for speakers in the 1994 H2 evaluation set. Boldface entries indicate significant values. The overall error rate is 23.5% and the trigram-hit ratio was 66%.

4.5. Effect of Speaking Rate

It has been shown that when the phone rate exceeds one standard deviation from the mean, word error rate increases significantly (*e.g.* [9]). It has been shown that the mean vowel rate for read speech in the CSR H1 corpus is approximately 16 vowels/second, with a standard deviation of 2 vowels/second. The observed vowel rate for the CSR H2 evaluation set is approximately 16.3 vowels/second with a standard deviation of 2.3 vowels/second. One particular speaker, 512, had a vowel rate of approximately 20 vowels/second and an error rate of 31.7%.The average word error rate for the fastest four speakers is 30%.

5. SUMMARY AND CONCLUSIONS

In this paper we describe a number of procedures that have been employed to explore the problem of large-vocabulary recognition of telephone-quality speech. We compared the recognition accuracy of several training procedures and we obtained best performance by training on high-quality 8-kHz-bandwidth speech. The error rate of the SPHINX-II system on speech over long-distance telephone lines in the 1994 H2 evaluation is 23.5%. We suggest several sources for this error rate including long non-speech segments, gaps in coverage by the language model, variability in the acoustics of the training set, and a higher out-of-vocabulary word rate in some cases.

ACKNOWLEDGMENTS

This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. We thank Evandro Gouvêa for his help on error analysis. We also thank Bhiksha Raj for his help on this research and Ravishankar Mosur and the rest of the speech group for their contributions to this work.

REFERENCES

1. Acero, A., *Acoustical and Environmental Robustness in Automatic SPeech Recognition*, Kluwer Academic Publishers, Boston, MA, 1993.
2. Bernstein, J. and Taussig, K.,"Macrophone: An American English Telephone Speech Corpus for the Polyphone Project". *ICASSP-94*, May 1994.
3. Carey, M. B., Chen, H. T., Descloux, A., Ingle, J. F., and Park, K. I., "1982/83 End Office Connection Study: Analog Voice and Voiceband Data Transmission Performance Characterization of the Public Switched Network", *AT&T BLTJ*, **63**, Nov. 1984.
4. Hwang, M.-Y., *Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition*, Ph.D. Thesis, Carnegie Mellon University, 1993.
5. Liu, F.-H., "*Environmental Adaptation for Robust Speech Recognition*", Ph.D. Thesis, Carnegie Mellon University, 1994.
6. Moreno, P. J. and Stern, R.M., "Source of Degradation of Speech Recognition in the Telephone Network". *ICASSP-94*, May 1994.
7. Moreno, P. J., Raj, B., Gouvêa, E. and Stern, R.M., "Multivariate Gaussian Based Cepstral Normalization for Robust Speech Recognition". *ICASSP-95*, Detroit, May 1995.
8. Rosenfeld, R., personal communication.
9. Siegler, M. A., and Stern, R. M., "On the Effects of Speech Rate in Large-Vocabulary Continuous Speech Recognition Systems, to appear in ICASSP-95.
10. TAS1010 Voiceband Channel Simulators Operations Manual. Telecom Analysis Systems, 1989.
11. Weintraub, M. and Neumeyer, L.,"Constructing Telephone Acoustic Models from a High-Quality Speech Corpus". *ICASSP-94*, May 1994.