From the Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Adelaide, Australia, Vol. I., pp. 109 - 112, April, 1994.

SOURCES OF DEGRADATION OF SPEECH RECOGNITION IN THE TELEPHONE NETWORK

Pedro J. Moreno and Richard M. Stern
Department of Electrical and Computer Engineering
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

ABSTRACT

In this paper we compare speech recognition accuracy for high-quality speech recorded under controlled conditions with speech as it appears over long-distance telephone lines. In addition to comparing recognition accuracy, we use telephone-channel simulation to identify the sources of degradation of speech over telephone lines that have the greatest impact on speech recognition accuracy. We first compare the performance of the CMU SPHINX-I system on the TIMIT and NTIMIT databases [3,8]. We found that other factors beyond a mere decrease in bandwidth cause the observed degradation in recognition accuracy, and that the environmental compensation algorithms RASTA [6] and CDCN [1] fail to compensate completely for degradations introduced by the telephone network. In the second part of this paper we attempt to identify the most problematic telephone-channel impairments using a commercial telephone channel simulator and the SPHINX-II system. Of the various effects considered, additive noise and linear filtering appear to have the greatest impact on recognition accuracy. Finally, we examined the performance of three cepstral compensation algorithms in the presence of the most damaging conditions. We found the compensation algorithms to be effective except for the worst 1% of the telephone channels.

1. INTRODUCTION

As speech recognition systems become more accurate and sophisticated, interest in telephone-based applications for them increases. It is well known that recognition accuracy percentages are lower for speech over the telephone network than for speech that is carefully recorded in a quiet environment. Nevertheless, the reasons for this degradation in recognition accuracy are not well understood. In this paper we describe a series of studies that attempt to compare the effects of several putative causes of telephone-channel degradation, to determine which of these impairments have the greatest impact on recognition accuracy.

2. RECOGNITION ACCURACY USING THE TIMIT AND NTIMIT DATABASES

The NTIMIT database [8] has enabled researchers to perform controlled experiments that compare phonetic recognition accuracy obtained with high-quality speech with the accuracy for the same speech transmitted over long distance telephone lines. The TIMIT database is a continuous, speaker independent, phonetically-balanced and phonetically-labelled speech database. The NTIMIT database was created by transmitting sentences in the TIMIT database over telephone lines. Previous work on speech recognition systems has demonstrated that the use of speech over the telephone line increases the rate of recognition errors. For example Chigier [3] reports a reduction in accuracy of about 10%. In our work [11] we have also confirmed this result.

In Table 1 we compare phoneme recognition accuracy obtained using the TIMIT and NTIMIT databases. A version of SPHINX-I was used that was not optimized for phonetic recognition. We observe an absolute decrease in recognition accuracy of 11.4% as the TIMIT database is replaced by the NTIMIT database.

TRAIN	TEST	% ERROR
TIMIT	TIMIT	47.3
NTIMIT	NTIMIT	58.7
TIMIT	NTIMIT	68.7

Table 1: TIMIT and NTIMIT recognition results.

The initial set of subsequent experiments was designed to explore the hypothesis that this gap in recognition accuracy is caused by the more limited bandwidth of telephone speech. Specifically, Table 2 summarizes the recognition accuracy obtained by training and testing on (1) the original TIMIT database, (2) a version of the TIMIT database that is downsampled to 8 kHz (producing an effective bandwidth of 4 kHz), (3) a version of the TIMIT database that is both downsampled to 8 kHz and passed through a linear filter designed to approximate the frequency response of a typical telephone channel [2], and (4) speech from the NTIMIT database.

Database	Signal Bandwidth	% ERROR
Original TIMIT	0-8000 Hz	47.3
Downsampled TIMIT	0-4000 Hz	48.2
Filtered TIMIT	250-3400 Hz	50.3
Downsampled NTIMIT	250-3400 Hz	57.5

Table 2: Effect of bandwidth reduction on TIMIT.

We note that downsampling and filtering TIMIT only reduces recognition accuracy by 3.0%, while switching to the speech from the NTIMIT database reduces accuracy by an additional 7.2%. This suggests to us that bandwidth limitations and other linear filtering effects are not the sole source of the observed degradation in performance. Other possible source of degradation include the additive noise encountered in telephone channels and nonlinear attributes of the gain and phase response of the telephone channel. The RASTA [6] and CDCN [1] procedures are two examples of algorithms that have been developed to compensate for the effects of linear filtering and additive noise in office environments. Figure 1 summarizes recognition accuracy obtained on the TIMIT/NTIMIT databases using the implementation of RASTA that was part of the 1991 SRI ARPA speech recognition system [12] and the original implementation of the CDCN algorithm. The RASTA

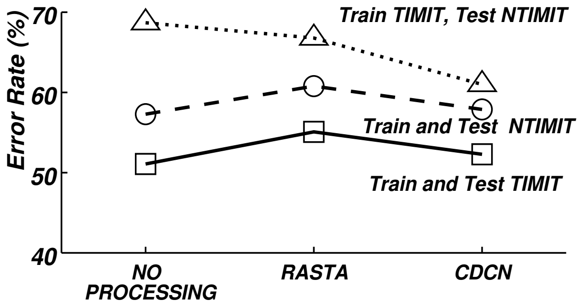

Figure 1: Effect of RASTA and CDCN on speech recognition error in the TIMIT and NTIMIT databases.

algorithm, which in this implementation compensates primarily for the effects of linear filtering, improves accuracy only in the "cross-condition" case, when the system is trained in one environment and tested in the other. CDCN provides greater improvement in the cross-condition case, but it does not improve recognition accuracy when the system is trained and tested in the same environment.

3. RECOGNITION ACCURACY USING TELEPHONE-CHANNEL SIMULATIONS

Extensive experimental and statistical analyses have been used to characterize some of the degradations encountered in the telephone network [2]. Among others, telephone networks produce the following impairments:

- Added low-frequency tones (frequently at 180 Hz)
- Additive stationary noise
- Impulse noise
- Amplitude and phase jitter
- Intermodulation distortion
- Unknown channel gain and phase response

We have compared the relative impact of each of these impairments by isolating them using a commercial telephone channel simulator, the TAS 1010 [13]. The use of the simulator enables us to isolate impairments and collect speech databases with selected degradations. Figure 2 depicts the experimental apparatus used for the collection of controlled degraded databases.

In order to test these channels using a task that produces a lower intrinsic error rate than phonetic recognition using the TIMIT/NTIMIT database, we used the CMU AN4 database [1], a continuous speaker-independent database consisting of strings of letters, numbers, and a few control words recorded under noiseless conditions using a close-talking microphone. The training set is transmitted through the simulator with all impairments set to zero with the exception of a low pass filter in the 0-4000 Hz frequency region. This is our baseline condition.

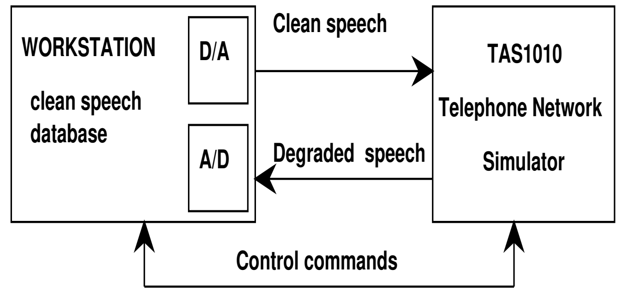

Figure 2: Experimental apparatus for the telephone network simulation experiments.

All experiments using this database were performed with the SPHINX-II speech recognition system [7], a more advanced version of the CMU recognition engine.

Effect of analysis bandwidth. The SPHINX-II system uses a front end based on Mel-Frequency Cepstral Coefficients (MFCC) to perform a frequency analysis in the 130- to 6900-Hz region. This choice of analysis bandwidth is appropriate when the system processes speech recorded through good-quality microphones such as the ARPA-standard Sennheiser HMD-414 close-talking microphone (referred to as CLSTLK). However when speech is recorded through telephone lines, a reduction in the analysis bandwidth normally yields lower recognition errors. As is seen in Fig. 3, the use of a reduced analysis bandwidth has little impact on rec-

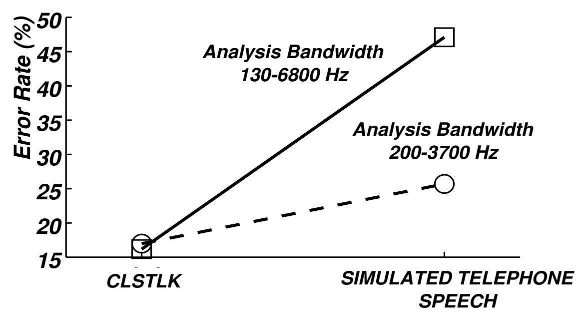

Figure 3: Effect of the analysis bandwidth on the recognition error rate for the AN4 database.

ognition accuracy using the CLSTLK mic, but it dramatically improves recognition error rate when the system is trained with high-quality speech and tested using simulated telephone speech. For this reason the results reported in the rest of the paper were all obtained with the reduced analysis bandwidth of 200-3700 Hz.

Similar results have also been observed in real telephone collected databases such as TIMIT and NTIMIT.

Effect of simulated impairments. Speech over the telephone network was simulated by transmitting clean speech through the TAS simulator with individual impairments set at the 50^{th}, 90^{th} and 99^{th} percentiles of degradation as determined by [2]. Table 3 summarizes the recognition error rates obtained with these various simulated impairments. Since statistical percentile ratings do not exist for the 180-Hz tone condition, three signal-to-tone intensity ratios were chosen arbitrarily.

Of the various impairments, additive stationary noise (C-message noise [5]), impulse noise, and interference by 180-Hz tones increased recognition error the most. SPHINX-II seems to be quite insensitive to the other impairments, even at the 99% levels. When all impairments are combined (except for the 180-Hz interfering tones), the results are still dominated by the additive C-message noise.

Impairment	Recognition Error		
	50%level	90%level	99%level
Clean Speech	13.0		
Baseline	18.3		
Amplitude jitter	17.0	18.7	17.7
Phase jitter	17.8	17.6	17.9
C-message noise	19.9	42.3	96.7
Impulse noise	17.6	18.5	22.9
2nd order intermodulation distortion	18.4	17.3	17.8
3rd order intermodulation distortion	18.0	18.3	17.7
All the above (average channel)	19.5	44.1	97.8
180-Hz tone interference	**Signal-to-Tone Ratio**		
	30 dB	**20 dB**	**10 dB**
	17.4	18.7	25.3

Table 3: Recognition error rates for the AN4 task obtained with various simulated telephone-channel impairments.

Effect of channel frequency response. Effects of the frequency response of the channel were evaluated by setting the simulator to match the frequency response of various actual channels including difficult channels such as the CCITT 1025, and more benign channels such as those with 4-dB or 12-dB high-frequency roll-off. The frequency response of these channels is shown in Fig. 4.

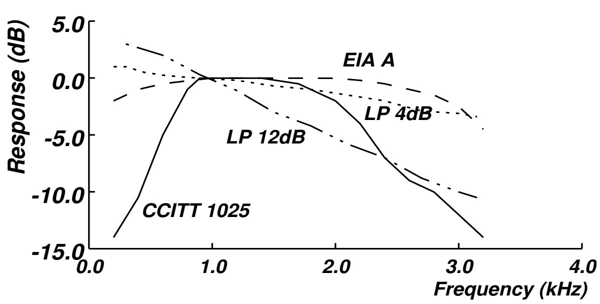

Figure 4: Frequency response of four different channels.

Table 4 summarizes results obtained using the four simulated frequency responses shown in Fig. 4. Increased error rate is generally observed for channels with a frequency response that is not flat. It was also found that the phase response of the simulated channels has almost no effect on recognition accuracy.

Channel	Recognition Error%
Baseline	17.2
CCITT 1025	32.0
EIA A channel	17.0
LP 4db channel	18.7
LP 12db channel	22.9

Table 4: Recognition error for different channels for the AN4 task. The system was trained in the baseline condition.

4. COMPENSATION ALGORITHMS

In a companion paper [10] we describe a number of algorithms to compensate for the effects of additive noise and linear filtering using additive cepstral correction vectors and cepstral highpass filtering. In this section we examine the ability of a subset of these algorithms, mean normalization, MFCDCN, PDCN and silence codebook adaptation, to compensate for the effects of the most damaging impairments from Tables 3 and 4.

Mean Normalization (MN). This algorithm computes the average cepstral mean for each sentence and subtracts it, which compensates for the effects of unknown linear filtering.

Silence Codebook Adaptation (SCA) and Multiple Fixed Codeword-Dependent Cepstral Normalization (MFCDCN). In SCA the silence models of the HMM are adapted to reduce the number of insertions produced during recognition over noisy channels. In our implementation of SCA, a separate codebook is created to describe silence segments, and the means of its codewords were adapted to describe the test data.

Before updating the silence codebook, incoming speech is processed with MFCDCN, which applies additive cepstral vectors based on SNR and other physical attributes of the speech to perform environmental compensation. Because of data limitations the MFCDCN compensation vectors were computed using the testing data itself. We do not observe a significant decrease in recognition accuracy when compensation vectors are obtained from independent speech samples (but with the same degradation).

Phone-Dependent Cepstral Normalization (PDCN) and MFCDCN. PDCN is similar to MFCDCN but performs environmental compensation based on the presumed phoneme identity, rather than on the basis of SNR. In the present study, PDCN is always used in combination with MFCDCN.

Experimental results. Results of our compensation experiments for C-message noise, composite channel effects, and various isolated impairments are summarized in Figs. 5, 6, and 7. As can be seen from the figures, combinations of SCA, MFCDCN, and PDCN are reasonably effective in ameliorating the effects of channel impairments at the 90% level, and somewhat less effective for impulse noise. It is clear that none of the procedures considered can cope with C-message noise in the worst 1% of telephone channels as simulated.

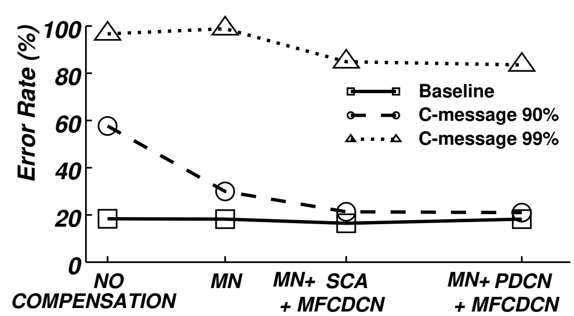

Figure 5: Performance of three compensation algorithms on speech contaminated with C-message additive noise.

5. SUMMARY

In this paper we compared error rates for speech recognition using clean speech and telephone-quality speech. We showed that pho-

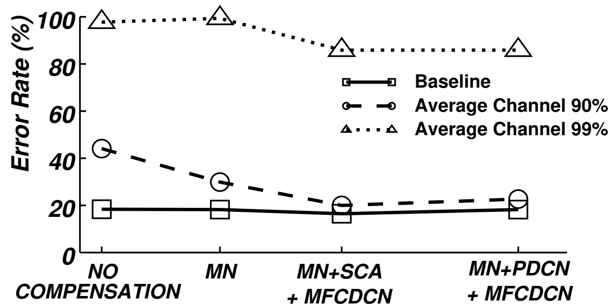

Figure 6: Performance of three compensation algorithms on speech passed through a simulated average telephone channel.

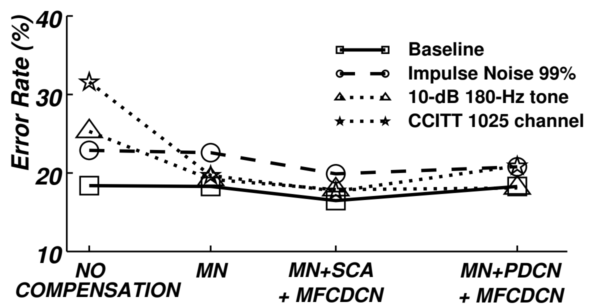

Figure 7: Performance of three compensation algorithms on speech contaminated with selected impairments.

neme error rates using the TIMIT/NTIMIT tasks increase by about 10% over telephone channels and that the bandwidth reduction found in telephone channels cannot alone account for this higher error rate. We also showed that compensation algorithms such as RASTA and CDCN are unable to recover all the distortion introduced by the telephone network.

Using a commercial telephone-channel simulator, we attempted to determine which channel impairments contribute the most to degradation in recognition accuracy. We found that the most significant impairments are the presence of stationary noise, impulse noise, and low-frequency tones, along with differences in channel frequency response between training and testing conditions.

We studied the effect of analysis bandwidth on the recognition system, demonstrating that a reduced analysis bandwidth improves recognition accuracy for telephone speech without greatly degrading the accuracy obtained with clean speech.

Finally, we examined the extent to which three cepstral compensation algorithms are able to neutralize the effects of the most damaging conditions. We found the compensation algorithms to be effective except in the case of the strongest additive stationary noises. Since the telephone simulations do not include some important phenomena including distortions caused by carbon-button microphones and stochastically-occurring impairments, we are continuing these analyses using speech recorded over actual telephone channels.

ACKNOWLEDGEMENTS

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. The authors thank Yoshiaki Ohshima and Fu-Hua Liu, along with the rest of the CMU speech group for many helpful discussions. We also gratefully acknowledge Telecom Analysis Systems for providing the telephone simulator used in these experiments, as well as Daniel Tapias (Telefónica I+D) for providing initial code for the TAS system. Pedro J. Moreno has been supported by a Fulbright fellowship.

REFERENCES

1. Acero, A. “Acoustical and Environmental Robustness in Automatic Speech Recognition”. Ph. D. Dissertation, ECE Department, CMU, Sept. 1990

2. Carey, M. B., Chen, H. T., Descloux, A, Ingle, J. F., and Park, K. I. “End Office Connection Study: Analog Voice and Voiceband Data Transmission Performance Characterization of the Public Switched Network”, *AT&T BLTJ*, 63, Nov. 1984.

3. Chigier, B. “Phonetic Classification on Wide-Band and Telephone Quality Speech “, ICASSP-91.

4. Davis, S. B., and Mermelstein, P. “Comparison of Parametric representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”. ICASSP-80.

5. Gaylor, W. D. “Telephone Voice Transmission. Standards and Measurements”. Prentice Hall, Englewood Clifss, N.J. 1989.

6. Hermansky, H., Morgan, N., Bayya, A., and Kohn, P. “Compensation for the Effect of the Communication Channel in Auditory-like Analysis of Speech (RASTA-PLP)”, *Proc. EUROSPEECH 1991*.

7. Huang, X., Alleva, F., Hon, H., Hwang, M.-Y., Lee, K., and Rosenfeld, R. "The SPHINX-II Speech Recognition System: An overview". Comp. Speech and Lang, **2**:137-148, 1993.

8. Jankowski, C., Kalyanswamy, A., Basson, S., and Spitz, J. “NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database “, ICASSP-90

9. Liu, F.-H., Stern, R. M., and Acero, A. “Efficient Joint Compensation of Speech for the Effects of Additive Noise and Linear Filtering”. ICASSP-92.

10. Liu, F.-H., Stern, R. M., Acero, A. and Moreno, P. J.“Environment Normalization for Robust Speech Recognition Using Direct Cepstral Comparisons”. ICASSP-94

11. Moreno P.J. “Speech Recognition in Telephone Environments”. M.S. Thesis., Dept. of ECE, CMU, Dec. 1992.

12. Murveit, H., Butzberger, M, and Weitraub, M., “Reduced Channel Dependence for Speech Recognition”, *Proc. DARPA Spoken and Natural Language*, Feb. 1992.

13. TAS1010 Voiceband Channel Simulators Operations Manual. Telecom Analysis Systems, 1989.