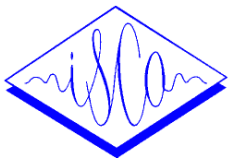

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A STUDY OF SPEECH RECOGNITION SYSTEM ROBUSTNESS TO MICROPHONE VARIATIONS: EXPERIMENTS IN PHONETIC CLASSIFICATION [1]

Jane Chang and Victor Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA
email: jwc@goldilocks.lcs.mit.edu

ABSTRACT

This paper presents experiments in phonetic classification conducted as part of a study on the effects of microphone variations on performance in speech recognition systems. The TIMIT corpus provides data recorded on a close-talking microphone, on a free field microphone and over telephone lines. The study focuses on the unmatched training and testing conditions under which degradation is most severe. Analysis of baseline performance characterizes the effects of microphone variations. Downsampling is shown to significantly improve performance for bandlimited conditions at the cost of some degradation for non-bandlimited conditions. Comparative analysis of microphone independent preprocessing techniques, including cepstral mean normalization, RASTA processing, spectral subtraction and codebook dependent cepstral normalization, reveals the effects and tradeoffs of different compensation techniques.

INTRODUCTION

Over the past decade, we have witnessed significant improvement in speech recognition technology – word error rates for large vocabulary, speaker independent, continuous speech recognition have been declining by half approximately every two years. Modern speech recognition systems rely heavily on automatic training techniques that determine the optimal parameter settings once a stochastic model has been formulated. In fact, the power of automatic learning is so pervasive that researchers have greatly simplified the signal representation (e.g., using Mel-frequency cepstral coefficients and their time derivatives) in the belief that model parameters can eventually capture acoustic invariance provided that a large amount of data is available for training.

Despite this apparent success, the deployment of speech recognition technology is still hampered by lack of robustness in system performance. It is not uncommon to have a system's error rate increase by ten-fold when tested using a microphone different from the one on which it was trained. Similarly, wide variations in recognition performance are often observed when the system is used either by a speaker or in an acoustic environment substantially different from what the system was exposed to during training. Over the past few years, researchers have begun to address the robustness issues with regard to microphone and ambient noise, resulting in compensation techniques such as spectral normalization, subtraction and filtering. These techniques generally attempt to decrease the system's sensitivity to input variations by effectively inserting a "speech enhancement" module in the recognizer's front-end to clean up the signal, which is then fed to the otherwise unaltered recognition system. Most of these techniques have been demonstrated to be effective, by varying degrees, on the recognition tasks for which the systems were evaluated.

What techniques are most effective in combating mismatches between training and testing conditions? The answer is far from unequivocal for several reasons. The preprocessing techniques described in the literature were usually evaluated on different recognition tasks using disparate speech corpora, making inter-technique comparisons difficult. In some instances when the same task and corpus were used, the metric for comparison is word error rate. Since word error rate is a function of acoustic, lexical and language modeling, it is difficult to isolate the contribution of the techniques to acoustic modeling and to generalize the results across domains.

This study addresses the issue of microphone robustness for matched and unmatched training and testing conditions. The primary objectives are to improve our understanding of the effects of microphone variations by focusing on system performance at the sub-word level, and to use this understanding as a basis for comparing different "microphone-independent" preprocessing techniques. In the remainder of this paper, we will first describe the task, corpus and system used. We will then make general comments on the spectral characteristics of the data recorded on different microphones. Baseline performance for matched and unmatched training and testing conditions will be presented next, followed by a comparison of several preprocessing techniques. We conclude with a summary of our salient findings and a discussion of future work. Due to space limitations, this paper presents only classification results. Interested readers are referred to [1] for more recognition results and detailed discussions.

TASK, CORPUS AND SYSTEM

We have chosen phonetic classification and recognition tasks as the basis for a comparative study of microphone variabilities and a standardized benchmark of compensa-

[1]This research was supported by the Department of Defense under Contract MDA904-93-C-4180. J. Chang receives support from AT&T Bell Laboratories.

tion techniques. Being word and domain independent, phonetic experiments reduce the confounding effects of system variables such as vocabulary and language modeling. It is our hope that increased microphone robustness at the phonetic level should propagate up to word and sentence levels, leading to better overall system performance and generalizing to different task domains. Experiments are conducted on matched and unmatched training and testing conditions with particular emphasis placed on the conditions where the training microphone is of higher quality than the testing microphone, since this is the most likely scenario for technology deployment.

Experiments are conducted using TIMIT [2], a phonetically rich, continuous speech corpus with time-aligned orthographic and phonetic transcriptions, making it possible to gather statistics on specific speech sounds and to conduct phonetic studies. TIMIT is particularly useful for this study because there exist essentially three simultaneous recordings made by each subject using three different transducers. The original TIMIT corpus released by NIST was recorded using a Sennheiser HMD-414 microphone. This data was subsequently passed through the telephone channel with the timing information preserved by researchers at NYNEX, resulting in the NTIMIT corpus [3]. The third set of data, less known to the research community, was recorded in stereo with the Sennheiser using a Bruel and Kjaer (B&K) Model 4165 microphone. This data has not been released by NIST and was only recently retrieved from archive tapes. Unfortunately, about 3% of the standard NIST training and testing sentences recorded on the B&K could not be recovered. As a result, the training and testing sets used for all three microphones are the 97% of TIMIT data common to all three microphones, containing 139,257 phonetic tokens in 3,603 sentences from 451 speakers for training and 12,978 phonetic tokens in 383 sentences from 48 speakers for testing.

The phonetic classification and recognition systems use the same set of representations and components so that direct comparisons among various conditions can be made. We have adopted rather simple representations since we are interested in the *relative*, rather than absolute, error rates. Nevertheless, our systems achieved baseline performance comparable to those reported in the literature on the same task and data. The systems use a signal representation consisting of 14 Mel-frequency cepstral coefficients (MFCCs). The feature vector for each phone segment is made up of 3 averages (each covering one-third of the segment), 2 time derivatives (one for each neighboring third of the three sub-segments) and log duration. Full covariance Gaussian models are used for the 39 phonetic classes commonly used in the literature [4] to facilitate easy comparison.

GENERAL CHARACTERISTICS

Observable acoustic variations come from diverse sources, each affecting the speech signal in a different way. For example, convolutional distortion may be introduced by speakers' vocal tracts, room acoustics, and microphone transfer functions. Environmental noise can add distortions to the signal, and transmission channel can impose bandwidth limitations. The Sennheiser and B&K corpora were recorded in a relatively noise-free environment [2]. The Telephone data

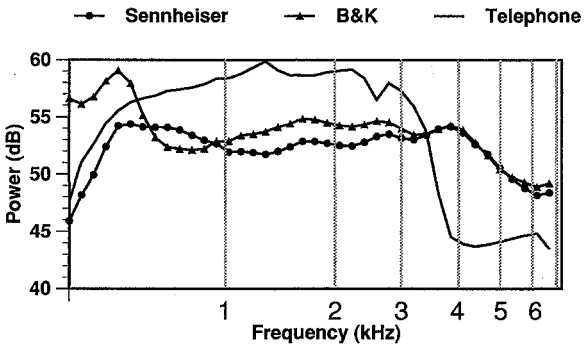

Figure 1: The mean MFSC vector computed across the entire training set for the Sennheiser (S), B&K (B) and Telephone (T) corpora.

was obtained by playing the Sennheiser data through an artificial mouth, recording the resulting signal using a telephone handset, and then transmitting it over a telephone network [3].

The Sennheiser is a noise-canceling, pressure-gradient microphone with a flat response in the frequency range of interest (approximately 150–7000 Hz). Since it is a close-talking, headset-mounted microphone, the Sennheiser records from a relatively constant distance and position near the mouth and is not very sensitive to environmental noise. An unfortunate side effect is that signals emitted from the nose as well as the throat are also attenuated, resulting in significantly lower energy during voice bar and nasal murmur. The B&K is a free field pressure microphone, also with a flat frequency response in the frequency range of interest. Signals recorded with the B&K are more likely to be corrupted by environmental noise, since the microphone is placed some distance away from the mouth. Furthermore, the data tend to show more variability due to the fact that the distance and position of the microphone are hard to maintain. The Telephone data are characterized by the *aggregate* of the Sennheiser and the telephone handset, as well as the noise and bandlimiting effects, to a frequency range of about 3 kHz, of a telephone network.

Figure 1 shows the general spectral characteristics of the three corpora. The spectral vectors are obtained by averaging the Mel-frequency spectral coefficients (MFSCs) [5], computed at 5 ms intervals, for all the data in the training sets. Since the three data sets are identical in speaker and content, the mean spectral vectors shed direct light on the only difference between them i.e., the ways the data were recorded. As Figure 1 shows, the Sennheiser and B&K data differ primarily at low frequencies. The positive slope of Sennheiser spectrum at low frequency reflects the microphone's noise-canceling capability. The low frequency peak in the B&K spectrum is presumably due to the combined effects of the presence of low frequency noise and energy in non-oral resonances. Telephone data differs most significantly in the high frequency region where it suffers from bandlimiting. As we shall see, these spectral differences help to explain the nature of the additional classification and recognition errors incurred under unmatched training and testing conditions.

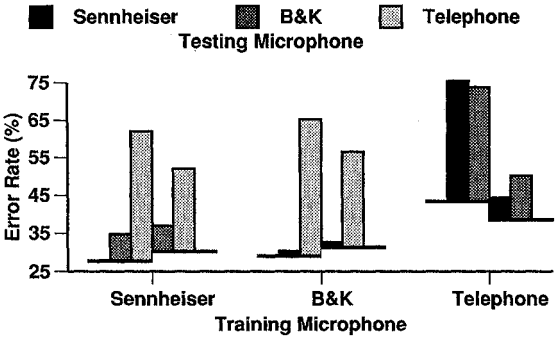

Figure 2: Classification error rates for matched and unmatched training and testing conditions (see text for explanation).

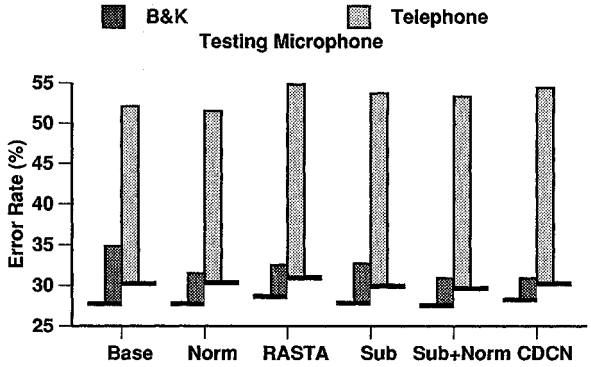

Figure 3: Classification error rates for various preprocessing techniques (see text for explanation).

BASELINE PERFORMANCE

We performed initial experiments to establish baseline performance for matched and unmatched training and testing conditions. Figure 2 shows classification results. For each of the three training conditions on the horizontal axis, there are two sets of results depending on whether the data is sampled at 16 kHz (on the left) or is downsampled to 8 kHz (on the right). The bold horizontal baselines indicate performance for matched training and testing and the shaded vertical bars indicate the *additional* degradation incurred when the testing microphone does not match the training microphone. We introduce the notation [X:Y] to denote the condition of training on microphone X and testing on microphone Y.

Our results show matched training and testing produces the best performance for each microphone. There is a moderate degradation from the Sennheiser to the B&K and a much larger degradation to the Telephone, especially for unmatched training and testing conditions. In the remainder of this paper, we restrict our discussion to the two conditions of interest, namely conditions [S:B] and [S:T].

Condition [S:B] Phonetic classification error rate for condition [S:S] is 27.6%. [2] The error rate increases by 25% (to 34.5%) for condition [S:B]. Closer examination of confusion statistics reveals that one quarter of the additional confusions occur between voiced fricatives and stops and their unvoiced counterparts, another quarter occur within vowels, and one sixth occur between weak events and closures/silences. Variations in low frequency tilt and noise can explain these voicing, formant and low energy confusions. For example, the Sennheiser cancels voicing while the B&K records both voicing and noise at low frequencies.

Condition [S:T] Without downsampling, the error rate more than doubles (to 61.8%) for condition [S:T] from the condition [S:S]. While this is largely due to the loss of high frequency information in telephone transmission, the signal representation also contributes to the increase. Specifically, the computation of the MFCC signal representation at 16 kHz requires taking the logarithm of the small spectral values outside of the telephone bandwidth and applying a cosine transform, thereby corrupting all of the telephone cepstral coefficients. Downsampling the data to 8 kHz avoids this additional degradation and reduces the error rate by 16% (to 51.8%) for condition [S:T], as shown in Figure 2. Although this procedure increases the error rate somewhat for the non-bandlimited conditions, it significantly improves the performance for the bandlimited conditions, often by a large amount.

Even after downsampling, the error rate for condition [S:T] is still much greater than condition [S:B], suggesting additional sources of degradation other than bandlimiting, such as higher levels of network distortion and noise [7]. Without high frequency information, the low energy events, including bandlimited fricatives, stops and affricates in addition to closures and silences, are especially susceptible to these microphone effects. Indeed, examination of confusion statistics shows that one third of the additional errors involve closures and silences, one sixth occur within weak events, and another one sixth occur within vowels.

PREPROCESSING TECHNIQUES

Front-end preprocessing techniques that compensate for distortion and noise prior to speech recognition are the most common and direct techniques towards microphone robustness. Part of this study analyzes several microphone independent preprocessing techniques. Microphone independent techniques do not require any of the microphone specific information, data and training required by microphone dependent techniques and are therefore less constrained. Figure 3 shows classification error rates for different preprocessing techniques under conditions [S:B] and [S:T]. For each technique, the two bold horizontal baselines indicate performance for condition [S:S] under 16 kHz and 8 kHz sampling rates. The vertical bars represent the additional degradation incurred under unmatched testing.

Normalization techniques compensate for convolutional differences by estimating correction vectors from long term statistics of slowly varying microphone effects. These techniques vary in domain, weighting and amount of data across which averages are computed. Of the variations we implemented, cepstral mean normalization [8] is most simple and effective, taking advantage of the additive cepstral correlate to convolution. RASTA [9] processing is another normaliza-

[2] Note that this baseline error rate is comparable to those reported in the literature [6], giving us some assurance that our signal representation is reasonable.

tion technique that uses a highpass filter to remove slowly varying microphone effects.

Spectral subtraction [10] techniques compensate for additive microphone effects by estimating correction vectors from short term noise and signal levels. These techniques vary in domain and amount of data across which noise and signal levels are computed. We found subtraction to be most effective in the MFSC domain. Subtraction and normalization can be combined to compensate for both additive and convolutional effects by cascading preprocessing blocks. Codebook dependent cepstral normalization (CDCN) [8] is a more complex preprocessing technique that compensates jointly, rather than in cascade, for convolutional and additive effects.

Condition [S:B] All of the preprocessing techniques improved the baseline condition [S:B] (34.5%) without overly degrading the condition [S:S]. Mean normalization achieved an error rate of 31.2%, compensating for almost half of the additional errors incurred by unmatched testing. In our experiments, we did not find RASTA (32.2%) or subtraction (32.4%) to be as effective, the latter case partly because there are more convolutional than additive effects for the condition [S:B].

In systems that allow more complex processing, the most effective technique was subtraction followed by normalization, which reduced the error rate to 30.6%. The combination takes advantage of the relative merits of both techniques and achieves the best result in this study. While subtraction improves classification on closures/silences and consonants, normalization accounts for a wider range of errors, including vowels, consonantal voiced pairs, and closures/silences. CDCN, a more complex technique, did not perform as well as disjoint subtraction and normalization in classification.

Condition [S:T] Although downsampling is shown to be very effective, none of the preprocessing techniques are very effective in improving the baseline downsampled result (51.8%). In fact, only mean normalization (50.6%) lowered the error rate by improving classification on closures/silences and vowels. This suggests the nature of telephone degradation significantly varies from the convolutional and additive models that are effective for the B&K.

SUMMARY AND FUTURE WORK

This paper describes several experiments designed to help understand the effects of different microphones and the comparative advantages of preprocessing techniques proposed in the literature. By focusing on the phonetic classification task and using the TIMIT corpus, we are able to make some definitive statements regarding the effects different microphones have on the speech signal without external confounding factors such as vocabulary and language model.

The differences between close-talking, noise-canceling microphone recordings and free field microphone recordings made in a noise isolated environment occur mainly at low frequencies. Our experiments show an increase in phonetic classification error rate of about 25% when the system is trained on the Sennheiser data and tested on the B&K data, largely due to mis-classification of vowels, consonant pairs that differ in voicing, and closures/silences. The error rate increases by more than two-fold when tested on the Telephone data with the same bandwidth. Downsampling significantly improves performance for bandlimited conditions at the cost of degradation for non-bandlimited conditions.

Our comparison of various preprocessing techniques indicates that spectral subtraction followed by normalization is the most effective means for increasing microphone robustness for condition [S:B], reducing the error rate by nearly one-third. On the other hand, standard preprocessing techniques are not very effective when dealing with Telephone data, suggesting that further study is necessary in order to better understand the nature of telephone degradation.

We have also conducted experiments on phonetic *recognition*, and the results are reported in [1]. In general, phonetic recognition results follow the same trend, except that the error rate, taking into account substitution, insertion and deletion, is higher by about 50%. Errors, mainly substitutions and deletions, become relatively more concentrated in closures/silences, suggesting difficulties in segmentation. On the condition [S:B], preprocessing improves recognition performance, but on the condition [S:T], performance is so degraded that preprocessing and even downsampling do not improve results as much as hoped.

Future work will include exploring how to improve performance in the SUMMIT recognition system beyond the realm of preprocessing techniques. We will investigate the use of acoustic features that are more resilient to input variations. For example, segment duration is microphone invariant, and can potentially help compensate for low frequency variations and distinguish voicing, vowels, and closures/silences. We will also pursue the development of models and the corresponding training and search strategies at the higher levels of the recognition process that account for corrupted models of the speech input, including adaptive techniques. For example, one can potentially make the system more robust by training it on combined corrupted and uncorrupted data.

REFERENCES

[1] J. Chang, "Speech recognition system robustness to microphone variations", S.M. Thesis, MIT, expected 1994.

[2] W. Fisher, G. Doddington and K. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status", *Proc. DARPA Speech Recognition Workshop*, 93–99, Feb 1986.

[3] C. Jankowski, A. Kalyanswamy, S. Basson and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database", *Proc. ICASSP*, 109–112, 1990.

[4] K. Lee and H. Hon, "Speaker-independent phone recognition using Hidden Markov Models", *Tran. ASSP*, 1641–1648, Nov 1989.

[5] H. Meng, "The Use of Distinctive Features for Automatic Speech Recognition", S.M. Thesis, MIT, 1991.

[6] B. Chigier, "Phonetic classification on wide-band and telephone quality speech", *Proc. DARPA Speech and Natural Language Workshop*, 291–295, 1992.

[7] P. Moreno and R. Stern, "Sources of degradation of speech recognition in the telephone network", *Proc. ICAASP*, I-109–112, 1994.

[8] F. Liu, R. Stern, X. Huang and A. Acero, "Efficient cepstral normalization for robust speech recognition", *Proc. DARPA Human Language Technology Workshop*, Mar 1993.

[9] H. Hermansky, N. Morgan, A. Bayya and P. Kohn, "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)", *Proc. Eurospeech*, 1367–1370, 1991.

[10] D. Compernolle, "Increased noise immunity in large vocabulary speech recognition with the aid of spectral subtraction", *Proc ICAASP*, 27.6.1–4, 1987.