A STUDY OF SPEECH RECOGNITION SYSTEM ROBUSTNESS TO MICROPHONE VARIATIONS[1]

Jane Chang and Victor Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA

ABSTRACT

This study seeks to improve our understanding of the effects of microphone variations on speech recognition systems. The TIMIT corpus provides data recorded on close talking and far field microphones and over telephone lines. The SUMMIT system is configured for phonetic classification and recognition. At the last ICSLP, we presented an analysis of the data and experiments in phonetic classification using a baseline system and various preprocessing techniques. In this paper, we present experiments in phonetic recognition using an improved baseline system and compensation techniques that require varying amounts of microphone specific data.

INTRODUCTION

Over the past decade, we have observed significant improvement in speech recognition technology – word error rates for large vocabulary, speaker independent, continuous speech recognition have been decreasing by half approximately every two years. Despite this apparent success, lack of robustness in system performance still hampers the deployment of speech recognition technology. We often observe severe performance degradations when a system is used by speakers or under conditions, such as environment and microphone conditions, that are substantially different from those used during training. Over the past few years, researchers have begun to address these robustness issues, resulting in a wide range of techniques that compensate for and reduce system sensitivity to such input variations. These compensation techniques include preprocessing techniques that clean up the signal before input to the recognizer and training techniques that account for input variations within the recognizer itself.

In this study, we address the issue of microphone robustness with a focus on realistic mismatched conditions when the testing microphone is lower in quality than the training microphone. We seek to improve our understanding of the effects of microphone variations on speech recognition systems at the subword level and to use this understanding to determine which techniques are more effective in compensating for microphone variations in our system. The focus on the phonetic rather than word level isolates the microphone effects, reduces the confounding effects of corpus and system dependent variables, facilitates the direct comparison of results, and allows the generalization of results across domains. The following sections discuss the data, baseline system and compensation techniques used in this study.

DATA

The TIMIT [3] acoustic phonetic corpus provides continuous speech data with time aligned phonetic transcriptions. TIMIT is particularly useful for studies of microphone variations because it provides three different recordings of the same data. The original release of TIMIT was recorded using a Sennheiser HMD-414. Subsequently, the Sennheiser data was recorded over telephone lines and released as NTIMIT [4]. 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) 4165. With help from NIST, we recovered 97% of the B&K utterances. For consistency, we only use utterances that are common to all three microphones. As a result, our training sets consist of 97% of the NIST training utterances, our testing sets consist of all the NIST core test utterances, and our development sets consist of 383 of the remaining utterances.

This study focuses on the mismatched conditions when training on the Sennheiser and testing on the B&K and Telephone. In contrast to the Sennheiser, the B&K does not have noise canceling abilities and records from a more variable distance and position farther from the mouth. As a result, the B&K is more sensitive to non-oral resonances emitted from the nose and throat, as well as any environmental noise present in the recording booth [3]. The Telephone is characterized by the combination of the Sennheiser with a telephone handset and channel that introduce noise and bandlimiting effects [4]. Figure 1 shows general spectral characteristics of the data. The Sennheiser

[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.

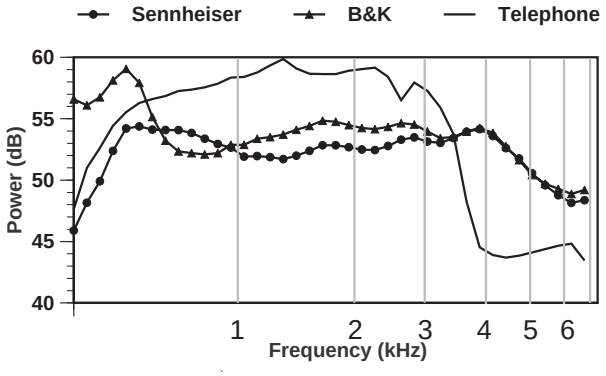

Figure 1: Mean Mel Frequency Spectral Coefficient (MFSC) vectors computed over the entire training sets for the Sennheiser, B&K and Telephone.

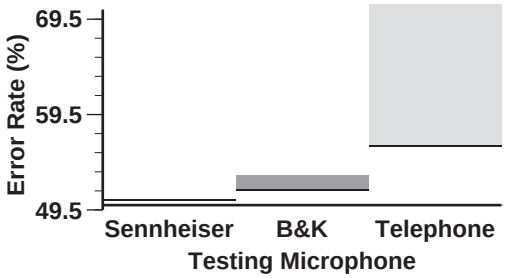

Figure 2: Baseline recognition error rates in percent. For each testing microphone on the x-axis, the line indicates the matched error when training and testing on that microphone, while the bar indicates the mismatched error when training on the Sennheiser and testing on that microphone.

and B&K differ mostly at low frequencies, where the slope of the Sennheiser reflects its noise canceling ability, and the peak in the B&K reflects the presence of energy due to non-oral resonances and noise. The Telephone shows the effects of normalization and noise at lower frequencies and bandlimiting at higher frequencies.

BASELINE SYSTEM

The SUMMIT [10] system is a segment based speech recognition system that explicitly detects phonetic segment boundaries in order to extract features in relation to specific acoustic events. For our comparative experiments, we are interested in relative rather than absolute performance. Therefore, in configuring SUMMIT for phonetic recognition, we use rather simple components to maintain consistency, facilitate training and otherwise reduce confounding effects, at the expense of achieving optimal performance[2]. The system uses a Mel Frequency Cepstral Coefficient (MFCC) representation. The segmentation algorithm [10] is based solely on spectral change with no probabilistic modeling. The features extracted for each segment consist of 3 averages over each third of the segment, 2 derivatives with neighboring segments, and duration. The models consist of context independent diagonal Gaussian mixtures and a bigram language model. Performance is evaluated on 56 classes, with all closures collapsed into one class.

As described, the recognition system is consistent with the classification system used in our previous experiments [2]. Original experiments using this system show that recognition errors are larger in number but similar in kind to the *classification* errors we had studied in detail [1]. In this paper, we present experiments using an improved system that incorporates two known findings. First, the system uses Cepstral Mean Normalization (CMN). Studies [1, 7] show that CMN significantly reduces degradation under mismatched training and testing conditions without degrading the matched condition. Second, the system handles wideband and narrowband speech. Studies [1, 5] show that downsampling significantly reduces degradation under mismatched training and testing conditions that also differ in bandwidth. Our system produces both wideband and narrowband models in training and uses the model that matches the bandwidth of the input signal in testing.

Figure 2 shows baseline recognition error rates. We introduce the notation (X, Y) to denote the condition of training on X and testing on Y. When training and testing on the same microphone, the increase in error from 49.9% under (S, S) to 51.1% under (B, B) and 55.6% under (T, T) reflects the decrease in quality from the Sennheiser to the B&K and Telephone. When training on the Sennheiser, the increase in error from 49.9% under (S, S) to 52.5% under (S, B) and 70.3% under (S, T) reflects the increase in mismatch between the Sennheiser and the B&K and Telephone. As the Sennheiser and B&K are relatively similar, testing on the B&K rather than the Sennheiser incurs a relatively small 5% error increase. Furthermore, differences between the microphones at low frequencies can explain many of the additional errors involving voicing, formants and weak events [1]. As the Sennheiser and Telephone are relatively different, testing on the Telephone rather than the Sennheiser incurs a relatively large 41% error increase. Other than bandlimiting, the Telephone introduces higher levels of distortion and noise [5], and without high frequency information, recognition is more sensitive to these effects.

Regardless of the microphone, most of the additional errors under mismatched conditions are deletions, especially of weak events such as stop closures.

[2] The lowest phonetic recognition error rate achieved by SUMMIT using context dependent models is approximately 30% [8].

This suggests a lack of robustness in the segmentation, classification and search components. In our previous study of classification [2], we forced the system to use the segment boundaries derived from the time aligned phonetic transcription and bypassed the search component. In this paper, we study the relative robustness of the segmentation and classification components by conducting a set of experiments in which we "mix and match" segmentation networks with classification models while keeping the search component intact. For example, when training on the Sennheiser and testing on the B&K and Telephone, we force the system to use the matched segmentation network proposed when testing on the Sennheiser. Similarly, when training and testing on the Sennheiser, we force the system to use the mismatched segmentation network proposed when testing on the B&K and Telephone. The results of these experiments suggest that classification contributes much more than segmentation to the overall degradation under mismatched conditions and that the segmentation component is relatively robust to microphone variations. Nevertheless, we note that segmentation and classification cannot really be separated. Well matched boundaries and models can result in fewer errors, while poorly matched boundaries and models can result in large numbers of deletions, such as when testing on the Telephone. Furthermore, search and other recognition parameters also affect segmentation and classification, as we will discuss in the following section.

COMPENSATION TECHNIQUES

Different compensation techniques use different amounts of microphone specific data. Preprocessing techniques can compensate for input variations before recognition without the use of microphone specific data. The baseline (BASE) uses CMN, which subtracts an estimate of the convolutional distortion in the input signal. Among the other preprocessing techniques with which we experimented [1], we present results for log spectral subtraction (SUB) [9], which subtracts an estimate of additive noise, and Codebook Dependent Cepstral Normalization (CDCN) [7], which subtracts an estimate of both convolutional and additive effects.

Other techniques use a relatively small amount of microphone specific data to compensate for input variations during recognition. As suggested, search parameters that control the tradeoff between deletions and insertions can compensate for large numbers of deletions. Although acoustic models are microphone dependent, duration and language models are not. Therefore, parameters that control the relative weights of acoustic, duration and language models may compensate for acoustic mismatch. Other parameters, such as those in segmentation and classification, can also compensate for microphone variations. We present results for microphone adaptation (ADAPT), which iteratively adapts these recognition parameters for each testing microphone in a supervised manner using the development set, described above as approximately 10% the size of the training set.

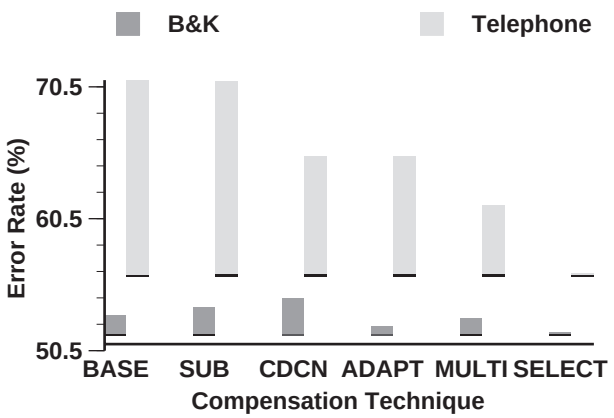

Figure 3: Compensated recognition error rates in percent. For each compensation technique on the x-axis, the lines indicate baseline matched errors for (B, B) and (T, T), while the bars indicate mismatched errors under (S, B) and (S, T) using that technique.

Training techniques use a relatively large amount of data to train microphone specific models. For consistency, we present results using the same training parameters for all techniques. For example, we maintain the same total amount of training data and maximum number of mixtures per model. Multi-style training (MULTI) [6] pools one third of each microphone training set into a single model. Microphone selection (SELECT) uses each microphone training set to produce a separate model and selects the highest scoring model for each test utterance.

Figure 3 shows compensated recognition error rates in percent. In general, from left to right, the compensation techniques use increasing amounts of microphone specific data to achieve increasing reductions in mismatched errors, approaching *matched* errors for the B&K and Telephone. Except for multi-style training which is discussed below, the techniques do not significantly change errors for the Sennheiser.

Preprocessing techniques are general and effective techniques for increasing microphone robustness that do not require microphone specific data and can be applied to any microphone. The simple technique of CMN is at least as effective as log spectral subtraction and most of the other preprocessing techniques with which we experimented [1]. One exception is

the significantly more complex technique of CDCN, which reduces the degradation from (S, S) to (S, T) by 28%, achieving a 64.5% error when testing on the Telephone.

Microphone adaptation is an effective technique for increasing microphone robustness that only requires a relatively small amount of microphone specific data. Among the adapted parameters, the search parameters that control the tradeoff between deletions and insertions compensate for the largest percentage of errors. The parameters that control model weights and boundary thresholds are also effective. Similar to CDCN, microphone adaptation reduces the degradation from (S, S) to (S, T) by 28%. In addition, microphone adaptation reduces the degradation from (S, S) to (S, B) by 57%, achieving a 51.7% error when testing on the B&K. Slight improvements can be achieved by combining microphone adaptation with preprocessing techniques such as CDCN. Further experiments can explore the use of smaller amounts of data.

Training techniques are the most effective techniques but require large amounts of microphone specific data. Multi-style training pools different microphone data and attempts to capture acoustic features that are more consistent across microphones in general. As a result, multi-style training further reduces error for the Telephone to 60.8%, at the expense of a slight degradation in the matched condition. Microphone selection effectively runs three recognizers in parallel. The simple selection algorithm based on the highest recognition score selects the matched model for most test utterances, and the small percentage of mismatches that mostly occur between the relatively similar Sennheiser and B&K do not significantly change error. As a result, microphone selection essentially achieves microphone robustness for the TIMIT microphones. Further improvements can be achieved by combining microphone selection with other compensation techniques that improve performance under matched conditions. Experiments in multi-style training show that more data and mixtures can reduce the degradation under the matched condition. This suggests the combination of multi-style training for unknown microphones with microphone selection for known microphones. Although we do not have unknown microphone data, experiments using the combination even on known microphone data result in slight improvements over microphone selection alone.

SUMMARY AND FUTURE WORK

In this paper, we use the TIMIT corpus and SUMMIT system to study the effects of microphone variations on phonetic recognition. Using an improved baseline system, we experiment with compensation techniques that require varying amounts of microphone specific data to achieve varying degrees of microphone robustness. Preprocessing techniques that do not use any data and adaptation techniques that use a small amount of data can significantly increase robustness and may offer sufficient compensation for relatively small mismatches, such as between the Sennheiser and B&K. Training techniques that use a large amount of data can achieve microphone robustness and offer significant improvements over other techniques for relatively large mismatches, such as between the Sennheiser and Telephone.

In future experiments, we will further explore training issues and examine the quantity and quality of data required. We will also address the issue of robust feature extraction and investigate features that are less microphone dependent. In addition, we will incorporate improvements to the SUMMIT system, such as improving the segmentation component by incorporating probabilistic models.

REFERENCES

[1] J. Chang, *Speech recognition system robustness to microphone variations*, SM Thesis, MIT, 1995.

[2] J. Chang and V. Zue, "A study of speech recognition system robustness to microphone variations: experiments in phonetic classification", *Proc.* ICSLP, 995–998, 1994.

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

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

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

[6] R. Lippman, E. Martin and D. Paul, "Multi-style training for robust isolated-word speech recognition", *Proc.* ICASSP, 705–708, 1987.

[7] F. Liu, R. Stern, A. Acero and P. Moreno, "Environment normalization for robust speech recognition using direct cepstral comparisons", *Proc.* ICASSP, II:61–64, 1994.

[8] M. Phillips and J. Glass, "Phonetic transition modeling for continuous speech recognition", *J. Acoust. Soc. Am.*, 95:2877, 1994.

[9] D. Van Compernolle, "Increased noise immunity in large vocabulary speech recognition with the aid of spectral subtraction", *Proc.* ICASSP, 1143–1146, 1987.

[10] V. Zue, J. Glass, D. Goodine, H. Leung, M. Phillips, J. Polifroni and S. Senneff, "Recent progress on the SUMMIT system", *Proc.* DARPA *Speech and Natural Language Workshop*, 380–385, 1990.