AUGMENTED CEPSTRAL NORMALIZATION FOR ROBUST SPEECH RECOGNITION

Alex Acero, Xuedong Huang
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052, USA

ABSTRACT

We proposed an augmented cepstral mean normalization algorithm that differentiates noise and speech during normalization, and computes a different mean for each. The new procedure reduced the error rate slightly for the case of same-environment testing, and significantly reduced the error rate by 25% when an environmental mismatch exists over the case of standard cepstral mean normalization.

1. INTRODUCTION

Accuracy of speech recognition systems rapidly degrades when deployed in acoustical environments different than those used in training. A great deal of attention has been paid recently to this problem [1][4], in an effort to deploy the technology in the field. When these mismatches occur, the speech recognizer could become unusable.

There are two main ways to reduce the mismatch: by adapting the HMM models or by compensating the input features. HMM model adaptation [2][7] is powerful because there are more parameters in the system that can be changed to reduce the acoustical mismatch. Since there exists a great constraint on how these parameters can be modified, which is limited by a simple model of the degradation, this adaptation should not require a lot of speech data. Nevertheless, it has been found that due to inaccuracies in the model, the use of more data typically results in higher accuracy. An advantage of this approach is that it provides a graceful degradation under very severe mismatch conditions. One of the problems of these approaches is the large number of computations required to adapt the model, which is not amenable to the rapid adaptation needed in rapidly changing environments, such as telephone applications.

Approaches that operate by compensating the input features [1][5][7] have the limitation of having to make transformations without the acoustic or language knowledge used in the search process, possibly using a inaccurate correction vector. On the other hand, they typically require little computation, achieve rapid environment adaptation, and if the mismatch is not very severe they can perform as well as the model adaptation approaches.

Both in feature normalization and in model adaptation, it is advantageous to use stereo recordings of the clean and noisy speech [1][5][7], because of the superior segmentation obtained with the clean speech often results in improvements in recognition accuracy. However, in many situations these recordings are not available, so algorithms have to be devised that operate only with the noisy data [6].

Cepstral Mean Normalization (CMN) [5] has been successfully used as a simple yet effective way of normalizing the feature space. Not only does it provide with an error rate reduction under mismatched conditions, but also it has been shown to yield a small decrease in error rate under matched conditions. Those benefits, together with the fact that it is very simple to implement, is the reason why many current systems have adopted it.

A problem of CMN is that it does not discriminate between silence and voice when computing the utterance mean, and therefore the mean is affected by the amount of noise included in the calculation. For improved speech recognition accuracy, we propose an augmented CMN procedure that differentiates noise and speech during normalization, and computes a different mean for each one.

2. ALGORITHM DESCRIPTION

Let's have the cepstrum vectors for an utterance represented as $X = \{\mathbf{x}_i\}$ where $0 \le i \le N$. Our algorithm will transform the input feature into a normalized feature $Z = \{\mathbf{z}_i\}$.

Let's further define **n** is the average cepstral vector for the noise frames in utterance *X*, and **s** is the average cepstral vector for the speech frames in *X*. Since in the training phase we will have a speech database containing a number of utterances, we can define $\mathbf{n}_{avg}$ as the average cepstral vector for all noise frames and $\mathbf{s}_{avg}$ as the average cepstral vector for all speech frames.

The proposed technique consists of adding a correction vector to all noise frames and a different correction vector to all speech frames, so that the average cepstrum vector for noise frames after compensation is identical for all utterances. The same applies to speech frames.

Since a perfect noise/speech discrimination is not always feasible, we can generalize this approach by incorporating p_i , the *a posteriori* probability of frame *i* being noise. With this in mind, the new cepstral normalization algorithm consists of subtracting a correction vector $\mathbf{r}_i$ to each incoming cepstrum vector $\mathbf{x}_i$:

$$\mathbf{z}_i = \mathbf{x}_i - \mathbf{r}_i$$

where the correction vector $\mathbf{r}_i$ is given by

$$\mathbf{r}_i = p_i(\mathbf{n}-\mathbf{n}_{avg}) + (1-p_i)(\mathbf{s}-\mathbf{s}_{avg})$$

It is clear that for all normalized utterances, the average noise cepstral will be $\mathbf{n}_{avg}$, and the average speech will be $\mathbf{s}_{avg}$. Therefore, this is a generalization of the CMN algorithm to two classes: noise and speech.

The use of the *a posteriori* probability p_i allows a smooth interpolation between noise and speech, much like the SDCN and ISDCN algorithms [1]. The HMM probabilities could be used to estimate p_i, though we found that a simpler modeling based exclusively on the signal energy is just as effective in moderate SNR environments. Moreover, our results showed that typically only a few frames within a sentence would have values of p_i significantly different than either *0* or *1*. Therefore, we constantly updated a threshold separating speech from noise, based on a histogram of log-energies. This results in a very simple implementation that is also very effective [3].

Sankar and Lee [8] also proposed the use of two different cepstral means for noise and speech. They showed that if HMM information is available, both noise and speech cepstral means could be estimated by Maximum Likelihood (ML). While a ML estimation procedure for $\mathbf{n}$ and $\mathbf{s}$ using the HMM parameters should be in principle better than their sample means, it appears that the difference in practice is small, unless the sentence is very short or the noise level is very high. In fact, in [8] the use of ML estimation versus the sample mean did not result in any significantly different recognition accuracy for a system using cepstral bias removal.

3. EXPERIMENTAL RESULTS

We evaluated this algorithm with the *DARPA 5000-word Wall Street Journal* task, using the *si_dev5* test set. Since we wanted to conduct our experimental results on desk-top microphones with standard PC sound-cards, we down-sampled the test set from 16kHz to 11kHz. Our recognition system was a compact version of Whisper [3]. Briefly, Whisper is a PC-based continuous-speech speaker-independent system based on semi-continuous HMM system with context-dependent tri-phone models. Since Whisper allows us to trade CPU and memory usage for accuracy, we chose for these experiments a configuration that included 7000 senones.

Our baseline recognition system included cepstral mean normalization. The results comparing both male and female sets with the new algorithm are shown in Table 1. Use of this algorithm results in a 5% decrease in error rate for this test set.

	Male	Female
Baseline CMN	12.3%	8.1%
Augmented CMN	11.7%	7.7%

Table 1. Comparison of the recognition error rates of the baseline system using standard CMN and the augmented CMN for the male and female segments of *si_dev5*. This data set is recorded with a close-talking microphone.

To evaluate the effectiveness of the proposed algorithm to combat mismatches in acoustical environments, we collected a database consisting of one male speaker recording 56 sentences from the *5000-word Wall Street Journal* task with a desk-top microphone in a PC environment. In Table 2 we show the performance of the proposed algorithm. Use of this algorithm in a mismatched environment resulted in more than 25% decrease in error rate.

	Desk-top mic
Baseline CMN	21.7%
Augmented CMN	15.7%

Table 2. Comparison of the recognition error rates of the baseline system using standard CMN and the augmented CMN for a 56-utterance test set recorded with a desk-top microphone.

4. DISCUSSION

CMN is currently used in many *state-of-the-art* systems because it reduces slightly the error rate for clean speech and moderately for mismatched acoustic conditions, and it is simple. We found that our Augmented CMN brings an additional 25% reduction in error rate for mismatched conditions and a more modest 5% for clean matched conditions.

One advantage of using sentence-based normalization methods such as CMN or Augmented CMN instead of adaptation methods, is that additional improvement in recognition accuracy can be obtained even for clean conditions. Presumably, this improvement occurs because some of the variability present in the training data is reduced, since the training utterances are also normalized.

REFERENCES

[1] Acero, A. “Acoustical and Environmental Robustness in Automatic Speech Recognition”. *Kluwer Academic Publishers*. 1993.

[2] Gales M. and Young S. “An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise”. *ICASSP92*.

[3] Huang X., Acero A., Alleva F., Hwang M., Jiang L. and Mahajan M. “Microsoft Windows Highly Intelligent Speech Recognizer”. *ICASSP95*.

[4] Juang B. H. “Speech Recognition in Adverse Environments”. *Computer, Speech and Language*, vol 5, pp 275-294. 1991.

[5] Liu F., Stern R., Huang X. and Acero A. “Efficient Cepstral Normalization for Robust Speech Recognition”. *Proceedings of ARPA Human Language Technology Workshop*, March 1993.

[6] Moreno P., Raj B., Gouvea E. and Stern R. “Multivariate Gaussian-Based Cepstral Normalization for Robust Speech Recognition. *ICASSP95*.

[7] Neumeyer L. and Weintraub M. “Robust Speech Recognition in Noise Using Adaptation and Mapping Techniques”. *ICASSP95*.

[8] Sankar A. and Lee C.H. “Robust Speech Recognition Based on Stochastic Matching”. *ICASSP95*.