APPROACHES TO MICROPHONE INDEPENDENCE IN AUTOMATIC SPEECH RECOGNITION

Pedro J. Moreno, Uday Jain, Bhiksha Raj, Richard M. Stern

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

ABSTRACT

This paper describes a series of cepstral-based compensation procedures that render the SPHINX-II system more robust with respect to acoustical changes in the environment. The first algorithm, RATZ (MultivaRiate gAussian based cepsTral normaliZation) requires stereo-data for computing compensation terms, and is similar in philosophy to MFCDCN [ref] (in fact MFCDCN can be thought of as a discrete case of RATZ). We also describe a second algorithm, an improved version of CDCN, that does not require stereo training data and yet achieves performance levels comparable to the RATZ and other stereo algorithms. Use of the various compensation algorithms in consort produces a reduction of error rates for SPHINX-II by as much as 20.0% percent relative to the rate achieved with cepstral mean normalization alone, in both development test sets and in the context of the 1994 ARPA CSR evaluations.

1. INTRODUCTION

Robustness with respect to environmental variability remains a continuing problem with speech recognition technology. For example, the use of microphones other than the ARPA standard Sennheiser HM-414 headset (CLSTLK) severely degrades the performance of systems like the SPHINX II system, even in a relatively quiet environments [*e.g.* 1,2]. Applications such as speech recognition in automobiles, at offices, on a factory floor, or outdoors demand an even greater degree of environmental robustness.

In this paper we describe cepstral domain algorithms that considerably improve the performance of SPHINX II in acoustically adverse environments. Furthermore we discuss their application to the microphone independence (Spoke 5) task in the 1994 ARPA CSR evaluation.

Traditionally algorithms such as MFCDCN, which has performed well in previous evaluations, have depended on stereo data for training of compensation vectors. On the other hand algorithms such as CDCN, which do not require stereo training data, have either been computationally expensive or have not improved performance appreciably. An example of the latter is CMN which is now part of the SPHINX II baseline system.

The two new algorithms that we describe are RATZ and N-CDCN. RATZ requires stereo training data while N-CDCN does not. Both algorithms achieved a similar level of performance in the Spoke 5 evaluation. The algorithms are described in greater detail in the next section.

In Section 2 we describe the various novel environmental compensation algorithms that were used. In Section 3 we describe the performance of each of these algorithms on the development set. In Section 4 we describe their performance on the 1994 CSR evaluation set. Finally, in section 5 present our conclusions

2. NEW ENVIRONMENTAL COMPENSATION ALGORITHMS

In this section we describe the novel approaches that were attempted on the tests. In addition the well known MFCDCN [ref.] algorithm was also tried on the development set.

2.1. RATZ

RATZ (multivaRiate gAussian based cepsTral NormaliZation) is a new algorithm used to compensate speech for the effects of unknown recording environments. RATZ combines some of the best features of empirical compensation techniques such as MFCDCN and approaches which use structural models of degradation like CDCN. It performs compensation using empirical comparisons, like MFCDCN, but uses the more formal style of representation of densities that is used in CDCN.

RATZ assumes that the statistics of the clean speech cepstra can be defined hierarchically: the SNRs of the speech frames are defined to have a mixture gaussian distribution, and the cepstra under each SNR level are further assumed to have a mixture gaussian distribution. The parameters describing this distribution, namely the means, variances, priors, and the means, variances and priors for the SNR distributions are computed using generic EM methods. A subset of the WSJ1 and WSJ0 speech corpora was used for training these parameters.

The corresponding distribution for the noisy speech is computed by assuming invariance of the a posteriori probabilities of the gaussian modes across the clean and noisy speech distributions and computing them on the clean speech counterparts of each of the noisy speech vectors.

Noisy speech observations are compensated by using an MMSE technique to shift them back to the clean speech statistics.

Training: Fifteen environments of stereo speech distributed by NIST in previous years were used to compute environment specific statistics. The effect of the environment in the speech statistics is assumed to be reflected as a shift in the means and scalings of the variances of the distribution of clean speech as described by RATZ. Additionally a 16th environment with zero shifts for the correction statistics is generated to account for close-talking microphones**.**

Testing: Given a noisy utterance, parametrized by a sequence of cepstral vectors, the previously learned statistics describing the clean speech, and the correction statistics for each of the 16 known environments, RATZ applies a MMSE technique to estimate the unobserved clean vectors. The estimated clean speech vectors are decomposed as the noisy speech vectors minus a correction factor. RATZ applies each of the correction statistics serially and takes the top 3 most likely environments. The final correction factors are a weighted sum of the top 3 correction factors.

2.2. N-CDCN

N-CDCN (New-Codebook Dependent Cepstral Normalization) is a new and improved version of its predecessor, CDCN. The older CDCN, though able to achieve a respectable amount of error reduction [ref.] without requiring stereo data, had not found use in current systems due to the disadvantage of requiring task dependant features that had to be pre-computed. This new version alleviates this problem while retaining CDCN's high level of performance compensating for the combined effects of additive noise and unknown channel distortion. Note that CDCN does not require stereo data to achieve noise/channel compensation. Compensation is achieved on a sentence by sentence basis.

The algorithm follows these steps:

- Assumes a structural model of the degradation. Clean speech is contaminated by additive stationary noise after being filtered by an unknown linear channel. This can be expressed as:

$$y[n] = x[n] * h[n] + m[n]$$

- A statistical description of the clean speech acoustic cepstral space, as parametrized by a mixture of multivariate gaussians, is estimated by EM methods. A subset of the WSJ1 and WSJ0 speech corpora was used for this purpose.
- Given a noisy utterance, parametrized by a sequence of cepstral vectors, and given the previously learned statistics describing the clean speech, N-CDCN iteratively estimates noise and channel vectors that maximize the likelihood of observing the noisy utterance.
- Finally, a MMSE technique is used to estimate the unobserved clean speech vectors given the observed noisy speech along with the previously estimated noise and channel vectors and the statistics describing clean speech.

The following enhancements over the old CDCN are also implemented in N-CDCN:

- No retraining of the HMM's is needed. CDCN in its original implementation introduced the concept of "universal acoustic space". This forces all speech vectors, including the training set, to be mapped to this artificial space. N-CDCN bypasses this space by directly modelling the clean speech space.
- No prior assumptions are made about the parameters describing the statistical model of clean speech. In the original CDCN variances and prior probabilities were not computed and had to be empirically obtained for the database.
- The estimation formulas for the noise and channel have been greatly improved in accuracy and speed. Better initial values for the channel and noise estimates are provided.

3. PERFORMANCE IN DEVELOPMENTAL TESTING

In this and the following section we describe the results of a series of experiments that compare the recognition accuracy of the various algorithms described in Sec. 2 using the ARPA CSR Wall Street Journal task. Stereo pairs from WSJ0 and WSJ1 were used for the training corpus, and the system was tested using the utterances from secondary microphones in the 1993 development test set. This test set has a closed vocabulary of 5000 words.

3.1. Compensation Strategies

MFCDCN: This algorithm in the past has been the algorithm of choice for an environment independence task was used. We looked at possible extensions to this algorithm. Iterative MFCDCN and PDCN (Phone Dependent Cepstral Normalization) did not provide an appreciable improvement over stand alone MFCDCN.

Multiple Decoders: MFCDCN learns the correction codebooks for a limited number of environments and depends on the joint effect of these codebooks to provide correction vectors for any unseen environments. The usual procedure is the interpolation of the correction vectors from the top three environments that best match the previously unseen environment. We attempted a new algorithm to deal with the unseen environment problem. Test utterances from a new environment were processed using just one correction codebook and a set of hypothesis files was produced. This process was repeated for all environments for which correction codebooks existed. The decoder then performed a final pass on the hypothesis files and generated a final transcription with the best combination of acoustic and language model score, for each utterance.

The gain using this procedure while appreciable for some speakers, was not significant. Furthermore this procedure is computationally very expensive requiring the use of multiple decoder runs.

RATZ and N-CDCN: The two new algorithms recently developed were used to process the development set data and the results obtained were comparable to those obtained my MFCDCN. However since RATZ is a generalization of the MFCDCN framework we believe that it is more adaptable and robust in the case of environments that do not exist in the training corpus. N-CDCN achieved a performance level equal to that obtained using MFCDCN or RATZ. Additionally it had the advantage of not requiring stereo training data or any retraining of the clean acoustic space. The use of these algorithms and the results obtained will be discussed in the next section.

4. PERFORMANCE USING THE 1994 CSR EVALUATION DATA

We summarize in this section the results of experiments using the 1994 ARPA CSR Spoke 5 test set obtained using SPHINX II. The test data consisted of speech recorded on ten microphones that had not been used previously in ARPA evaluations. Microphones were not shared across the test speakers.

The test utterances were processed with RATZ correction codebooks which were trained on stereo data provided by ARPA. At this stage the decoder generated hypothesis files for each utterance using four sets of models. The four sets used were two male and two female models. A final pass was used to generate the transcription for each utterance by balancing the acoustic and language model score.

A similar set of procedures was followed for processing the data using N-CDCN.

.

Compensation method	**CONDITIONS (% error rate)**			
	P0	**C1**	**C2**	**C3**
RATZ	10.2	12.7	6.9	7.0
N-CDCN	10.1	12.7	7.5	7.0

Table 1: Word error rates for Spoke6 evaluation using SPHINX-II.

5. SUMMARY AND CONCLUSIONS

In this paper we describe a number of procedures that improve the recognition accuracy of the SPHINX-II system in unknown acoustical environments. We find that our new blind compensation algorithm, N-CDCN, has performed as well as stereo based compensation algorithms. We believe that the reason for this similarity in performance is due to the

inability of stereo based algorithms in learning limitation of stereo based algorithms in capturing all possible environments given a previous finite number of environments in which they are trained.

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 and the rest of the speech group for their contributions to this work.

REFERENCES

1. Juang, B.-H., “Speech Recognition in Adverse Environments”, *Computer Speech and Language*, **5**:275-294, 1991.
2. Acero, A., *Acoustical and Environmental Robustness in Automatic SPeech Recognition*, Kluwer Academic Publishers, Boston, MA, 1993.
3. Liu, F.H., Acero, A., and Stern, R.M., “Effective Joint Compensation of Speech for the Effects of Additive Noise and Linear Filtering”, *ICASSP-92*, pp. 865-868, March 1992.
4. Liu, F-H., “Environmental Adaptation for Robust Speech Recognition”. Ph.D. Thesis, ECE Department, CMU, July 1994.
5. Moreno, P.J., Raj, B., Gouvêa, E. and Stern, R.M., “Multivariate Gaussian Based Cepstral Normalization”, to appear in *ICASSP-95*, Detroit, May 1995.