TECHNICAL RESEARCH REPORT

Comparison Studies of Several Microphone Robustness Techniques

by M.K. Sönmez, Y.H. Kao,
P.K. Rajasekaran, and J.S. Baras

T.R. 94-30

Sponsored by
the National Science Foundation
Engineering Research Center Program,
the University of Maryland,
Harvard University,
and Industry

Comparison Studies of Several Microphone Robustness Techniques

M.K. Sönmez* Y.H. Kao P.K. Rajasekaran J.S. Baras†
Texas Instruments Incorporated
Dallas, TX 75265
{sonmez,yhkao, raja}@csc.ti.com, {kemal,baras}@src.umd.edu

Abstract

We study the effectiveness of various microphone robustness techniques from the viewpoint of speech recognition, utilizing the ARPA-sponsored Wall Street Journal (WSJ) data base[1]. Two of the techniques considered are being introduced in this paper: two cepstral normalization algorithms utilizing the artificial neural network techniques Self Organizing Map (SOM) and Learning Vector Quantization (LVQ). The algorithms obtained are low-complexity non-parametric counterparts of the parametric approaches Codeword-dependent Cepstral Normalization (CDCN) and Fixed CDCN (FCDCN). The other techniques considered are Cepstral Mean Normalization (CMN), RASTA, SNR-dependent Cepstral Normalization (SDCN), Interpolated SDCN (ISDCN), CDCN, FCDCN; some of these techniques require one or more of the following information: stereo data, SNR estimate, single microphone data for adaptation, and knowledge of the microphone used for the specific data under test. We determine the effectiveness in several ways: (i) scattergram plot of the speech frame parameter vector (usually a cepstral vector), (ii) adjusted deviation ratio, measured from scattergram, and (iii) correctness of classifying a test vector into a vector code book. All these measures have direct correlation with speech recognition performance, which will be measured with experiments to be conducted.

1 Introduction

Robustness has proved to be a very important concern for speech recognition systems in recent years as studies have shown that even the performance of speaker independent systems are greatly affected by the type of microphone or the acoustical environment they operate in when these are different from the ones used during training.[2] Many techniques have emerged to improve the robustness of speech recognition systems; in this study we limit our attention to techniques that operate in the cepstral domain. These range from simple cepstral filtering techniques such as CMN and RASTA to more sophisticated and computationally intensive methods such as CDCN. We introduce two cepstral normalization techniques which make use of data-driven transformations of vector quantization codebooks. Specifically, these transformations are realized by the SOM [3] and the LVQ [3] algorithms in unsupervised and supervised manner, respectively. Our goal is to compare various cepstral normalization techniques including the artificial neural network techniques which are introduced in this paper on a large enough database. The metrics that we use, the scattergram, the deviation ratio, are functions of the statistics of the cepstral vectors and have direct correlation with speech recognition performance. The comparison obtained this way is recognizer independent: however, we plan to conduct recognition experiments as well.

2 Training and Test Corpora

The speech data is a subset of the WSJ database. It consists of 10 sentences from 30 speakers, half female half male, recorded with the Sennheiser HMD-414 (CLSTLK) and the Crown PZM6FS omnidirectional desktop microphone in a stereo manner. The training data consists of 20 speakers, in stereo form for algorithms that require stereo data. The testing data is divided into two, first part consisting of 5 speakers is utilized to adjust algorithm parameters and to prevent overtraining though it is not seen

*M.K. Sönmez is with Texas Instruments Incoporated and the University of Maryland.

†Martin Marietta Chair in Systems Engineering at the University of Maryland

during the training. The second part is the evaluation set, unseen till all the development is complete, on the basis of which the algorithms are compared. The stereo form of the testing data is utilized to generate the scattergrams and compute scattergram-dependent similarity metrics such as the deviation ratio.

3 Cepstral Normalization Techniques

The techniques considered are:

1. Cepstral mean normalization
2. RASTA [4]
3. SNR-Dependent Cepstral Normalization [5]
4. Blind SDCN [6]
5. Interpolated SDCN [5, 7]
6. Codeword-dependent Cepstral Normalization [5]
7. Fixed CDCN [7]
8. Self-Organizing Map
9. SOM and Learning Vector Quantization

In the next section, we describe cepstral normalization by SOM and LVQ. Other techniques are described adequately in the literature.

4 Self-Organizing Map and Learning Vector Quantization

Self-Organizing Map and Learning Vector Quantization are adaptive vector quantization algorithms which offer non-parametric alternative techniques to parametric codeword-dependent cepstral normalization algorithms CDCN and Fixed CDCN, respectively. In CDCN, the existence of a universal acoustic space which is the distribution of speech frames under a normalized clean environment is assumed and the environmental parameters of the transformation of vector quantization codebooks from the universal space to the current environment are estimated by maximum likelihood. Due to the estimation involved, the computational complexity of CDCN is high. In SOM, the transformation is obtained via a data-driven learning process which starts from the training codebook and gradually transforms it into the appropriate codebook for the current environment using SNR-dependent local modifications upon which a set of correction vectors are computed simply as the difference between two codebooks. The algorithm is computationally very low-cost but may require a longer adaptation period than CDCN. The FCDCN algorithm and the LVQ fine-tune the CDCN and the SOM, respectively with supervised training data. Due to the dependence on SNR, our implementation of SOM and LVQ have siginificant modifications over Kohonen's algorithms [3]. Specifically, we use the topology induced by the Euclidean distance on the cepstral space to define the neighborhoods for the training codebook. The learning rate is a decaying function over frames which is dependent on the frame SNR. We are also experimenting with product codebooks, for example, making the correction on only the first few cepstral coefficients thereby reducing the dimension of the SOM. The correction of the first few coefficients have been observed to account for almost all of the improvement in recognition performance [5].

5 Experiments and Preliminary Results

The full experiment which is still in progress, will consist of comparisons of supervised and unsupervised methods listed above and most importantly shed light on the pros and cons of the parametric and non-parametric approaches such as adaptation speed, computational complexity, robustness. So far, preliminary experiments have been conducted which compare the SOM and LVQ with cepstral filtering techniques. These experiments have demonstrated a significant performance improvement over cepstral filtering. The results are summarized in Fig. 1 and Table 1. Fig. 1 shows the scattergrams of the third cepstral coefficient, with various techniques. Table 1 lists adjusted deviation ratios [8, 9] for the cepstral coefficients 2 and 3. The adjusted deviation ratio is a quantification of the scattergram, i.e. the sum of the distances of the points to the line $y = x$ normalized by their variance. The apparent improvements should result in a higher recognition performance. A recognition experiment will be carried out later on. We have received the code for CDCN from CMU and now are in the process of

setting up the CMU algorithms for the comparison experiment.

Technique	$DR(\mathbf{c}[2])$	$DR(\mathbf{c}[3])$
No normalization	1.0	1.0
RASTA	0.92	0.13
CMN	0.99	0.11
SOM	0.86	0.08
LVQ	0.59	0.07

Table 1: Adjusted deviation ratios

6 Conclusion

We have described an experimental effort to compare the performances of cepstral normalization algorithms, two of which are introduced in this work. The SOM and LVQ based normalization algorithms are non-parametric counter-parts to the well-known CDCN and FCDCN algorithms. The experiment will therefore emphasize the pros and cons of parametric and non-parametric approaches as well as the performance gain obtained by these algorithms over simple cepstral filtering at some computational cost.

References

[1] Wall Street Journal ref.

[2] Acero, A. and Stern, R.M., "Environmental Robustness in Automatic Speech Recognition", *ICASSP-90*, April 1990, pp. 849-852.

[3] Kohonen, T., "The Self-organizing Map", Proc. of the IEEE, v.78, no. 9, pp. 1464-1480.

[4] Hermansky, H., Morgan, N., Bayya, A., Kohn, P., "Compensation for the Effect of the Communication Channel in Auditory-Like Analysis of Speech (RASTA-PLP)", *Proc. of the Second European Conf. on Speech Comm. and Tech.*, September, 1991.

[5] Acero, A., *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Ph. D. Thesis, Carnegie-Mellon University, September 1990.

[6] Liu, F.H., Acero, A., and Stern, R.M., "Efficient Joint Compe nsation of Speech for the Effects of Additive Noise and Linear Filtering",

[7] Acero, A. and Stern, R.M., "Robust Speech Recognition by Normalization of the Acoustic Space", *ICASSP-91*, May 1991, pp. 893-896.

[8] Kao, Y.H., Baras, J.S., Rajasekaran, P.K., "Robustness Study of Free-Text Speaker Identification and Verification", *ICASSP 93*, Minneapolis, MO, April 1993, II-379.

[9] Kao, Y.H., *Robustness Study of Free-Text Speaker Identification and Verification*, Ph. D. Thesis, University of Maryland, 1992.

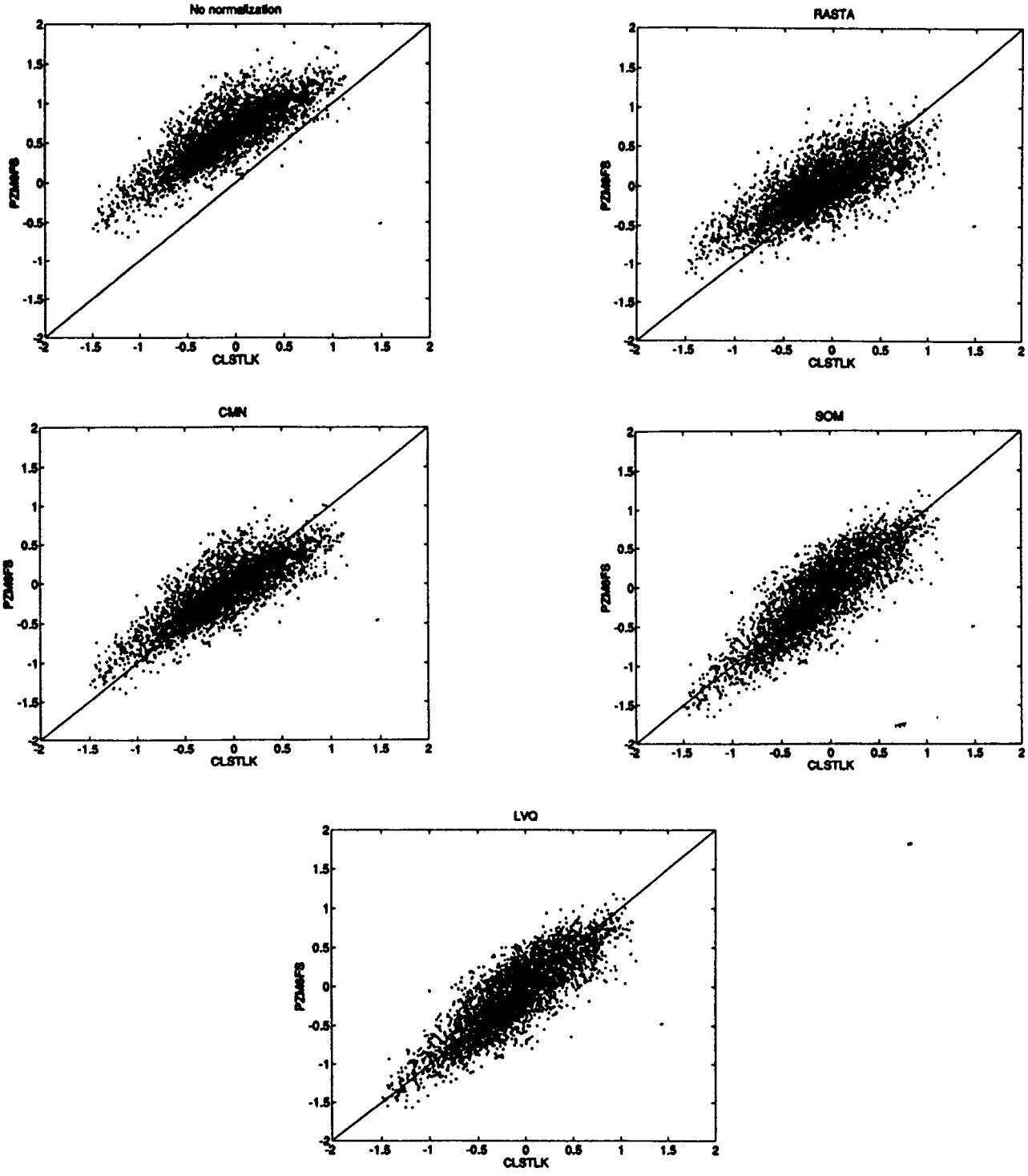

Figure 1: Scattergrams for the third cepstral coefficient with various normalization techniques