FEATURE COMPENSATION EMPLOYING ONLINE GMM ADAPTATION FOR SPEECH RECOGNITION IN UNKNOWN SEVERELY ADVERSE ENVIRONMENTS

Wooil Kim and John H. L. Hansen

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, Texas, USA
{wikim,John.Hansen}@utdallas.edu, http://crss.utdallas.edu

ABSTRACT

This study proposes an effective feature compensation method to improve speech recognition in real-life speech conditions, where (i) severe background noise and channel distortion simultaneously exist, (ii) no development data is available, and (iii) clean data for ASR training and the latent clean speech in the test data are mismatched in the acoustic structure. The proposed feature compensation method employs an online GMM adaptation procedure which is based on MLLR, and a minimum statistics replacement technique for non-speech segments. The DARPA Tank corpus is used for performance evaluation, which includes severe real-life noisy conditions. The clean Broadcast News (BN) corpus is used for training the speech recognition system in this study. Experimental results show that the proposed feature compensation scheme outperforms GMM-based FMLLR and the ETSI AFE for DARPA Tank data, achieving a +5.56% relative improvement compared to FMLLR. These results demonstrate that the proposed feature compensation scheme is effective at improving speech recognition performance in unknown real-life adverse environments.

***Index Terms*—** robust speech recognition, feature compensation, GMM adaptation, minimum statistics replacement, DARPA Tank corpus

1. INTRODUCTION

One of the primary factors degrading the performance of speech recognition systems in actual environments is acoustic mismatch between training and operating conditions of the speech recognizer. Background noise, microphone mismatch, communication channel, and speaker variability are major sources of such mismatch. Recently, as mobile devices such as smart phones become more popular, speech recognition technology via mobile platforms is more challenging, since a range of background noise and time-varying channel effects make recognition conditions more difficult. This paper focuses on an effective feature compensation scheme for robust speech recognition in unknown severely adverse environments.

To minimize the acoustic mismatch, extensive research has been conducted in recent decades, which includes many types of speech/feature enhancement methods such as Spectral Subtraction, Cepstral Mean Normalization (CMN), and variety of feature compensation schemes [1]-[9]. Various model adaptation techniques have been successfully employed such as the Maximum A Posteriori (MAP), Maximum Likelihood Linear Regression (MLLR), and Parallel Model Combination (PMC) [10]-[12]. Recently, missing-feature methods have shown promising results [13]-[15].

As real-life conditions for speech recognition become more adverse, the front-end technique is required to be more effective in addressing unknown severely noisy environments. In real-life environments, the speech signal is corrupted in a more complicated way by coupling the background noise and channel distortion effects. In addition, development data is often unavailable, which can be used for model adaptation that reflect the test condition, so an effective online adaptation over the input speech is generally required. Some front-end techniques have verified their effectiveness in a restricted frame work condition[1], where the clean training data and the latent clean speech in the test data were obtained in an identical recording condition and they share the same vocabulary configuration. A series of our preliminary experiments showed some of those front-end techniques were not effective when the acoustic structures of the clean speech used for training and generating test data are mismatched.

In this paper, we propose an effective feature compensation method employing an online Gaussian Mixture Model (GMM) adaptation. A noise-corrupted speech GMM is obtained by the MLLR adaptation technique, and it is used for clean speech reconstruction. Minimum statistics are deter-

This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen.

[1] Clean speech samples are collected in an identical condition, and a part of them are used as training data and others used for generating the noise-corrupted test speech.

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

mined during the input speech and then replaced with the non-speech segments to further improve the reconstructed speech. The proposed feature compensation method is evaluated over the DARPA Tank corpus, using a speech recognizer trained on the Broadcast News (BN) corpus [16] to observe the effectiveness in a completely unknown acoustic environment. The DARPA Tank data includes real-life severe adverse conditions which make speech recognition highly challenging.

This paper is organized as follows. Sec. 2 presents details of the proposed feature compensation method including the online GMM adaptation, clean speech reconstruction, and minimum statistics replacement technique. Representative experimental procedures and their results are presented and discussed in Sec. 3. Finally, in Sec. 4 we draw the main conclusions of our work.

2. FEATURE COMPENSATION EMPLOYING ONLINE GMM ADAPTATION

The proposed feature compensation method employs an online GMM adaptation method over the noise-corrupted input utterance. The noise-corrupted speech GMM is obtained via adaptation and used for clean speech reconstruction. As an initial stage, a K-component GMM representing the clean speech signal $\mathbf{x}$ in the cepstral domain is estimated off-line from the clean training data, which is given by,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{1}$$

2.1. Step 1: Online GMM adaptation

In the proposed method, the conventional MLLR adaptation technique [11] is employed for online GMM adaptation over the input speech. Here, MLLR adaptation for a GMM is briefly described. Only the mean vectors are updated by model adaptation in this study. The kth mean vector is assumed to be transformed using the following equation,

$$\tilde{\boldsymbol{\mu}}_k = \mathbf{A}\boldsymbol{\mu}_k + \mathbf{b}, \tag{2}$$

where $\mathbf{A}$ is a regression matrix and $\mathbf{b}$ is an additive bias term. Eq. (2) can be extended by introducing an extended mean vector $\mathbf{m}_k = [\mathbf{1}, {\boldsymbol{\mu}_k}^T]^T$ as follows:

$$\tilde{\mathbf{m}}_k = \mathbf{W}\mathbf{m}_k, \tag{3}$$

where $\mathbf{W} = [\mathbf{b}, \mathbf{A}]$.

As is well known, the extended transform matrix $\mathbf{W}$ can be obtained by equating the partial derivative to zero so as to maximize the auxiliary function $Q(\lambda, \tilde{\lambda})$. The resulting equation is as follows:

$$\sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_k(t)\boldsymbol{\Sigma}_k^{-1}\mathbf{y}(t)\mathbf{m}_k^T = \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_k(t)\boldsymbol{\Sigma}_k^{-1}\mathbf{W}\mathbf{m}_k\mathbf{m}_k^T, \tag{4}$$

where $\mathbf{y}(t)$ is the noise-corrupted input speech with a total number of frames T, and $\gamma_k(t)$ is the posterior probability of being in the kth Gaussian component at time t. By transforming the mean vector, a noise-corrupted speech GMM is obtained as $\{\omega_k, \tilde{\boldsymbol{\mu}}_k, \boldsymbol{\Sigma}_k\}$.

2.2. Step 2: Clean speech reconstruction

In the proposed method, a constant bias transform of the mean parameters of the speech model in the cepstral domain is assumed under the noisy environment. This is the assumption generally taken by other data-driven methods [17], and is represented as follows,

$$\tilde{\boldsymbol{\mu}}_k = \boldsymbol{\mu}_k + \mathbf{r}_k. \tag{5}$$

The bias term $\mathbf{r}_k$ is used for clean speech reconstruction, which is based on the Minimum Mean Squared Error (MMSE) estimator as follows [9][17],

$$\tilde{\mathbf{x}}(t) = \int_{\mathcal{X}} \mathbf{x}p(\mathbf{x}|\mathbf{y}(t))d\mathbf{x} \cong \mathbf{y}(t) - \sum_{k=1}^{K} \mathbf{r}_k\, p(k|\mathbf{y}(t)). \tag{6}$$

It might be considered that the linear transform of Eq. (2) for mean vector can be used for the clean speech reconstruction, instead of using another assumption of the bias transform as Eq. (5). However, Eq. (2) is obtained with the criterion which maximizes the likelihood for the updated (i.e., transformed) model parameters in the MLLR algorithm. Therefore, to apply the transform matrix to the feature space does not guarantee the maximum likelihood. Our proposed method can be compared to the feature space MLLR (FMLLR) [18], where the input feature vector is transformed instead of the model parameters. As such, the performance of the proposed feature compensation method will be compared to the FMLLR in Sec. 3.

2.3. Minimum statistics replacement for non-speech segments

We found that to replace non-speech segments with minimum statistics consistently improves speech recognition performance for severely noise-corrupted speech. The minimum statistics are determined over the entire duration T of the input speech in the log-spectral domain using the following equation,

$${\alpha_n}^{\{l\}} = \min\{y_n^{\{l\}}(1), y_n^{\{l\}}(2), \ldots, y_n^{\{l\}}(T)\}, \tag{7}$$

where $y_n^{\{l\}}(t)$ is the nth element of the log-spectrum $\mathbf{y}^{\{l\}}(t)$, which can be obtained by an inverse Discrete Cosine Transform (DCT) of the cepstrum (i.e., $\mathbf{y}^{\{l\}} = \mathbf{C}^{-1}\mathbf{y}$). The obtained minimum statistics vector $\boldsymbol{\alpha}^{\{l\}}$ is converted to the cepstral domain as $\mathbf{C}\boldsymbol{\alpha}^{\{l\}}$, and then replaced for frames which are detected as non-speech segments in the cepstral domain.

This study employs a simple non-speech detection method utilizing single Gaussian models for speech and non-speech, which are estimated from every input speech.

3. EXPERIMENTAL RESULTS

The Broadcast News (BN) speech corpus (*F0*: baseline broadcast speech and *F1*: spontaneous broadcast speech) [16] was used for training the Hidden Markov Model (HMM) of the speech recognizer. A total of 16.8 hours of speech were used, which consists of 18049 utterances. The speech samples were down-sampled to 8 kHz to be the same as the DARPA Tank corpus which were used for performance evaluation in this study. The SPHINX3 [19] was employed as our HMM based speech recognizer, and each HMM represents a tri-phone which consists of 3 states with an 8-component GMM per state, which is tied with 4120 states. A conventional Mel-Frequency Cepstral Coefficients (MFCC) feature front-end is employed in the experiment, which was suggested by the European Telecommunication Standards Institute (ETSI) [20]. An analysis window of 25 msec in duration is used with a 10 msec skip rate for 8 kHz speech data. The computed 23 Mel-filterbank outputs are transformed to 13 cepstrum coefficients including c0 (i.e., c0-c12). The first and second order time derivatives are also added during decoding, so the feature vector used for recognition is 39-dimensional.

Table 1 presents the performance evaluation over the DARPA Tank corpus. The DARPA Tank data used in this study includes 10 speakers' conversations (2.5 hours with 1489 utterances) during military field training games in various situations (tank, jeep, personnel carriers, etc.) using vehicles and monitors outfitted with laser tag transmitter/receiver technology. The data reflects various types of actual severe background noise and communication channel effects. The vocabulary consists of 1190 words and a trigram is adapted on the DARPA Tank data transcription using a Broadcast News language model as an initial model.

Here, the proposed system was also compared to the existing conventional front-end algorithms using the BN speech recognition system. Spectral Subtraction (SS) [21] combined with CMN was selected as one of the conventional algorithms. This represents one of the most commonly used techniques for additive noise suppression and removal of channel distortion respectively. We also evaluated a feature compensation method, VTS (Vector Taylor Series) for performance comparison, where the noisy speech GMM is adaptively estimated using the EM algorithm over each test utterance [17]. The Advanced Front-End (AFE) algorithm developed by ETSI was also evaluated as one of the state-of-the-art methods, which contains an iterative Wiener filter and blind equalization [22]. Here, FMLLR [18] was also compared, which was implemented as a GMM-based version for a fair comparison to the proposed feature compensation method.

Table 1. Recognition performance in WER (%) of the proposed system for DARPA Tank data with relative improvement to FMLLR.

	WER	(Relative)
No processing	82.27	
SS + CMN	54.26	
VTS + CMN	56.29	
AFE + CMN	54.83	
GMM-FMLLR + CMN	53.07	
OGAFC + CMN	**51.11**	**(+3.69)**
OGAFC + CMN + MSR	**50.12**	**(+5.56)**

A series of experiments showed that a combination of a target front-end system and CMN provides more performance improvement in WER than employing either technique individually. It is also noted that the HMM of the ASR system was trained over the CMN-processed BN corpus, except for the "No processing" and AFE+CMN cases. The ETSI AFE showed the best performance when using the speech recognizer trained over the AFE-processed training data, therefore the ASR system for the AFE+CMN was trained over the AFE+CMN processed BN corpus. A clean speech GMM with 1024 components was obtained by training over the same CMN-processed BN corpus, and used for the proposed method and the GMM-based FMLLR.

For the DARPA Tank data, the GMM-based FMLLR showed the best performance as 53.07% in WER among conventional methods. The proposed system (OGAFC+CMN) results in 51.11% and 50.12% without/with the minimum statistics replacement technique, showing +3.69% and +5.56% relative improvement respectively compared to FMLLR+CMN. It can be seen that the proposed system achieves significantly better performance compared to the ETSI AFE algorithm for the DARPA Tank evaluation. The evaluation results shown in Table 1 demonstrate that the proposed feature compensation method is highly effective at improving speech recognition performance in realistic conditions, where convolutional distortion and background noise are simultaneously present and no a prior knowledge of the environments is available.

The plots in Fig. 1 present distributions of the 0th MFCC component of the training data (BN) and test data (DARPA Tank). Here it can be seen that the c0 distribution of the DARPA data with the proposed OGAFC+CMN applied becomes more matched to the BN data compared with the case of AFE+CMN. This result is also consistent with speech recognition WER performance seen in Table 1, where OGAFC+CMN outperforms AFE+CMN.

4. CONCLUSIONS

This study has proposed an effective feature compensation method to improve speech recognition in unknown severely

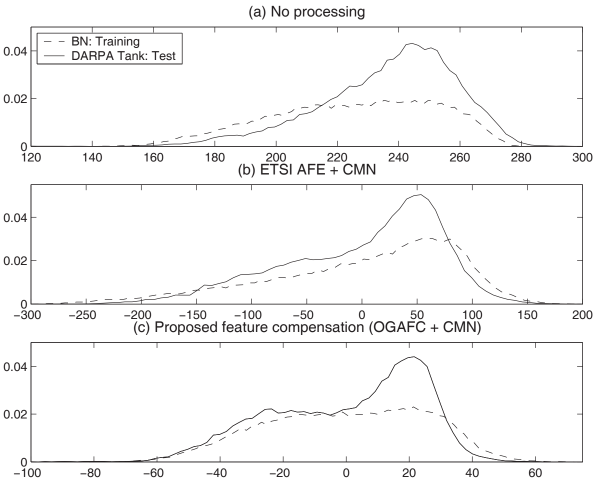

Fig. 1. *Distributions of the 0th MFCC feature component (i.e., c0) for BN and DARPA Tank data with (a) no processing, and applying (b) AFE+CMN and (c) the proposed OGAFC+CMN.*

adverse environments. The proposed feature compensation method employed an online GMM adaptation procedure which was based on MLLR, and minimum statistics replacement technique for non-speech segments. As a front-end to speech recognition, the proposed method was evaluated on the DARPA Tank database which includes severe real-life noisy conditions. Experimental results showed that the proposed feature compensation scheme outperformed GMM-based FMLLR and ETSI AFE for the DARPA Tank data. These results demonstrated that the proposed front-end scheme is highly effective at improving speech recognition performance in real-life speech recognition conditions, where (i) severe background noise and channel distortion simultaneously exist, (ii) no development data is available, and (iii) clean data for ASR training and the latent clean speech in the test data are mismatched in the acoustic structure.

5. REFERENCES

[1] S.F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol.27, pp.113-120, 1979.

[2] Y. Ephraim and D. Malah, “Speech Enhancement Using Minimum Mean Square Error Short Time Spectral Amplitude Estimator,” *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol.32, no.6, pp.1109-1121, 1984.

[3] J.H.L. Hansen and M. Clements, “Constrained Iterative Speech Enhancement with Application to Speech Recognition,” *IEEE Trans. on Signal Proc.*, vol.39, no.4, pp.795-805, 1991.

[4] J.H.L. Hansen, “Morphological Constrained Enhancement with Adaptive Cepstral Compensation (MCE-ACC) for Speech Recognition in Noise and Lombard Effect,” *IEEE Trans. on Speech and Audio Proc.*, vol.2, no.4, pp.598-614, 1994.

[5] P.J. Moreno, B. Raj, and R.M. Stern, “Data-driven Environmental Compensation for Speech Recognition: A Unified Approach,” *Speech Communication*, 24(4), pp.267-285, 1998.

[6] N.S. Kim, “Feature Domain Compensation of Nonstationary Noise for Robust Speech Recognition,” *Speech Communication*, 37, pp.231-248, 2002.

[7] V. Stouten, H. Van hamme, and P. Wambacq, “Joint Removal of Additive and Convolutional Noise with Model-based Feature Enhancement,” *ICASSP2004*, pp.949-952, 2004.

[8] A. Sasou, T. Tanaka, S. Nakamura, and F. Asano, “HMM-Based Feature Compensation Methods: an Evaluation Using the Aurora2,” *ICSLP2004*, pp.121-124, 2004.

[9] W. Kim and J.H.L. Hansen, “Feature Compensation in the Cepstral Domain Employing Model Combination,” *Speech Comm.*, 51(2), pp.83-96, 2009.

[10] J.L. Gauvain and C.H. Lee, “Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,” *IEEE Trans. on Speech and Audio Proc.*, vol.2, no.2, pp.291-298, 1994.

[11] C.J. Leggetter and P.C. Woodland, “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs,” *Computer Speech and Language*, 9, pp.171-185, 1995.

[12] M.J.F. Gales and S.J. Young, “Robust Continuous Speech Recognition Using Parallel Model Combination,” *IEEE Trans. on Speech and Audio Proc.*, vol.4, no.5, pp.352-359, 1996.

[13] M. Cook, P. Green, L. Josifovski, and A. Vizinho, “Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data,” *Speech Comm.*, 34(3), pp.267-285, 2001.

[14] B. Raj and R.M. Stern, “Missing-Feature Approaches in Speech Recognition,” *IEEE Signal Processing Magazine*, vol.22, no.5, pp.101-116, 2005.

[15] W. Kim and J.H.L. Hansen, “A Novel Mask Estimation Method Employing Posterior-Based Representative Mean Estimate for Missing-Feature Speech Recognition,” *IEEE Trans. on Audio, Speech and Language Proc.*, vol.19, no.5, pp.1434-1443, July 2011.

[16] D. Graff, “An overview of Broadcast News corpora,” *Speech Comm.*, 37(1), pp.15-26, 2002.

[17] P.J. Moreno, *Speech recognition in noisy environments*, Ph.D. Thesis. Carnegie Mellon University, 1996.

[18] M.J.F. Gales, “Maximum Likelihood Linear Transformations for HMM-based Speech Recognition,” *Computer Speech and Language*, vol.12, pp.75-98, 1998.

[19] http://cmusphinx.sourceforge.net

[20] *ETSI standard document*, ETSI ES 201 108 v1.1.2 (2000-04), 2000.

[21] R. Martin, “Spectral Subtraction based on Minimum Statistics,” . *EUSIPCO-94*, pp. 1182-1185, 1994.

[22] *ETSI standard document* ETSI ES 202 050 v1.1.1 (2002-10), 2002.