PCMM-BASED FEATURE COMPENSATION SCHEMES USING MODEL INTERPOLATION AND MIXTURE SHARING

Wooil Kim[1], Ohil Kwon[2] and Hanseok Ko[1]

[1]Dept. of Electronics and Computer Engineering, Korea University, Seoul, Korea
[2]Hyundai Autonet Co. Ldt., Ichon, Korea
{wikim, hsko}@korea.ac.kr, koi@haco.co.kr

ABSTRACT

In this paper, we propose an effective feature compensation scheme based on the speech model in order to achieve robust speech recognition. The proposed feature compensation method is based on parallel combined mixture model (PCMM). The previous PCMM works require a highly sophisticated procedure for estimation of the combined mixture model in order to reflect the time-varying noisy conditions at every utterance. The proposed schemes can cope with the time-varying background noise by employing the interpolation method of the multiple mixture models. We apply the 'data-driven' method to PCMM for more reliable model combination and introduce a frame-synched version for estimation of environments posteriori. In order to reduce the computational complexity due to multiple models, we propose a technique for mixture sharing. The statistically similar Gaussian components are selected and the smoothed versions are generated for sharing. The performance was examined over Aurora 2.0 and speech corpus recorded while car-driving. The experimental results indicate that the proposed schemes are effective in realizing robust speech recognition and reducing the computational complexities under both simulated environments and real-life conditions.

1. INTRODUCTION

The difference between the training and operating environments is a significant factor affecting and mostly degrading the performance of speech recognition system. Background noise and channel distortion are typical sources of such performance degradation. Putting these two environments on an equal footing is one of the most essential issues in the development of real-world applications using speech recognition technology.

Spectral subtraction, CMN (Cepstral Mean Normalization) and model-based feature compensation are some of the prominent examples of the efforts employed to bring the operating environment closer to the training environment at the pre-processing level of the speech recognition system. Another approach recently introduced is not directed at removing the noise components, but generating a speech model matched to the noisy environment at the training or decoding step. MAP (Maximum A Posteriori) and MLLR (Maximum Likelihood Linear Regression) adaptation techniques and PMC (Parallel Model Combination) method are included in this category [1-3].

In this paper, we focus on the Gaussian mixture model (GMM)-based feature compensation method of rendering improved recognition under the combined adverse conditions of additive background noise and channel distortion [4]. Among GMM-based methods, we have interests in PCMM (Parallel Combined Mixture Model)-based feature compensation method, which is known to have an effective performance without training procedure with the noisy speech database [5]. In this paper, the proposed PCMM scheme enables to cope with the time-varying noisy environments adaptively by employing the interpolation of multiple environment models. In addition, we propose a technique of mixture sharing, in order to reduce the computational complexity due to multiple models.

The paper is organized as follows. We first review the PCMM-based feature compensation scheme and identify the relevant issues in Section 2. We then describe the proposed scheme in Section 3 and 4. The representative experimental procedures and results are presented and discussed in Section 5. Finally, in Section 6, we make our concluding remarks.

2. PCMM-BASED FEATURE COMPENSATION

PCMM-based feature compensation is also based on GMM-based method, which was first proposed by Acero and soon afterwards, Moreno designed a data-driven method called RATZ (Multivariate Gaussian Based Cepstral Normalization). In RATZ, the statistical transformation of the clean speech's cepstral distribution under the noisy condition is estimated from the noisy speech database and then noisy input feature vectors are compensated using these statistics [3][4].

Distribution of clean speech cepstrum can be modeled with K Gaussian mixture as follows.

$$p(\mathbf{x}) = \sum_{k=1}^{K} c_k N(\mathbf{x}; \boldsymbol{\mu}_k, \Sigma_k) \tag{2.1}$$

It is assumed that noisy environment causes the shift in the means and the compression or expansion of the covariance matrices of cepstral distributions. Therefore, we can express the distribution of noisy speech as

$$p(\mathbf{y}) = \sum_{k=1}^{K} c_k N(\mathbf{y}; \boldsymbol{\mu}_k + \mathbf{r}_k, \Sigma_k + \mathbf{R}_k). \tag{2.2}$$

Noisy input feature vectors are then compensated based on the MMSE (Minimum Mean Squared Error) estimator.

$$\hat{\mathbf{x}}_{MMSE} \cong \mathbf{y} - \sum_{k=1}^{K} \mathbf{r}_k \, p[k \mid \mathbf{y}] \tag{2.3}$$

In the PCMM-based feature compensation, the correction factors $\mathbf{r}_k, \mathbf{R}_k$ are estimated from the noisy speech mixture model generated by combining the clean speech model and noise model, whereas RATZ obtains them from the noisy speech database training.

For generation of noisy mixture model, PMC method is employed. In PMC, the noise-corrupted speech model is generated using the clean speech model and noise model independently. Therefore, it is known to exhibit an outstanding advantage in that it does not require additional training procedure with noisy speech database [2]. In the previous work [5], the 'log-normal approximate' method was employed to implement the on-line combination of mixture model. In this paper, we apply the 'data-driven' method, which is more reliable to estimate the acoustic model, because the on-line model estimation is not employed in the proposed schemes. The 'data-driven' method can reduce the error due to approximation in combining the log-normal functions by actually adding the synthesized speech data and noise data.

Considering the variations in mean and covariance assumed in RATZ, we can compute the correction factors as follows.

$$\begin{aligned} \mathbf{r}_k &= \tilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k \\ \mathbf{R}_k &= \tilde{\Sigma}_k - \Sigma_k \end{aligned} \tag{2.4}$$

Where $\tilde{\boldsymbol{\mu}}_k$ and $\tilde{\Sigma}_k$ denote the mean and variance of kth component in the noise-corrupted mixture model, which is generated through the model combination.

In the previous work, to cope with the time-varying background noise, the noise model adaptation was applied. To reflect the adapted noise on the mixture model, we need to accomplish the mixture model combination at every utterance. It requires huge computational complexities because of conversion between linear spectrum, log spectrum and cepstral domain. In this paper, we propose more efficient PCMM-based method by employing multiple models which reflect the environments to be expected. Utilizing the multiple models estimated at off-line can be effective to compensate input feature adaptively under the time-varying noisy condition and it eliminates the on-line model combination procedure.

3. INTERPOLATION OF MULTIPLE MODELS

The feature compensation with a single noisy model assumes the environment where the recognition is going to be performed is known, enabling the algorithm to make use of previously trained correction factors. However in more realistic conditions, this might not be possible. In the method of multiple models, the posterior probabilities of each of E possible environments are estimated over the input noisy speech [4]. Utilizing the multiple models which reflect the nosy environments to be expected can be solution to coping with the time-varying noise situation.

In this paper, we modify the estimation procedure into the frame-synched version for frame-by-frame processing as follows. Given $\mathbf{Y}_1^t = [\mathbf{y}_1, \mathbf{y}_1, ..., \mathbf{y}_t]^T$, the posterior probability of the environment i over $\mathbf{Y}_1^t$ can be re-written as

$$P[i \mid \mathbf{Y}_1^t] = \frac{P[i]p(\mathbf{Y}_1^{t-1} \mid i)p(\mathbf{y}_t \mid i)}{\sum_{e=1}^{E} P[e]p(\mathbf{Y}_1^{t-1} \mid e)p(\mathbf{y}_t \mid e)} \tag{3.1}$$

where $p(\mathbf{Y}_1^{t-1} \mid i) = p(\mathbf{Y}_1^{t-2} \mid i)p(\mathbf{y}_{t-1} \mid i) = \prod_{\tau=1}^{t-1} p(\mathbf{y}_\tau \mid i)$.

From the (2.3), the clean feature can be restored frame-synchronously as follows.

$$\hat{\mathbf{x}}_{MMSE} \cong \mathbf{y}_t - \sum_{e=1}^{E} P[e \mid \mathbf{Y}_1^t] \sum_{k=1}^{K} \mathbf{r}_{e,k} P[k \mid e, \mathbf{y}_t] \tag{3.2}$$

When the background noise types are known to be a specific number such as the car-driving condition, multiple-model method can be more effective than the adaptation techniques or on-line estimation of noise components in terms of computational complexity.

If we utilize a clean mixture model as one of multiple models, the performance of recognition system can be maintained under the high SNR conditions. In the addition, the interpolation of clean model and noisy model brings the effect of adaptation under time-varying or unknown SNR conditions.

4. MIXTURE SHARING

The computation amounts in the implementation of GMM-based feature compensation methods depend dominantly on the number of the Gaussian components to be computed. Therefore, the computational complexities are proportional to the number of multiple models used for the model interpolation method.

In this paper, we propose a technique of sharing the statistically similar components among the noisy mixtures in an effort to reduce the computational load. The Gaussian components similar each other are selected in terms of Kullback-Leibler distance and then the common components for sharing are generated through smoothing the similar ones. The selection procedure is as a following pseudo algorithm.

Step 0 : $\mathbf{D} = \{d_1, d_2, ..., d_K\}, \mathbf{C}_S = \phi$

$$d_k = \sum_{e=2}^{E} kl\_dist(g_{1,k}, g_{e,k}),\ 1 \le k \le K \tag{4.1}$$

Step 1 : $k = \arg\min d_i \in \mathbf{D}$

Step 2 : $\mathbf{C}_S = \mathbf{C}_S \cup \{k\}, \mathbf{D} = \mathbf{D} - \{d_k\}$

Step 3 : if $N(\mathbf{C}_S) = K_S$, then stop.

Else, then go to **Step 1**.

Where d_k is sum of Kullback-Leibler distances of the kth Gaussian component of each environment mixture $g_{e,k}$ from the kth one of the first environment $g_{1,k}$ and $N(\cdot)$ denotes the number of elements in a set. Finally, we obtain the set $\mathbf{C}_S$ which contains K_S number of the indices of Gaussian components to be shared. The parameters of the smoothed Gaussian components for sharing are computed by following equations.

$$\tilde{\boldsymbol{\mu}}_{Sk} = \frac{1}{E}\sum_{e=1}^{E} \tilde{\boldsymbol{\mu}}_{e,k},\ \tilde{\Sigma}_{Sk} = \frac{1}{E}\sum_{e=1}^{E} \tilde{\Sigma}_{e,k},\ k \in \mathbf{C}_S \tag{4.2}$$

Therefore, we can replace the likelihood function of Gaussian components included in the set $\mathbf{C}_S$ by the sharing components.

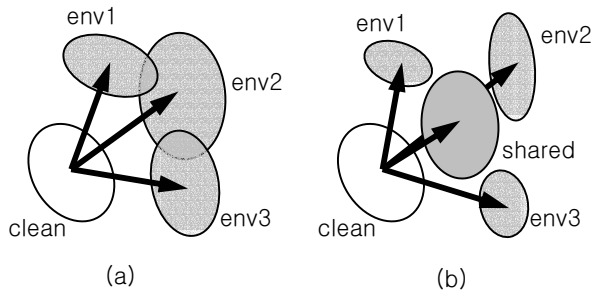

Fig 1. Concept of the mixture sharing.

$$p(\mathbf{y} \mid e,k) = \begin{cases} p(\mathbf{y} \mid \tilde{\boldsymbol{\mu}}_{Sk}, \tilde{\boldsymbol{\Sigma}}_{Sk}) & \text{if } k \in \mathbf{C}_S \\ p(\mathbf{y} \mid \tilde{\boldsymbol{\mu}}_{e,k}, \tilde{\boldsymbol{\Sigma}}_{e,k}) & \text{otherwise} \end{cases} \quad (4.3)$$

The correction factors whose indices are included in set $\mathbf{C}_S$ can be shared also.

$$\mathbf{r}_{e,k} = \begin{cases} \mathbf{r}_{Sk} = \tilde{\boldsymbol{\mu}}_{Sk} - \boldsymbol{\mu}_k & \text{if } k \in \mathbf{C}_S \\ \mathbf{r}_{e,k} = \tilde{\boldsymbol{\mu}}_{e,k} - \boldsymbol{\mu}_k & \text{otherwise} \end{cases} \quad (4.4)$$

By sharing the components, we can reduce calculation amount over $E \times K$ number of Gaussian likelihood functions to $K_S + E(K - K_S)$, it leads to computational reduction as much as $(E-1)K_S$. Fig. 1 illustrates the concept of the mixture sharing technique.

5. EXPERIMENTS AND RESULTS

5.1 Performance testing on Aurora 2.0

We followed the Aurora2.0 evaluation procedure for the performance verification. Along with all identical conditions suggested in the Aurora2.0 procedure, we used c0 instead of log-energy for the convenience of PCMM implementation.

First, we examined the performance of the baseline system in comparison with that of the existing preprocessing algorithms, with regard to the environmental robustness of their speech recognition. The typical algorithms include spectral subtraction (SS), and cepstral-mean-normalization (CMN). In spectral subtraction, the background noise is estimated using the minimum statistics method with a time delay of about 250msec. PCMM1 and PCMM2 denote the PCMM-based feature compensation methods using the combined model with the gain matched to the testing SNR conditions. In PCMM1 and PCMM2, 'log-normal' and 'data-driven' methods are employed for model combination respectively. For the clean speech modeling, the 128-Gaussian mixture is estimated using the clean training data which is identical to the data used in HMM modeling. The noise model is estimated with a single Gaussian distribution.

Table 1 shows the performance of the baseline system and the existing algorithms. From the results, PCMM with the noise model matched to the testing noisy conditions shows the superior performance to spectral subtraction and CMN or combination of these techniques. From the fact that PCMM2 outperforms PCMM1 slightly, we can see that 'data-driven' model combination is more reliable than 'log-normal' in the model estimation. In the entire experiments, we employed 'data-driven' model combination for the PCMM implementation.

Table 1. Word accuracy for baseline system to car noise condition in Aurora 2.0. (%)

	Baseline	SS	SS+CMN	PCMM1	PCMM2
Clean	98.84	98.63	98.87	98.84	98.84
20dB	96.42	97.38	97.76	97.91	97.94
15dB	87.62	93.98	95.53	97.11	97.08
10dB	61.71	81.42	86.22	93.86	93.86
5dB	26.87	50.16	59.59	81.81	82.20
0dB	10.38	17.66	25.05	53.09	53.92
-5dB	8.41	5.99	14.43	21.77	22.10
Avg.	56.60	68.12	72.83	84.76	85.00

Table 2. Word accuracy for the proposed schemes to car noise condition in Aurora 2.0. (%)

	IP	SS+IP	SS+IP+ CMN	SS+IP64 +CMN	SS+IP96 +CMN
Clean	98.84	98.84	98.87	98.87	98.87
20dB	97.88	97.88	98.18	98.03	97.32
15dB	97.35	97.17	97.91	97.73	96.93
10dB	93.29	94.57	95.35	94.75	94.42
5dB	81.78	86.46	87.56	86.28	87.06
0dB	53.80	63.41	65.37	62.90	64.54
-5dB	22.79	29.17	28.18	24.52	25.56
Avg.	84.82	87.90	88.87	87.94	88.05

Under the identical condition with baseline test, we accomplished the performance evaluation of the proposed scheme. For the interpolated PCMM, we generated three different SNR noisy mixture models, that is, 17dB, 7dB and -2dB. Considering the clean speech model as one of environments, the number of the multiple models is four. For the comparison, we examined the performance in the following combinations.

1) IP : Interpolated PCMM-based feature compensation.
2) SS+IP : Spectral Subtraction + Interpolated PCMM
3) SS+IP+CMN : Spectral Subtraction + Interpolated PCMM + Cepstral Mean Normalization
4) SS+IP64+CMN: SS + Interpolated PCMM with 64 Gaussian components shared + CMN
5) SS+IP96+CMN: SS + Interpolated PCMM with 96 Gaussian components shared + CMN

As shown in Table 2, we can see that the proposed feature compensation schemes are effective under the noisy condition and these figures present their superior performances over the existing algorithms in Table 1. The PCMM with the interpolated models shows a quite similar performance to PCMM with the SNR-matched single model in Table 1. This proves that the interpolated PCMM is effective in the adaptive feature compensation under the SNR changing environment at every utterance. We could obtain the improved results by applying spectral subtraction before PCMM processing. It shows that the interpolated PCMM is suitable to the unknown SNR situations produced by the noise subtraction. The interpolated PCMM with the proposed mixture sharing technique (IP64, IP96) shows the slightly lower performance compared to the non-sharing case. From these results, the mixture sharing is useful to reduce the computational complexities with holding the original performance.

Table 3 and Table 4 show the recognition performance over the all sets of clean-condition training and multi-condition testing in Aurora 2.0. The results in Table 4 show that the proposed feature compensation schemes have consistent performances under various kinds of additive background noises and even the channel distortions such as in Set C. Table 5 shows the relationship of the performance and computational complexity reduction. The figures in the parenthesis in the first column are the percentages of relative improvement compared to the case of non-sharing, SS-IP-CMN. The figures in the parenthesis in the second column are percentages of the number of Gaussian components to be computed to the full components. In case of 64-component sharing, we can obtain 25% reduction in computation with only 3.16% decrease in performance. When 96 components shared, 37.5% computational reduction is achieved with just 4.18% decrease of performance

5.2 Performance testing on real car-driving conditions

To verify the effectiveness of the proposed schemes in the practical situations, we accomplished the recognition testing on the speech corpus collected under real car-driving conditions. We used Car01 and CarNoise01 released by SITEC (Speech Information Technology & Industry Promotion Center) [6]. Car01 contains the Korean speech utterances recorded under car-driving with the speed of 80km/h and CarNoise01 contains the noise samples recorded at the various driving situations.

For recognition testing, we choose 548 vocabulary set in Car01, which consists of control command words in the car. 4,384 utterances recorded via a head-set microphone (channel 1) are used for clean HMM training and 1,096 utterances for noisy condition testing, which are recorded via a directional microphone located at the center of driver's sun visor (channel 4).

Table 6 shows the performance of the baseline system and conventional methods over the Car01 samples. The performances of the proposed schemes are shown in Table 7. PCMM denotes the PCMM-based method with a single model and the noise model for the model combination was estimated from the noise samples in CarNoise01, which are recorded while driving at speed of 80km/h. For interpolated PCMM (IP), we used three kinds of noise models which are estimated from the noise samples of 50km/h, 80km/h and 100km/h. From the results, we can see that the proposed schemes are also effective in the real-life situations. The results in Table 8 prove that the proposed mixture sharing technique is very useful to reduce the computational complexity under real car-driving conditions.

6. CONCLUSIONS

In this paper, we have proposed an efficient feature compensation algorithm based on PCMM method. We employed interpolation of multiple environment models and proposed a technique for mixture sharing to reduce computational complexity. The experimental results show that the proposed scheme is considerably effective in both the simulated adverse environments and real car-driving condition.

7. ACKNOWLEDGEMENT

This work was supported by grant No. 20006-302-04-2 from International Collaborative Research Program of Korea Science and Engineering Foundation.

Table 3. Word accuracy for the baseline system to all sets of clean-condition training and multi-condition testing in Aurora 2.0. (%)

	Set A	Set B	Set C	Avg.
Baseline	59.59	57.18	66.81	60.07
SS	67.70	65.00	74.85	68.05
SS-CMN	73.32	76.72	74.44	74.90

Table 4. Word accuracy for the proposed schemes to all sets of clean-condition training and multi-condition testing in Aurora 2.0. (%)

	Set A	Set B	Set C	Avg. (Relative Imp. %)
IP	85.35	83.75	70.53	81.75 (52.56)
SS+IP	86.43	84.00	78.07	83.79 (58.41)
SS+IP+CMN	87.72	85.66	83.71	85.96 (64.31)
SS+IP64+CMN	86.72	84.82	82.75	85.17 (62.28)
SS+IP96+CMN	86.33	84.63	82.59	84.90 (61.62)

Table 5. Relationship of performance and reduction in the Gaussian number on the testing of all sets in Aurora 2.0. (%)

	Relative improvement	Components to be computed
SS+IP+CMN	64.31	512
SS+IP64+CMN	62.28 (96.84%)	384 (75.0%)
SS+IP96+CMN	61.62 (95.82%)	320 (62.5%)

Table 6. Word accuracy for the baseline system to the real car-driving condition, Car01 testing. (%)

Clean (ch1)	Noisy (ch4)	SS (ch4)	SS-CMN (ch4)
94.16	58.76	82.94	88.96

Table 7. Word accuracy for the proposed schemes to the real car-driving condition, channel 4 of Car01. (%)

PCMM	IP	SS-IP-CMN	SS-IP64-CMN
88.96	88.96	91.33	91.24

Table 8. Relationship of performance and reduction in the Gaussian number on channel 4 of Car01 testing. (%)

	Relative improvement	Components to be computed
SS-IP-CMN	78.98	512
SS-IP64-CMN	78.76 (99.72%)	384 (75%)

8. REFERENCES

[1] X. Huang, A. Acero and H. Hon, *Spoken Language Processing*, Prentice Hall PTR, 2001.

[2] M. J. F Gales and S. J. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination," *IEEE Trans. on Speech and Audio Processing*, Vol.4, No.5, pp.352-359, Sep. 1996.

[3] P. J. Moreno, B. Raj and R. M. Stern, "Data-driven Environmental Compensation for Speech Recognition: A Unified Approach," *Speech Communication*, Vol.24. No.4, pp.267-285, July 1998.

[4] P. J. Moreno, *Speech Recognition in Noisy Environments*, PhD Thesis, Carnegie Mellon University, 1996.

[5] W. Kim, S. Ahn and H. Ko, "Feature Compensation Scheme Based on Parallel Combined Mixture Model," *Proc. Eurospeech2003*, pp.677-680, Sep. 2003.

[6] http://www.sitec.or.kr