Improved Acoustic Modeling Based on Selective Data-driven PMC*

Wooil Kim · Sunmee Kang*** · Hanseok Ko*****

ABSTRACT

This paper proposes an effective method to remedy the acoustic modeling problem inherent in the usual log-normal Parallel Model Composition intended for achieving robust speech recognition. In particular, the Gaussian kernels under the prescribed log-normal PMC cannot sufficiently express the corrupted speech distributions. The proposed scheme corrects this deficiency by judiciously selecting the "fairly" corrupted component and by re-estimating it as a mixture of two distributions using data-driven PMC. As a result, some components become merged while equal number of components split. The determination for splitting or merging is achieved by means of measuring the similarity of the corrupted speech model to those of the clean model and the noise model. The experimental results indicate that the suggested algorithm is effective in representing the corrupted speech distributions and attains consistent improvement over various SNR and noise cases.

Keywords: Additive Noise, Robust Speech Recognition, PMC, Log-normal, Data-driven

1. Introduction

The difference between training and operating environments is a significant factor affecting, in fact usually degrading the performance of speech recognition system. How to make both conditions equal is one of the most essential issues in the development of the actual applications of speech recognition technology and vigorous research has been pursued to realize this goal. For example, classical noise removal or speech enhancement methods, as an effort to bring the operating environment closer to the training environment, have been used at the pre-processing level of speech recognition system. Spectral subtraction, Wiener filter, AEC (Adaptive Echo Cancellation) and HMM (Hidden

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

** Dept. of Electronics and Computer Engineering, Korea University

*** Dept. of Computer Science, Seokyeong University

**** Dept. of Electronics and Computer Engineering, Korea University

Dept. of Electrical and Computer Engineering, Johns Hopkins University

Markov Model) based noise suppression are some of the prominent examples [1]. While these methods can be implemented independently from the recognition system and have shown considerable effects in noise canceling, they cannot guarantee attaining the intelligibility essential to recognition, and in some cases they actually produce spectral distortions in the restored speech signal. Another approach is to compensate for the environmental influence at the feature extraction step. CMS (Cepstral Mean Subtraction) is a representative technique belonging to this category. The third approach, which is essentially the focus of this paper, is the compensation method based on acoustic speech model. It aims at not removing the noise components but generating the speech model matched to noisy environment. MAP (Maximum A Posteriori) and MLLR (Maximum Likelihood Linear Regression) adaptation techniques and PMC (Parallel Model Combination) method are included in this category [1][2].

In this paper, we focus on the PMC method in an effort to improve recognition performance under additive noisy conditions. In PMC, the goal is to estimate a new speech model compensated for by noise components identical to the ensuing noisy conditions by using a clean speech model and a noise model independently. It exhibits an outstanding advantage in that it does not require any noise-corrupted speech samples for training and shows reasonably improved performance. In this paper, we identify the inherent problem found in the log-normal approximation PMC and propose a remedial algorithm to cope with that. Also, to overcome the limitation of spectral modeling in log-normal technique, we selectively employ the data-driven PMC method.

The paper is organized as follows. We first review the basic concept of PMC method in Section 2 and identify the acoustic modeling problem existing in PMC and then describe the proposed algorithm in Section 3. The experimental procedures and results are presented and discussed in Section 4. Finally, in Section 5, we make concluding remarks and discuss future works.

2. Parallel Model Composition

In PMC, assuming that a recognition system has achieved optimal performance when the training and testing conditions are identical, the clean speech model is transformed to the corrupted speech model matched to the actual noisy environment. To generate the noise-corrupted speech model, the clean speech model and noise model are used independently. PMC is known to exhibit an outstanding advantage in that it does not require additional training procedures with noisy speech database [2].

2.1. Log-normal approximation

The composition of the speech model and noise model is accomplished along a mismatch function. The mismatch function of static speech feature is as follows.

$$O_i^l(\tau) = F(S_i^l(\tau), N_i^l(\tau)) = \log(g\exp(S_i^l(\tau)) + \exp(N_i^l(\tau))) \tag{2.1}$$

In (2.1), $O_i^l(\tau), S_i^l(\tau), N_i^l(\tau)$ denote log spectrum's ith component of corrupted speech, clean speech and noise signals respectively. The additive relation of clean speech and noise in time is available in the linear spectral domain. Equation (2.1) shows such relation using log spectrum of clean speech and noise. Gaussian form in log spectrum is converted to log-normal distribution in the linear spectrum by transform. In log-normal approximation, it is assumed that the addition of two log-normal distributions comes to have a log-normal form also. The mean and covariance of corrupted speech are computed by equation (2.2) under such assumptions. In (2.2), $\hat{\mu}, \mu, \tilde{\mu}$ refer to mean vectors of corrupted speech, clean speech and noise and $\hat{\Sigma}, \Sigma, \tilde{\Sigma}$ denote covariance matrices of them respectively in log-normal distributions.

$$\begin{aligned} \hat{\mu} &= g\mu + \tilde{\mu} \\ \hat{\Sigma} &= g^2\Sigma + \tilde{\Sigma} \end{aligned} \tag{2.2}$$

The mean and covariance of linear spectrum that has log-normal distribution are obtained from the mean and covariance of log spectrum by (2.3).

$$\begin{aligned} \mu_i &= \exp(\mu_i^l + \Sigma_{ii}^l/2) \\ \Sigma_{ij} &= \mu_i\mu_j[\exp(\Sigma_{ij}^l) - 1] \end{aligned} \tag{2.3}$$

Finally, we calculate the mean and covariance of corrupted speech's log spectrum by the following equations.

$$\begin{aligned} \hat{\mu}_i^l &\approx \log(\hat{\mu}_i) - \frac{1}{2}\log\left(\frac{\hat{\Sigma}_{ii}}{\hat{\mu}_i^2} + 1\right) \\ \hat{\Sigma}_{ij}^l &\approx \log\left(\frac{\hat{\Sigma}_{ii}}{\hat{\mu}_i\hat{\mu}_j} + 1\right) \end{aligned} \tag{2.4}$$

2.2. Data-driven PMC

In log-normal approximation, the processing is applied to one-to-one between Gaussian kernel of the clean speech model and one of the noise model. In order to

maintain the system's framework of the clean speech model, each kernel function of the corrupted speech must be estimated as a single Gaussian function. Generally speaking, when two signals have Gaussian distribution functions and additive relations in their exponential terms, the closer their means are, the more different the distribution of their composition is from single Gaussian form [3]. Figure 1 shows the trend in the distributions of corrupted speech in log spectral domain. The solid line is distribution plot of actual signal and dotted line is estimate by single Gaussian function. When the mean of noise is 0, the estimated Gaussian function of noisy speech is similar with the actual signal's distribution. However, the estimate digresses significantly away from the actual plotting as the mean of noise becomes closer to that of clean speech.

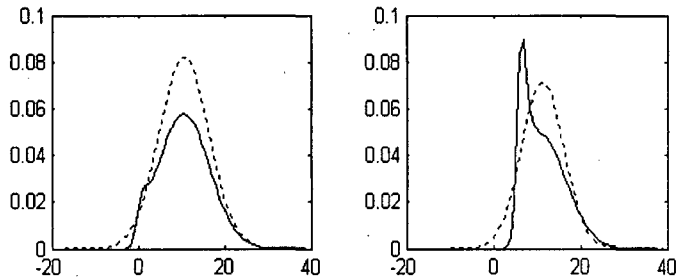

Figure 1. Log spectrum distributions (solid) and maximum likelihood Gaussian distributions (dotted) of corrupted speech when mean and variance of clean speech are 10.5 and 36. (a) noise mean is 0 with variance 0. (b) noise mean is 6 with variance 0.

The data-driven PMC utilizes artificially generated "observations" to cope with the problem appeared in the model combination addressed above. In particular, clean speech and noise "observations" generated from each acoustic model are added in the linear spectral domain and then, the noise-corrupted speech model is estimated with the synthesized observations [3]. Instead of one-to-one processing at each Gaussian kernel, the observations are generated from a mixture function of Gaussian kernels at each state and the model is re-estimated with those samples, so the distribution of corrupted speech can be modeled more adequately by multiple components.

3. Proposed Method

As shown in Figure 1, if two signals have an additive relationship in the linear-spectral domain and their means are similar to each other in the log-spectral domain where they have Gaussian distributions, the composition of the two signals has a distribution considerably different from the maximum likelihood Gaussian distribution

model. Data-driven PMC has been introduced to overcome the modeling limitation of log-normal method, but it has a huge computational load to generate observations and estimate parameters in every state of all acoustic models. In this paper, in an effort to solve the problem in composing the distribution of log-normal PMC and to reduce the computational load, we propose a modified PMC method in which the data-driven technique is selectively employed.

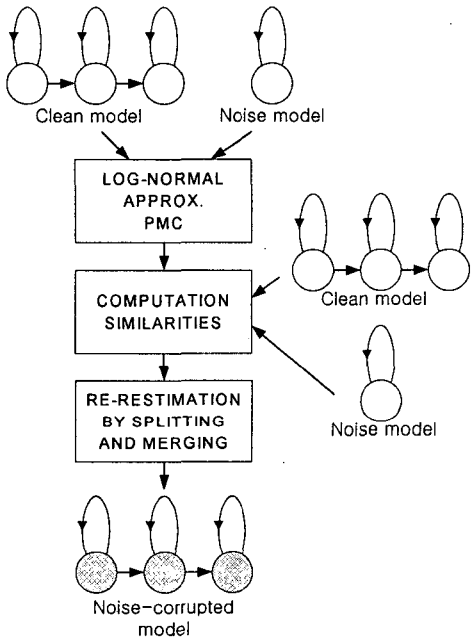

Figure 2. Block diagram of the proposed algorithm

First, the log-normal approximation PMC is applied to every Gaussian kernel of clean speech model using noise model. Then, among the Gaussian kernels in estimate of noise-corrupted model, we choose the components that fail to model the corrupted speech adequately and re-estimate them into two split Gaussian distributions by employing data-driven PMC. In the PMC scheme, the original system's framework has to be maintained, so it is impossible to change the number of kernels in each state. Therefore, it needs to merge the same number of components as the split ones. The components for splitting and merging are selected by means of similarity measure with clean speech and noise distributions. If a distribution kernel of corrupted speech is *significantly similar to that of noise speech*, it is reasonable to presume that a corresponding clean component is vulnerable to noise corruption. In this case, merging the noise-like components into one kernel at the same state is expected to affect little in the system's performance. Since the components that appear *similar in both the clean model and the corrupted model* are

anticipated to be robust to noisy, they have to be maintained to guarantee a well-behaved performance. Among the kernels of the corrupted model, the interest is in the components whose *distributions' difference is prominent between the clean model and the noise model.* They represent the components *corrupted "fairly" by the background noise* and may fail to reflect the corrupted speech adequately as addressed in Section II. In the proposed method, we selectively find the components for splitting or merging by using a similarity measure. In the case of splitting, the components are re-estimated into two split Gaussian functions. In the case of merging, the components are merged by the same number as that of the split cases.

To measure the similarity between distribution functions, we employ the Kullback-Leibler divergence as follows [4]:

$$I(p;q)=\frac{1}{2}\left\{\log\left(\frac{\Sigma_q}{\Sigma_p}\right)+\sum_{d=1}^{D}\left(\frac{(\mu_{pd}-\mu_{pd})^2}{\sigma_{qd}^2}+\frac{\sigma_{qd}^2}{\sigma_{pd}^2}\right)-2\right\} \tag{3.1}$$

Figure 2. shows a block diagram of the proposed algorithm.

4. Experimental Results

As a baseline, we constructed an isolated word recognition system using HTK [5]. Speech signals are analyzed within a 25 ms frame with 10 ms lapped into 39 th order feature vector that has 13 th order MFCCs including log energy and their 1 st and 2 nd derivatives. The number of Mel filter banks is 24. Context-independent 44 PLU (Phone-like unit) models are used and each HMM model is of 3-state left-to-right structure without any skip path. Every state has a continuous output probability function, a mixture consisting of 8 Gaussian components with diagonal elements only in the covariance matrix. The vocabulary to be recognized consists of 452 Korean words.

In the experiment, we used speech samples contained in Korean Speech DB PBW452 constructed for common use [6]. A total of 7,232 samples generated by 8 adult males recording two times are used for training the acoustic models using Baum-Welch algorithm. Data used for recognition testing are the 904 utterances of two males who were not part of the training set.

As shown in Table 1, we examined the baseline system's performance using clean speech samples and obtained 94.47% as recognition rate. In the experiments, the corrupted speech samples are generated by adding artificial white Gaussian noise to clean speech samples along given SNR. The 10 dB corrupted speech samples showed 16.81% in recognition using the clean speech model. It indicates that the difference between training and testing conditions brings drastic degradation in performance of the recognition

system. For recognition with the training model using 10 dB corrupted speech data, we obtained 90.27% using 10 dB noisy samples. Although the performance falls due to some units that lose discriminative properties under noisy conditions, the identical conditions of training and testing should guarantee the optimal results.

Table 1. Recognition test results - white noise case. (%)

	Clean	10 dB	5 dB	0 dB
Clean(no processing)	94.47	16.81	2.11	0.44
Matched		90.27	85.07	74.34
Matched(static)		81.75	65.71	39.71
Data-driven PMC		78.98	62.28	34.40
Log-normal PMC		76.88	60.51	30.09
Proposed PMC		**77.88**	**61.50**	**31.08**

Table 2. Number of Gaussian components undertaken for data-driven PMC.

Fully data-driven	Proposed PMC		
	10 dB	5 dB	0 dB
1,056(100%)	79(7.48%)	135(12.78%)	158(14.96%)

Table 3. Recognition test results - other noise cases. (%)

	Pink noise		Speech babble	
	10 dB	0 dB	10 dB	0 dB
Clean(no processing)	35.18	0.88	45.46	3.32
Log-normal PMC	77.32	22.90	84.18	36.50
Proposed PMC	**77.99**	**23.56**	**84.18**	**36.84**

Since the compensation is applied to only static parameters in the experiments of PMC, we trained the model with only static parameters matched to the noisy condition and attained 81.75% recognition rate. This result is to be used as the benchmark for comparison on the performance of the compensation methods.

We compared the proposed method to the log-normal approximation PMC in terms of recognition rate. The acoustic model for noise is estimated as one state with single Gaussian distribution function. The model compensation was performed over the mean and variance of a static parameter. We obtained 76.88% and 30.09% for 10 dB and 0 dB respectively from the model compensated for by the log-normal PMC. The increasing rates reflect the fact that the log-normal PMC is effective in compensating and generating the condition-matched model. The proposed PMC method brought 77.88% at 10 dB and 31.08% at 0 dB, which show an increment of about 1.0%. The improved

results indicate that the splitting and re-estimation process of the proposed scheme indeed contributes to the appropriate modeling of the corrupted speech.

Table 2 shows a comparison of the number of Gaussian components undertaken for "observations" generated and the parameters estimated by the data-driven method. Since the data-driven processing requires a huge computational load, the results in Table 2 confirms the system's efficiency. Only 7.48% components are used compared to the fully data-driven PMC under 10dB SNR. The computation complexity of Gaussian comparison processing is so trivial, therefore an additional 7.48% computational load of the fully data-driven led to the improved performance (77.88%), which is a middle figure between log-normal and data-driven method.

For other kinds of noise, we investigated the consistency in effectiveness of the proposed algorithm and presented the results in Table 3. Pink noise and babble noise contained in NOISEX-92 are used for the experiments. When using pink noise, the experiments attained improved results at 0.67% and 0.66% under 10 dB and 0 dB respectively. For speech babble noise, we obtained 0.34% increase in recognition rate at 0 dB SNR. From the experiments, we validated the improvement consistency in recognition performance for pink noise and babble noise cases. In the case of babble noise, from the fact that the recognition rate is comparatively high with the clean model, it is believed that the substantially small improvement is due to the rare case of less influence of noise to speech.

5. Conclusions

In this paper, we have proposed an algorithm that remedies the acoustic modeling problem inherent in the log-normal PMC aimed at robust speech recognition. By employing a selective data-driven PMC, the proposed scheme re-estimates the corrupted speech model by splitting and merging the results of the log-normal procedure. The results show that the suggested algorithm is effective in representing the corrupted speech distributions and achieves consistent improvement over various SNR and nose cases.

References

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

[2] Gales, M. J. F. & S. J. Young. 1996. "Robust Continuous Speech Recognition Using Parallel Model Combination." *IEEE Trans. on Speech and Audio Processing*,

Vol. 4, No. 5, 352-359.
[3] Gales, M. J. F. 1995. *Model-based Techniques for Noise Robust Speech Recognition.* Ph. D Thesis. Cambridge University.
[4] Kakizawa, Y., R. H. Shumway & M. Taniguchi. 1998. "Discrimination and Clustering for Multivariate Time Series." *Journal of the American Statistical Association*, Vol. 93, No. 441, 328-340.
[5] Young, S. 2000. *The HTK book (for HTK version 3.0).* Microsoft Corp.
[6] Kim, B., J. Kim, S. Kim & Y. Lee. 1997. "A Study on the Design and the Construction of a Korean Speech DB for Common Use." *Journal of Acoust. Soc. of Korea*, Vol. 16, No. 4, 35-41.

Received: Jan. 29, 2002.
Accepted: Mar. 8, 2002.

▲ Wooil Kim
Dept. of Electronics and Computer Engineering, Korea University
5Ka-1, Anam-dong, Sungbuk-ku, Seoul, 136-701, Korea.
Tel: +82-2-927-4724 (O) Fax: +82-2-3291-2450
H/P: 016-208-8546
E-mail: wikim@ispl.korea.ac.kr

▲ Sunmee Kang
Dept. of Computer Science, Seokyeong University
16Ka-1, Chongnung-Dong, Sungbuk-ku, Seoul, 136-704, Korea
Tel: +82-2-940-7291 (O) Fax: +82-2-919-0345
H/P: 011-9760-7144
E-mail: smkang@skuniv.ac.kr

▲ Hanseok Ko
Dept. of Electronics and Computer Engineering, Korea University
5Ka-1, Anam-dong, Sungbuk-ku, Seoul, 136-701, Korea.
Tel: +82-2-3290-3239 (O) Fax: +82-2-3291-2450
H/P: 011-9001-3239
E-mail: hsko@korea.ac.kr

Currently on Sabbatical Leave at:
Dept. of Electrical and Computer Engineering, Johns Hopkins University
Baltimore, MD, USA.
Tel: 410-516-7199 (O) 410-922-7907 (H)
E-mail: hsko@clsp.jhu.edu