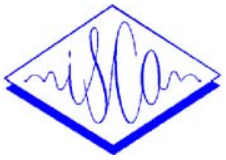

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

HMM COMPENSATION FOR NOISY SPEECH RECOGNITION BASED ON CEPSTRAL PARAMETER GENERATION

Takao Kobayashi†, *Takashi Masuko*†, *and Keiichi Tokuda*‡

†Precision and Intelligence Laboratory, Tokyo Institute of Technology, Yokohama, 226 Japan
‡Department of Computer Science, Nagoya Institute of Technology, Nagoya, 466 Japan
E-mail: tkobayas@pi.titech.ac.jp, masuko@pi.titech.ac.jp, tokuda@ics.nitech.ac.jp

ABSTRACT

This paper proposes a technique for compensating both static and dynamic parameters of continuous mixture density HMM to make it robust to noise. The technique is based on cepstral parameter generation from HMM using dynamic parameters. The generated cepstral vector sequences of speech and noise are combined to yield noisy speech cepstral vector sequence, and the dynamic parameters are calculated from the obtained cepstral vector sequence. Model parameters for noisy speech HMM are obtained using the statistics of the noisy speech parameter sequences. We use the mixture transition probability for estimating the parameters of the compensated model. Experimental results show the effectiveness of the proposed technique in the noisy speech recognition.

1. INTRODUCTION

The mismatch between training and testing conditions causes serious performance degradation in speech recognition systems. In general, retraining with the matched condition is the most effective way to improve the performance. However, it requires a high computational cost, and, moreover, it is not always possible due to lack of sufficient data for retraining.

Parallel model combination (PMC) [1] is one of the model compensation techniques that adapt HMM trained on clean speech data to make it robust to additive noise without requiring any data in the noisy environment. Although PMC with the Log-Normal approximation is computationally efficient and effective, the Log-Normal approximation is not always an appropriate assumption particularly in low SNR conditions. Moreover, in the original PMC formulation, the dynamic parameters are restricted to those calculated using simple differences. There have been proposed several approaches to overcome limitations of PMC [2]-[4].

We have proposed an alternative approach to compensation of both static and dynamic parameters of single Gaussian HMMs [5]. The approach is based on cepstral parameter generation from HMM using dynamic parameters [6],[7]. The generated cepstral vector sequences of speech and noise are combined to yield a noisy speech cepstral vector sequence, and the dynamic parameters are directly calculated from the obtained cepstral vector sequence. Compensated mean for the noisy speech HMM are obtained using the statistics of the noisy speech parameter sequences. In this paper, we extend this framework to the continuous mixture HMM case. To estimate the parameters of the noisy speech model, we introduce the mixture transition probability.

The idea of using a series of speech and noise observations generated from the speech and noise models is similar to that of Data-driven PMC (DPMC) [2]. However, in our approach, the generated cepstral sequence is a realistic speech parameter sequence which enables us to synthesize a high quality speech [8]. As a result, it is expected that we can obtain good estimates of the parameters including dynamic features for the noisy speech model. In addition, the dynamic parameters are not restricted to those calculated using simple differences.

2. HMM-BASED PARAMETER GENERATION AND COMPENSATION

2.1. ML-based cepstral parameter generation from continuous HMM

Let $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$ be a speech parameter vector sequence. We assume that the speech parameter vector $\boldsymbol{o}_t$ at frame t consists of the static feature vector $\boldsymbol{c}_t$, e.g., cepstral or mel-cepstral coefficient vector, and its dynamic feature vectors $\Delta\boldsymbol{c}_t, \Delta^2\boldsymbol{c}_t$, i.e., delta and delta-delta cepstral coefficient vectors, respectively. That is, $\boldsymbol{o}_t = [\boldsymbol{c}_t', \Delta\boldsymbol{c}_t', \Delta^2\boldsymbol{c}_t']'$, where $\cdot'$ denotes matrix transpose and

$$\Delta\boldsymbol{c}_t = \sum_{\tau=-L_1}^{L_1} w_\tau \boldsymbol{c}_{t+\tau}, \quad \Delta^2\boldsymbol{c}_t = \sum_{\tau=-L_2}^{L_2} w_\tau^{(2)} \Delta\boldsymbol{c}_{t+\tau}. \quad (1)$$

For a given continuous HMM λ, we can obtain a vector sequence $\boldsymbol{O}$ that maximizes $P(\boldsymbol{Q}, \boldsymbol{O}|\lambda, T)$ with respect to the state sequence $\boldsymbol{Q} = \{q_1, q_2, \ldots, q_T\}$ and $\boldsymbol{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_T\}$ with the constraints of (1) [6],[7]. If the state sequence $\boldsymbol{Q}$ is explicitly known, the optimal cepstral vector sequence $\boldsymbol{C}$ is determined by solving a set of linear equations. Even though the

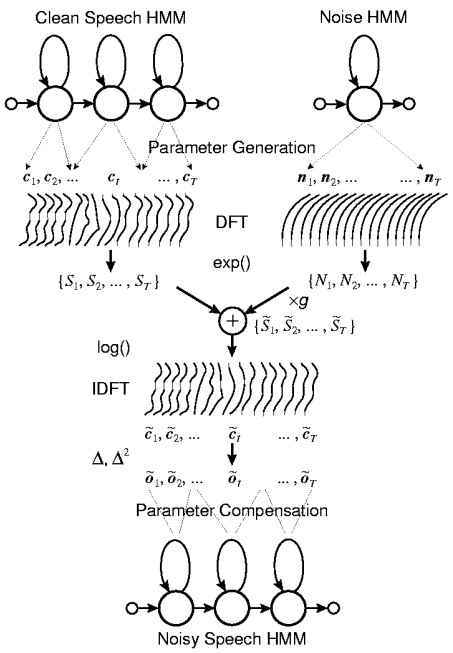

Figure 1: HMM-based parameter generation and compensation (PGC) [5].

state sequence is unknown, the optimal sequence is obtained using an efficient iterative algorithm.

2.2. Compensation for single Gaussian HMMs

An overview of the HMM-based parameter generation and compensation (PGC) technique [5] is presented in Fig. 1. It is assumed that speech and noise are independent and additive. It is also assumed that each state of HMMs has single Gaussian output distribution.

For a given speech HMM, we generate cepstral vector sequence $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_T\}$ using the ML-based parameter generation algorithm mentioned in **2.1**. From the generated cepstral vector sequence, we next obtain power spectrum sequence $\{S_1, \ldots, S_T\}$ by transforming the cepstrum vector $\boldsymbol{c}_t$ into linear power spectral domain at each frame t. Similarly, we also generate noise power spectrum sequence $\{N_1, \ldots, N_T\}$. Then we synthesize noisy speech spectrum sequence $\{\tilde{S}_1, \ldots, \tilde{S}_T\}$ by adding speech and noise spectra with a gain matching term g at each frame, and obtain noisy speech cepstral vector sequence $\{\tilde{\boldsymbol{c}}_1, \ldots, \tilde{\boldsymbol{c}}_T\}$ by transforming the noisy power spectrum into cepstral domain. From the obtained $\tilde{\boldsymbol{c}}_t$, dynamic parameters $\Delta\tilde{\boldsymbol{c}}_t$ and $\Delta^2\tilde{\boldsymbol{c}}_t$ are calculated using (1) and then noisy speech observation vector sequence $\{\tilde{\boldsymbol{o}}_1, \ldots, \tilde{\boldsymbol{o}}_T\}$ are composed.

Compensated mean vector $\tilde{\boldsymbol{\mu}}_k$ of the noisy speech HMM output distribution in state k is given by

$$\tilde{\boldsymbol{\mu}}_k = \boldsymbol{\mu}_k + (\tilde{\boldsymbol{m}}_k - \boldsymbol{m}_k) \tag{2}$$

where $\boldsymbol{\mu}_k$ is the mean vector of the clean speech HMM output distribution in the state k, $\boldsymbol{m}_k$ and $\tilde{\boldsymbol{m}}_k$ are the sample means of the generated clean and noisy speech parameter vectors within the state k, respectively.

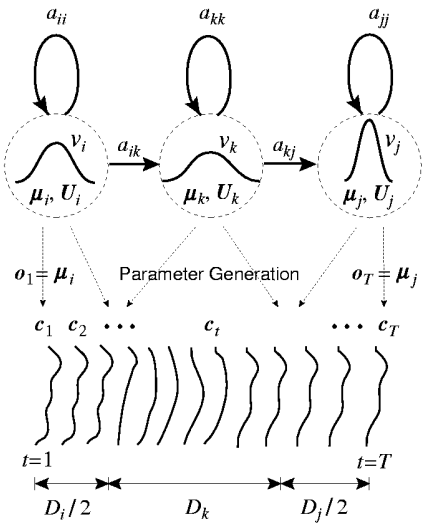

Figure 2: Parameter generation from mixture components of HMM.

3. PGC FOR CONTINUOUS MIXTURE DENSITY HMMS

In this section, we extend the PGC framework to the continuous mixture density HMM case. Since the sample mean of generated observation vectors are used for parameter compensation, it is desirable to obtain sufficient samples to get reliable estimation. For this purpose, we introduce transition probability between mixture components by considering the mixture components to be VQ codewords. The VQ-code transition probability was used for incorporating constraints into speaker independent HMM to adapt it to an input speaker [9]. Here we use this for estimating the mean vectors of the compensated model from the statistics of generated parameter sequences.

We consider the Gaussian mixture distribution to be a VQ codebook with mixture densities. Let v_k be a mixture component of any state output distribution having mean $\boldsymbol{\mu}_k$, and $d(\boldsymbol{o}_t, \boldsymbol{\mu}_k)$ be a distortion measure. We encode training speech parameter vector $\boldsymbol{o}_t$ by the index I_t of the codeword in such a way that

$$I_t = \arg\min_i d(\boldsymbol{o}_t, \boldsymbol{\mu}_i). \tag{3}$$

Then the mixture transition probability a_{kj}, from v_k to v_j is defined by

$$a_{kj} = P(I_{t+1} = j | I_t = k). \tag{4}$$

For the case of tied-mixture HMM or equivalently semi-continuous HMM, the transition probability is easily obtained using a set of shared mixture distributions as a VQ codebook. For the general continuous HMM case, VQ classification is done based on Viterbi

segmentation. In this case, the mixture components v_k and v_j can belong to the different states.

Using the transition probability of the mixture components, the compensation procedure is stated as follows:

1) At the end of the training stage of clean speech HMMs, obtain codeword index sequences for training data. From the obtained index sequences, calculate the transition probability between any possible pair of mixture components by

$$a_{kj} = N_{kj} \Big/ \sum_{i=1}^{K} N_{ki} \tag{5}$$

where N_{ki} is the number of transitions from the codeword index I_k to I_i and K is the VQ codebook size, i.e., the number of the mixture components.

2) For a set of mixture components $\{v_i, v_k, v_j\}$, generate cepstrum vector sequence $\{\boldsymbol{c}_t\}$ with the constraints that $\boldsymbol{o}_1 = \boldsymbol{\mu}_i$, $\boldsymbol{o}_T = \boldsymbol{\mu}_j$, and $T = D_i/2 + D_k + D_j/2$, where D_i, D_k, and D_j are the mean durations staying at the mixture components v_i, v_k, and v_j, respectively (Fig. 2). The mean duration D_k can be approximated by $D_k = 1/(1 - a_{kk})$.

3) Synthesize noisy speech parameter vector sequence $\{\tilde{\boldsymbol{o}}_t\}$ by combining generated speech and noise vector sequences in the same manner of the previous section.

4) Calculate the sample mean of the noisy speech parameter vector for the mixture component v_k by

$$\tilde{\boldsymbol{m}}_k = \sum_i \sum_j a_{ik} a_{kj} \tilde{\boldsymbol{m}}_{ikj} \Big/ \sum_i \sum_j a_{ik} a_{kj} \tag{6}$$

where

$$\tilde{\boldsymbol{m}}_{ikj} = \sum_{t=D_i/2+1}^{D_i/2+D_k} \tilde{\boldsymbol{o}}_t \Big/ D_k. \tag{7}$$

Similarly, calculate $\boldsymbol{m}_k$, i.e., the sample mean of the generated clean speech parameter sequences for the mixture component v_k. Finally obtain the compensated mean using (2).

4. EXPERIMENTAL RESULTS

To investigate effectiveness of the proposed technique, we performed experiments on a speaker-dependent 33-phoneme classification task using ATR Japanese Speech Database. The database consists of 5626 words uttered by a male speaker (MHT). We used even-numbered words in the database for training and odd-numbered words for testing. Each phoneme HMM is modeled by semi-continuous HMM (SCHMM) having 3 states with a VQ codebook size of 256. The topology for all models was left-right with no skips. Thirteen mel-cepstral coefficients including the 0-th coefficient were used as the static parameters. The mel-cepstral coefficients were obtained by a mel-cepstral analysis technique [10] on each 25.6ms frame of speech sampled at 10kHz with a Blackman window every 10ms. We used only delta parameters $\Delta\boldsymbol{c}$ as the dynamic features in the experiments. Delta mel-cepstral coefficients were calculated by (1) with $L_1 = 1$, $w_1 = -w_{-1} = 1/2$, and $w_0 = 0$. After the training of the HMMs, the codeword index sequences for training data were obtained using the VQ codebook of the SCHMMs. We used Mahalanobis distance as the distortion measure given by

$$d(\boldsymbol{o}_t, \boldsymbol{\mu}_k) = (\boldsymbol{o}_t - \boldsymbol{\mu}_k)' \boldsymbol{U}_k^{-1} (\boldsymbol{o}_t - \boldsymbol{\mu}_k) \tag{8}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{U}_k$ are the mean and covariance of the mixture component v_k in the VQ codebook, respectively. Then the mixture transition probabilities were calculated and stored for all possible codeword pairs.

Car noise or computer room noise from JEIDA Noise Database was added to clean speech to generate noisy speech data for testing. Noise was modeled by a single state HMM. Diagonal covariance matrices were used in both clean speech and noise models.

Table 1: Phoneme classification rates in car noise.

Model set	SNR (dB)					
	−6	0	6	12	18	24
Clean	40.5	56.0	68.4	76.0	80.0	86.4
Matched	70.4	82.2	86.9	88.7	90.9	91.7
PGC(Mean)	62.5	74.1	82.3	87.7	90.5	91.8
PGC(Fix)	66.5	76.6	83.8	88.7	91.0	92.6
PMC(Mean)	56.4	69.8	79.5	85.8	89.5	91.4
PMC(Fix)	61.6	71.6	80.4	87.1	90.5	92.3
PMC(Var)	58.8	72.2	81.3	86.7	89.8	91.5
DPMC(Mean)	60.1	72.4	81.6	86.7	89.4	91.1
DPMC(Var)	64.0	76.0	83.8	88.8	90.7	91.7

Table 2: Phoneme classification rates for DPMC with modified mismatch function in car noise.

Compensation	SNR (dB)					
	−6	0	6	12	18	24
Mean	60.9	73.2	82.0	86.9	89.6	91.3
Mean + Var	67.6	77.4	84.5	88.5	90.0	90.8

Table 1 shows the performance of the compensated models using the proposed technique (PGC), PMC, and DPMC for the phoneme classification task in car noise. In the table, the entries for "Matched" model are the results for the HMMs trained on noisy speech data with the same SNR as the input speech. The term "Mean" represents the mean compensation, and "Fix" means the use of a *fixed variance* [1] with the mean compensation. In addition, the term "Var" represents both mean and variance compensation. We set all off-diagonal terms of the covariance matrices of the noisy speech models to zero in all compensation techniques. Moreover, we used DPMC in a non-iterative fashion and with an assumption that the statistics of S_{t-1} and N_{t-1} are approximately the same as those of S_t and N_t, respectively.

It can be seen that similar classification perfor-

Table 3: Phoneme classification rates in computer room noise.

Model set	SNR (dB)					
	−6	0	6	12	18	24
Clean	16.2	26.5	40.2	54.4	66.0	74.9
Matched	48.9	64.1	77.3	85.5	89.8	91.5
PGC(Mean)	32.9	45.8	61.4	74.3	84.2	89.3
PGC(Fix)	40.8	55.2	68.0	78.7	85.7	89.1
PMC(Mean)	27.1	40.9	55.0	68.7	81.0	87.8
PMC(Fix)	33.0	47.6	62.0	74.0	83.6	88.4
PMC(Var)	32.1	44.9	58.7	72.2	83.2	88.7
DPMC(Mean)	28.5	42.9	57.8	72.2	82.2	87.7
DPMC(Var)	37.6	53.1	66.7	78.5	87.0	90.6

Table 4: Phoneme classification rates for DPMC with modified mismatch function in computer room noise.

Compensation	SNR (dB)					
	−6	0	6	12	18	24
Mean	28.7	43.2	58.8	73.1	82.6	87.4
Mean + Var	45.1	58.9	71.7	81.9	87.6	89.7

mance is achieved in high SNR conditions. However, in low SNR conditions, PGC provides higher performance than PMC or DPMC.

Since it seems that the mismatch function used in DPMC does not give good estimates of the delta parameters, we modified here the mismatch function for the delta parameters. We first generate a speech observation and further generate static parameters of the preceding and the following frames by

$$\boldsymbol{c}_{-1} = \boldsymbol{c} - \Delta\boldsymbol{c}, \quad \boldsymbol{c}_{+1} = \boldsymbol{c} + \Delta\boldsymbol{c} \tag{9}$$

where $\boldsymbol{c}$ and $\Delta\boldsymbol{c}$ are the generated static and delta parameter vectors, respectively. We next generate 3 noise observations independently and combine them with $\{\boldsymbol{c}_{-1}, \boldsymbol{c}, \boldsymbol{c}_{+1}\}$ to synthesize the noisy observations $\{\tilde{\boldsymbol{c}}_{-1}, \tilde{\boldsymbol{c}}, \tilde{\boldsymbol{c}}_{+1}\}$. Finally, we get the mismatch function for the delta parameters by $\Delta\tilde{\boldsymbol{c}} = (\tilde{\boldsymbol{c}}_{+1} - \tilde{\boldsymbol{c}}_{-1})/2$.

Table 2 shows the results using DPMC with the above mentioned mismatch function. It is shown that further improvement of the performance is achieved by the modification. It is also shown that the performance of PGC with the mean compensation and fixed variance is comparable with DPMC with both mean and variance compensation.

Table 3 and Table 4 show the results in computer room noise. It can be seen that similar results to those in car noise are obtained.

The computational cost of PGC depends on the VQ codebook size K. In SCHMM case, theoretically, we have to generate $K \times K$ parameter sequences per mixture component. However, the value of the transition probability $a_{ik}a_{kj}$ is equal to zero in most of the mixture component sets. In fact, the average number of generated sequences per mixture component was 856 in the experiment with the codebook size of $K = 256$. Furthermore, only slight performance degradation was observed when the 100 most significant transition probability mixture component sets per mixture component were used in the parameter generation. Even if the top 50 mixture component sets were used, the decrease in classification rate was about 1% and the computational load was significantly reduced.

5. CONCLUSION

We have proposed a technique for compensating both static and dynamic parameters of multiple mixture HMMs in noisy environments. The approach is based on cepstral parameter generation from HMM. The effectiveness of the technique has been investigated by phoneme classification experiments. Although the proposed technique requires the additional information of the mixture transition probability, it is a simple and easy task to obtain that information.

Acknowledgement The authors would like to thank Rui Yamada for helping with experiments. They also would like to thank Prof. Satoshi Imai for useful discussions.

REFERENCES

[1] M.J.F. Gales and S.J. Young, "Cepstral parameter compensation for HMM recognition in noise," *Speech Communication,* vol.12, pp.331–239, 1993.

[2] M.J.F. Gales and S.J. Young, "A fast and flexible implementation of parallel model combination," *Proc. ICASSP-95,* pp.133–136, 1995.

[3] R. Nagayama and H. Matsumoto, "HMM composition of noisy speech based on generation of spectrum sequence," *Proc. ASJ Spring Meeting* pp.7–8, 1994. (in Japanese)

[4] R. Yang, M. Majaniemi, and P. Haavisto, "Dynamic parameter compensation for speech recognition in noise," *Proc. EUROSPEECH-95,* pp.469–472, 1995.

[5] T. Kobayashi, T.Masuko, K. Tokuda and S. Imai, "Noisy speech recognition using HMM-based cepstral parameter generation and compensation," *Proc. ASA & ASJ 3rd Joint Meeting*, pp.1117–1122, Dec. 1996 / *J. ASA*, vol.100, no.4, pt.2, p.2790, Oct. 1996.

[6] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," *Proc. ICASSP-95,* pp.660–663, 1995.

[7] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," *Proc. EUROSPEECH-95,* pp.757–760, 1995.

[8] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," *Proc. ICASSP-96,* pp.389–392, 1996.

[9] S. Takahashi, T. Matsuoka and K. Shikano, "Phonemic HMM constrained by statistical VQ-code transition," *Proc. ICASSP-92,* pp.I-553–556, 1992.

[10] K. Tokuda, T. Kobayashi, T. Masuko and S. Imai, "Mel-generalized cepstral analysis — A unified approach to speech spectral estimation," *Proc. ICSLP-94,* pp.1043–1046, 1994.