

# Adaptation of Acoustic Model Using the Gain-Adapted HMM Decomposition Method

Akira Sasou<sup>1</sup>, Futoshi Asano<sup>1</sup>, Kazuyo Tanaka<sup>2,1</sup>, Satoshi Nakamura<sup>3</sup>

<sup>1</sup>National Institute of Advanced Industrial Science and Technology (AIST)

a-sasou@aist.go.jp, f.asano@aist.go.jp

<sup>2</sup>Institute of Library and Information Science, University of Tsukuba

ktanaka@ulिस.ac.jp

<sup>3</sup>ATR Spoken Language Translation Research Laboratories

satoshi.nakamura@atr.co.jp

## Abstract

In a real environment, it is essential to adapt acoustic models to variations in background noises in order to realize robust speech recognition. In this paper, we construct an extended acoustic model by combining a mismatch model with a clean acoustic model trained using only clean speech data. We assume the mismatch model conforms to a Gaussian distribution with time-varying population parameters. The proposed method adapts on-line the extended acoustic model to the unknown noises by estimating the time-varying population parameters using a Gaussian Mixture Model (GMM) and Gain-Adapted Hidden Markov Model (GA-HMM) decomposition method. We performed recognition experiments under noisy conditions using the AURORA2 database in order to confirm the effectiveness of the proposed method.

## 1. Introduction

One of approaches used to realize robust speech recognition is feature compensation that estimates clean speech features from noise-corrupted speech features[1, 2]. It is shown that a method based on a Gaussian Mixture Model (GMM) is effective for feature compensation[3]. This method estimates the expected value of mismatch feature and provides the clean speech features by subtracting the estimated mismatch feature from the noise-corrupted speech feature.

Another effective approach for robust speech recognition is model adaptation that adapts acoustic models to noisy conditions[4, 5, 6]. Model adaptation methods have an advantage in that these methods can adapt not only expected values but also distributions. When the noise environment can be modeled in advance, methods such as Parallel Model Combination are useful for adapting acoustic models. When the methods however are applied to an unknown noise environment, it is difficult to maintain sufficient recognition accuracy. In such a case, the acoustic models need to be adapted on-line to the unknown noise environment.

In this paper, we construct an extended acoustic model by combining a mismatch model with a clean acoustic model trained using only clean speech data. The proposed method adapts on-line the extended acoustic models to the unknown noise environment. The mismatch model is assumed to conform to a Gaussian distribution with time-varying population parameters. We adopt the feature compensation method based on GMM at the front end processing in order to eliminate the ex-

pected value of the mismatch feature. The features compensated by this method are distributed around true feature of the clean speech. As the background noise exhibits more non-stationary characteristics, the variance of the mismatch feature tends to become larger. As a result, the recognition accuracy is degraded. In order to eliminate the variance of the mismatch feature, we adopt the Gain-Adapted Hidden Markov Model (GA-HMM) decomposition method[7] that estimates on-line the variance of the mismatch feature from the feature compensated by the front end processing and then extracts the clean speech feature.

## 2. Modeling of the Mismatch Feature

Assuming that the speech and noise signals are uncorrelated, the Filter Bank Energy (FBE) of the noisy speech  $\mathbf{Y}_b(n)$  at frame  $n$  can be represented as a function of the clean speech  $\mathbf{X}_b(n)$  and the noise  $\mathbf{N}_b(n)$ ,

$$\mathbf{Y}_b(n) = \mathbf{X}_b(n) + \mathbf{N}_b(n). \quad (1)$$

This relation yields the expression of the noisy speech  $\mathbf{Y}_l(n)$  in the log-FBE domain:

$$\begin{aligned} \mathbf{Y}_l(n) &= \mathbf{X}_l(n) + \log [1 + \exp(\mathbf{N}_l(n) - \mathbf{X}_l(n))] \\ &= \mathbf{X}_l(n) + \mathbf{g}_l(\mathbf{N}_l(n), \mathbf{X}_l(n)) \\ &= \mathbf{X}_l(n) + \mathbf{G}_l(n) \end{aligned} \quad (2)$$

where  $\mathbf{G}_l(n)$  represents the mismatch feature and the subscript  $l$  denotes that the expression is in the log-FBE domain. Similarly, the noisy speech  $\mathbf{Y}_c(n)$  in the cepstral domain can be represented by

$$\begin{aligned} \mathbf{Y}_c(n) &= \mathbf{C}\mathbf{Y}_l(n) \\ &= \mathbf{X}_c(n) + \mathbf{C} \log [1 + \exp(\mathbf{N}_l(n) - \mathbf{X}_l(n))] \\ &= \mathbf{X}_c(n) + \mathbf{G}_c(n) \end{aligned} \quad (3)$$

where  $\mathbf{C}$  denotes Discrete Cosine Transform (DCT) matrix and the subscript  $c$  denotes the cepstral domain.

In this paper, we assume that the mismatch feature  $\mathbf{G}_c(n)$  is a stochastic variable that is independent of the clean speech feature  $\mathbf{X}_c(n)$  and conforms to a Gaussian distribution with time-varying population parameters. That is,

$$\mathbf{G}_c(n) \sim \mathcal{N}(\mu_{g_c}(n), \Sigma_{g_c}(n)) \quad (4)$$

where  $\Sigma_{g_c}(n) = \text{diag}(\sigma_{g_c,1}^2(n), \dots, \sigma_{g_c,D}^2(n))$  and  $D$  indicates the number of the cepstral coefficients.

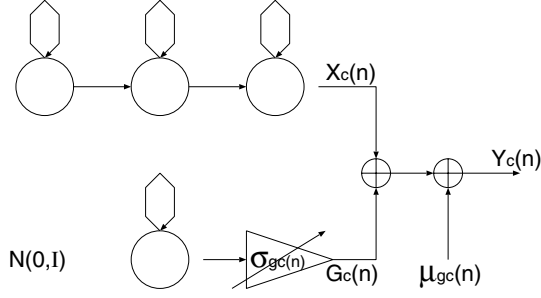


Figure 1: An extended acoustic model

By combining the clean acoustic model and the mismatch model according to equation (3), we obtain the extended acoustic model shown in Figure 1, in which we show as an example a clean acoustic model consisting of three states.

### 3. Expectation Estimation Based on GMM

In order to estimate an expectation of the mismatch feature  $\mu_{gc}(n)$  at the front end processing, we adopt a method based on a Gaussian Mixture Model (GMM). In the following, we describe the procedure of the compensation method.

The compensation method adopts a K-Gaussian mixture to describe the clean speech feature in the Log-FBE domain

$$p(\mathbf{X}_i) = \sum_{k=1}^K P(v_k) \mathcal{N}(\mathbf{X}_i; \mu_{x_l, k}, \Sigma_{x_l, k}) \quad (5)$$

where  $v_k$  is the  $k$ -th Gaussian pdf and the covariance matrix is assumed to be diagonal. The expectation of the mismatch feature  $\mu_{gl, k}$  for each pdf  $v_k$  is estimated using the noise-corrupted speech feature  $\mathbf{Y}_l(n)$  of the first 10 frames which are assumed to be silence

$$\mu_{gl, k} = \frac{1}{10} \sum_{n=1}^{10} \mathbf{g}_l(\mathbf{Y}_l(n), \mu_{x_l, k}). \quad (6)$$

Each Gaussian pdf  $v_k$  for  $\mathbf{Y}_l(n)$  is evaluated by

$$\begin{aligned} \mu_{y_l, k} &= \mu_{x_l, k} + \mu_{gl, k} \\ \Sigma_{y_l, k} &= \Sigma_{x_l, k} \end{aligned} \quad (7)$$

The expectation of the mismatch feature  $\mu_{gl}(n)$  is then evaluated according to

$$\mu_{gl}(n) = \sum_{k=1}^K P(v_k | \mathbf{Y}_l(n)) \mu_{gl, k} \quad (8)$$

where the post probability  $P(v_k | \mathbf{Y}_l(n))$  is given by

$$P(v_k | \mathbf{Y}_l(n)) = \frac{P(v_k) \mathcal{N}(\mathbf{Y}_l(n); \mu_{y_l, k}, \Sigma_{y_l, k})}{\sum_{k'=1}^K P(v_{k'}) \mathcal{N}(\mathbf{Y}_l(n); \mu_{y_l, k'}, \Sigma_{y_l, k'})}. \quad (9)$$

In order to reduce the variance caused by the estimation error of the compensation method, the estimated expectation of the mismatch feature  $\mu_{gl}(n)$  is smoothed by an LPF with the following transfer function:

$$H(z) = \frac{1 - z^{-8}}{8(1 - z^{-1})} \quad (10)$$

Finally, the expectation of the mismatch feature in the cepstral domain is obtained by

$$\mu_{gc}(n) = \mathbf{C} \mu_{gl}(n). \quad (11)$$

### 4. Variance Estimation Based on GA-HMM Decomposition Method

In the extended acoustic model in Figure.1, the two simultaneous HMMs are combined and one of the HMMs has time-varying output gain. In the case without the time-varying output gain, the conventional HMM decomposition method can be applied to this model. The extended acoustic model, however, needs to successively adapt the output gain to the observations. In this paper, we proposed the Gain-Adapted HMM decomposition method, which is obtained by extending the conventional HMM decomposition method.

In the case of the conventional HMM decomposition method applied to two simultaneous components, the recurrent relation for evaluating the most likely state sequence is given by

$$P_n(i, j) = \max_{u, v} P_{n-1}(u, v) \cdot a_{1u, i} \cdot a_{2v, j} \cdot b_{1, u} \otimes b_{2, v}(\tilde{\mathbf{Y}}(n)) \quad (12)$$

where  $P_n(i, j)$  is the probability, at time  $n$ , of the first component being in state  $i$  and the second in state  $j$ ,  $a_{1u, i}$  is the transition probability from state  $u$  to state  $i$  for the first component,  $a_{2v, j}$  is the transition probability from state  $v$  to state  $j$  for the second component, and  $b_{1, u} \otimes b_{2, j}(\tilde{\mathbf{Y}}(n))$  is the observation probability.

In the general case, in which each component has time-varying output gain, the GA-HMM decomposition method requires adaptation of each output gain in the calculation of the observation probability. In the following, we describe the method for evaluating the observation probability in the GA-HMM decomposition method.

The output pdf of the  $m$ -th component and the  $q$ -th state is assumed to be Gaussian Mixture:

$$\begin{aligned} b_{m, q}(\tilde{\mathbf{Y}}; \mathbf{G}_m) &= \sum_{k=1}^{K_{m, q}} P(v_{m, q, k}) \times \\ &\mathcal{N}(\tilde{\mathbf{Y}}; \mathbf{G}_m \mu_{m, q, k}, \mathbf{G}_m^T \Sigma_{m, q, k} \mathbf{G}_m) \end{aligned} \quad (13)$$

where  $\tilde{\mathbf{Y}} = [y_1, \dots, y_D]^T$ ,  $\mu_{m, q, k} = [\mu_{m, q, k, 1}, \dots, \mu_{m, q, k, D}]^T$ ,  $\Sigma_{m, q, k} = \text{diag}(\sigma_{m, q, k, 1}^2, \dots, \sigma_{m, q, k, D}^2)$ ,  $\mathbf{G}_m = \text{diag}(g_{m, 1}, \dots, g_{m, D})$  and  $v_{m, q, k}$  is a Gaussian pdf with the expectation  $\mathbf{G}_m \mu_{m, q, k}$  and the covariance matrix  $\mathbf{G}_m^T \Sigma_{m, q, k} \mathbf{G}_m$ . Assuming that the observation was emitted from the combined Gaussian pdfs of  $v_{1, u, k_1}$  and  $v_{2, v, k_2}$ , the gains  $\mathbf{G}_{1, (k_1, k_2)}$  and  $\mathbf{G}_{2, (k_1, k_2)}$  are evaluated as follows. One element  $y_d$  of the observation is thought of as conforming to the following Gaussian pdf:

$$\begin{aligned} P(y_d; g_{1, d}, g_{2, d}) &= v_{1, u, k_1, d} \otimes v_{2, v, k_2, d}(y_d) \\ &= \mathcal{N}\left(y_d; \sum_{m=1}^2 g_{m, d} M_m, \sum_{m=1}^2 g_{m, d}^2 V_m\right) \end{aligned} \quad (14)$$

where  $v_{m, \cdot, k, d}$  is the Gaussian pdf for  $d$ -th element of  $v_{m, \cdot, k}$ , and  $M_m$  and  $V_m$  are the expectation and variance of  $v_{m, \cdot, k, d}$ . We estimate the gains so that this post probability increases by using an adaptive estimation algorithm based on the gradient of equation (14). The gradient of the logarithmic probability of

equation (14) with respect to gain  $g_{l,d}$  is given by

$$\frac{\partial \ln P}{\partial g_{l,d}} = \frac{\{y_d - \sum_{m=1}^2 g_{m,d} M_m\} M_l - g_{l,d} V_l}{\sum_{m=1}^2 g_{m,d}^2 V_m} + \frac{\{y_d - \sum_{m=1}^2 g_{m,d} M_m\}^2 g_{l,d} V_l}{\{\sum_{m=1}^2 g_{m,d}^2 V_m\}}. \quad (15)$$

Each gain is then updated according to

$$g_{l,d} = \alpha \cdot \frac{\partial \ln P}{\partial g_{l,d}} + \beta \cdot \check{g}_{l,d} \quad (16)$$

where  $\alpha$  is the step size,  $\beta$  is the forgetting factor, and  $\check{g}_{l,d}$  is the previous estimate at the combined state  $(u, v)$ . This gain adaptation is achieved for all dimensions and the new estimates of the gains are stored in  $\mathbf{G}_{1,(k1,k2)}$  and  $\mathbf{G}_{2,(k1,k2)}$ , respectively. This process is iterated for all the combination of pdfs  $(k1, k2)$ . We then select the most probable combination of pdfs according to

$$\hat{k}1, \hat{k}2 = \arg \max_{k1,k2} b_{1,u} \otimes b_{2,v}(\tilde{\mathbf{Y}}(n); \mathbf{G}_{1,(k1,k2)}, \mathbf{G}_{2,(k1,k2)}). \quad (17)$$

If we use the observation probability of the selected combination  $(\hat{k}1, \hat{k}2)$  to evaluate equation (12),  $P_n(i, j)$  represents the likelihood of the extended acoustic model given the noise-corrupted speech feature. However, our speech recognition system adopts the likelihood of the clean acoustic model given the clean speech feature. In order to achieve this, we need to decompose the observation into the clean speech and mismatch features based on the gain-adapted components. In the following, we describe how to decompose the observation into the two components.

We first select the most probable combined state by

$$\hat{u}, \hat{v} = \arg \max_{u,v} P_{n-1}(u, v) \cdot a_{1u,i} \cdot a_{2v,j} \cdot b_{1,u} \otimes b_{2,v}(\tilde{\mathbf{Y}}(n)) \quad (18)$$

where  $b_{1,u} \otimes b_{2,v}(\tilde{\mathbf{Y}}(n))$  represents the selected observation probability for each  $(u, v)$  according to equation (17). The decomposition process is applied to the observation, assuming that the observation was emitted from the combined state  $(\hat{u}, \hat{v})$ . Let  $q_{m,d}$  denote the  $d$ -th element of the output from the  $m$ -th component. Then,  $q_{1,d}$  and  $q_{2,d}$  conform to  $v_{1,\hat{u},\hat{k}1,d}$  and  $v_{2,\hat{u},\hat{k}2,d}$ , respectively. One element,  $y_d$ , of the observation can be represented by

$$y_d = \sum_{m=1}^2 q_{m,d}. \quad (19)$$

Since the outputs of the components are mutually independent, the joint distribution of all the outputs is given by

$$Q(q_{1,d}, q_{2,d}) = \prod_{m=1}^2 \mathcal{N}(q_{m,d}; M_m, V_m) \quad (20)$$

where  $(M_1, V_1)$  and  $(M_2, V_2)$  are the population parameters of Gaussian pdfs  $v_{1,\hat{u},\hat{k}1,d}$  and  $v_{2,\hat{u},\hat{k}2,d}$ , respectively. We decompose the element  $y_d$  so that the joint occurrence probability is maximized for the condition of equation (19). The decomposed values are given by

$$\begin{aligned} \hat{q}_{1,d} &= \frac{V_2 M_1 + V_1 (y_d - M_2)}{V_1 + V_2} \\ \hat{q}_{2,d} &= y_d - \hat{q}_{1,d} \end{aligned} \quad (21)$$

This decomposition process is iterated for all dimensions.

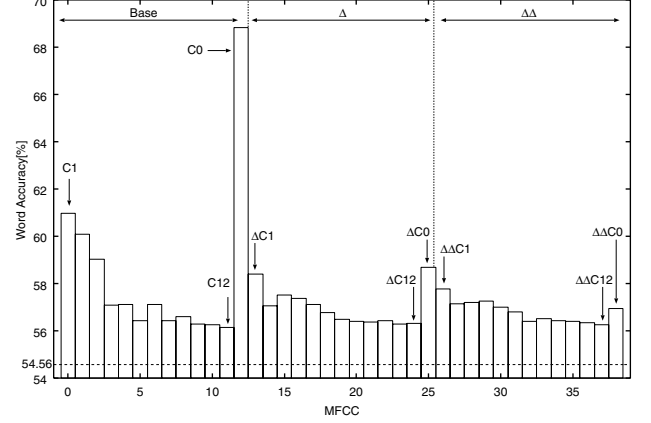


Figure 2: Recognition results based on 39-dimensional feature vectors, replacing each MFCC coefficient by that of the clean coefficient.

In the case of applying this GA-HMM decomposition method to the extended acoustic model in Figure 1, the observations are derived as  $\tilde{\mathbf{Y}}_c(n) = \mathbf{Y}_c(n) - \mu_{gc}(n)$ . The output gain of the clean acoustic model is fixed at 1. Let the first component ( $m = 1$ ) denote the clean acoustic model and let the second component ( $m = 2$ ) denote the mismatch model. The estimate of the clean speech feature is given by  $\tilde{\mathbf{X}}_c = [\hat{q}_{1,1}, \dots, \hat{q}_{1,D}]$ . The likelihood of the clean acoustic model given the clean speech feature can be evaluated using the following recurrent relation:

$$P_n(i, j) = \max_{u,v} P_{n-1}(u, v) \cdot a_{1u,i} \cdot b_{1,u}(\tilde{\mathbf{X}}_c) \quad (22)$$

## 5. Experiments

### 5.1. Experimental Setup

We used the AURORA2 database in order to evaluate the proposed method. The procedures for feature extraction are as follows. The frame length and period are set to 25 ms and 10 ms, respectively. After pre-emphasizing by  $1 - 0.97z^{-1}$ , the FFT is calculated. The inner products between the squared amplitudes of the FFT coefficients and the triangle windows of the Mel Filter Bank are calculated in order to generate the Mel-FBE feature. The natural logarithm of the Mel-FBE feature is then calculated. The expectation of the mismatch feature is evaluated by the method described in section 3. In this experiment, the number of Gaussian pdfs in equation (5) was set to 128. The GMM is trained using clean speech data for clean condition training. After subtracting the expectation of the mismatch feature, the MFCC is then obtained by applying DCT to the compensated feature. The MFCC is a 13-dimensional vector including the 0th coefficient. The delta and the delta-delta features are evaluated. By combining these features, the 39-dimensional feature vector is generated. The clean acoustic models for digits (1-9,zero,oh) were composed of 16 emitting states, with three mixtures per state. Those of sil and sp were composed of three emitting states and one emitting state, respectively, with six mixtures per state for both sil and sp. These clean acoustic models were trained using only clean speech data (clean training condition).

Generally, in comparison with the feature compensation

**Table 1.** Recognition results for Method-1

Aurora 2 Small Vocabulary	Clean training, multicondition testing													
	A						B				C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	98.96	99	99.02	99.2	99.05	98.96	99	99.02	99.2	99.05	99.05	99.18	99.12	99.06
20 dB	97.7	97.7	98.45	97.62	97.87	96.99	97.16	97.46	97.96	97.39	96.78	97.28	97.03	97.51
15 dB	96.16	94.98	97.55	95.5	96.05	92.82	95.04	96.09	96.79	95.19	94.63	94.71	94.67	95.43
10 dB	90.76	89.3	94.72	90.99	91.44	87.07	88.15	90.96	92.1	89.57	89.59	87.18	88.39	90.08
5 dB	77.99	72.73	84.91	78.99	78.66	69.97	73.88	78.88	80.19	75.73	77.43	71.55	74.49	76.65
0 dB	54.99	41.75	52.34	58.75	51.96	43.23	47.61	56.7	52.02	49.89	50.66	43.86	47.26	50.19
-5dB	27.14	15.24	18.25	32	23.16	15.04	22.55	23.86	21.17	20.66	24.72	20.56	22.64	22.05
Average	83.52	79.29	85.59	84.37	83.19	78.02	80.37	84.02	83.81	81.55	81.82	78.92	80.37	81.97

**Table 2.** Recognition results for Method-2

Aurora 2 Small Vocabulary	Clean training, multicondition testing													
	A						B				C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.02	99.09	98.99	99.23	99.08	98.89	99.06	98.99	99.23	99.04	99.05	99.15	99.10	99.07
20 dB	97.88	97.7	98.51	97.69	97.95	97.11	97.34	97.7	97.99	97.54	97.18	97.31	97.25	97.64
15 dB	96.38	95.59	97.52	95.62	96.28	94.04	95.68	96.54	96.95	95.80	95	94.92	94.96	95.82
10 dB	90.7	91.44	94.42	90.99	91.89	89.07	89.3	91.8	92.35	90.63	89.9	87.61	88.76	90.76
5 dB	78.81	76.75	84.31	78.9	79.69	73.84	74.4	79.93	80.62	77.20	77.8	72.13	74.97	77.75
0 dB	56.49	47.76	55.06	58.93	54.56	49.74	49.03	58.4	53.93	52.78	52.04	45.16	48.60	52.85
-5dB	27.45	18.41	14.35	31.26	22.87	19.65	22.04	23.86	20.06	21.40	23.7	18.14	20.92	21.89
Average	84.05	81.85	85.96	84.43	84.07	80.76	81.15	84.87	84.37	82.79	82.38	79.43	80.91	82.93

method, the model adaptation method requires numerous computations, especially when the number of models, mixture distributions or the dimension of the feature vector increases. The same problem arises in the GA-HMM decomposition method. We first conducted a preliminary experiment in order to investigate which coefficient in the 39-dimensional feature vector affects improvement of the recognition accuracy. Recognition was performed based on feature vectors generated by replacing a time series of one coefficient in the original 39-dimensional feature vectors by that of the clean coefficient. In this experiment, we used the speech data in test-set A and generated the original feature vectors without the compensation method based on the GMM. Figure 2 shows the recognition results. The average word accuracy was 54.56% when the original feature vectors were used. Based on these results, when  $c_0$  coincides with the clean coefficient, the recognition accuracy is drastically improved compared to the case for other coefficients. In the following experiments, we thus apply the GA-HMM decomposition method to  $c_0$  exclusively.

## 5.2. Experimental Results

We evaluated the following two methods:

- Method-1 : GMM
- Method-2 : GMM + GA-HMM decomposition method

Tables 1 and 2 show the results generated by Method-1 and Method-2, respectively. From these results, expectation compensation of mismatch effectively improved the recognition accuracy. By combining the expectation compensation and the variance compensation of mismatch, we can see that the average recognition accuracy tends to be improved.

## 6. Conclusion

In this paper, we constructed an extended acoustic model by combining the clean acoustic model and the mismatch model. The mismatch model is assumed to conform to a Gaussian distribution with time-varying population parameters. The proposed method can adapt on-line the extended acoustic model to

an unknown noise environment by estimating the time-varying population parameters of the mismatch model via GMM and GA-HMM decomposition methods and then estimate the clean speech feature.

In future studies, we plan to investigate the effectiveness of the proposed method when applied to speech recognition under unknown impulsive noise environments.

## 7. References

- [1] S.F.Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. on Acoust., Speech, Signal Process, ASSP-33, vol.27, pp.113-120, 1979.
- [2] B.S.Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J.Acoust. Soc. Am. Vol.55, pp.1304-1312, 1974.
- [3] J.C.Segura, A.de la Torre, M.C.Benitez, A.M.Peinado, "Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AU-RORA II database and tasks," Proc. of Eurospeech2001, Vol.I, pp.221-224, 2001.
- [4] M.Gales, P.C.Woodland, "Mean and variance adaptation within the MLLR framework," Computer Speech and Language, Vol.10, pp.249-264, 1996.
- [5] A.P.Varga, R.K.Moore, "Hidden Markov model decomposition of speech and noise," Proc. of ICASSP-90, pp.845-848, 1990.
- [6] M.J.F.Gales, S.J.Young, "Robust continuous speech recognition using parallel model combination," IEEE Trans. Speech and Audio Process.4, pp.352-359, 1996.
- [7] A.Sasou, K.Tanaka, "A waveform generation model based approach for segregation of monaural mixed sound," EURASIP Journal on Signal Processing, Vol.83/3, pp.561-574, Mar, 2003.