Feature Compensation based on Switching Linear Dynamic Model and Soft Decision

Woohyung Lim, Bong Kyoung Kim, Nam Soo Kim

School of Electrical Engineering and INMC
Seoul National University
whlim@hi.snu.ac.kr, bkkim@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract

In this paper, we present a new approach to feature compensation for robust speech recognition in noisy environments. We employ the switching linear dynamic model (SLDM) as a parametric model for the clean speech distribution, which enables us to utilize temporal correlations in speech signals. Both the background noise and clean speech components are simultaneously estimated by means of the interacting multiple model (IMM) algorithm. Moreover, we combine the SLDM algorithm with the spectral subtraction (SS) approach based on a soft decision. Performance of the presented compensation technique is evaluated through the experiments on AURORA 2 database.

1. Introduction

The performance of speech recognition systems deteriorates in the presence of background noise. One of the successful approaches to alleviate this type of performance degradation is the feature compensation technique in which noisy input features are compensated before being decoded using acoustic recognition models that were trained on clean speech. Even though a feature compensation algorithm can be designed without any prior knowledge of clean speech, it has been proven beneficial to employ a specific distribution model trained on an amount of clean speech data.

A prevailing way to characterize the clean speech distribution is to apply the Gaussian mixture model (GMM) which enables an exact approximation to the real distribution at a fixed time and can be easily trained using the *expectation-maximization* (EM) algorithm [1]-[4]. In [1], the clean speech feature vector is assumed to be distributed according to a GMM, and the Taylor series expansion technique is applied to linearize the observation function. Kim [2] further improved this GMM-based technique by incorporating a time-varying noise model for which the evolving noise can be tracked with the help of the interacting multiple model (IMM) algorithm.

A major drawback of the GMM for clean speech distribution modelling lies in its basic assumption that the feature vectors extracted at different times are statistically independent with each other. Since adjacent frames of speech usually possess a high correlation, a lot of valuable information is likely to be missed if we apply a GMM for feature compensation. Recently, a number of techniques have been developed to take advantage of the temporal correlations among adjacent frames of speech signal. In [5], an approach to capture the smooth time evolution of speech is proposed based on the switching linear dynamic model (SLDM).

As stated in [6], IMM-based approach causes undesirable insertion errors in recognition results in spite of quite a lot reduction of the recognition errors in the active speech regions. The SLDM-based IMM (SLDM-IMM) approach gives better estimation results for active speech regions than the conventional GMM-based IMM (GMM-IMM) approach [7], but in the non-speech regions, it causes the same type of errors as the GMM-IMM. One of the techniques to cope with this problem is the combination with the SS approach [6]. Because the SS approach more suppressed background noise during non-speech regions than the IMM -based approach, the combination of the SLDM-IMM algorithm and the SS improves the recognition performance effectively.

In this paper, we propose a new feature compensation technique in which the SS is combined with the SLDM-IMM results depending on the speech absence probability (SAP). The SAP is a byproduct of a usual speech enhancement technique. We construct a more rigorous dynamic model structure for the distributions of clean speech and combine it with a time-varying noise model. Both the noise and clean speech components are simultaneously estimated during feature compensation using an online implementation of the IMM algorithm. The performance of the proposed method is evaluated on AURORA2 database.

2. SLDM for Clean Speech Feature Distributions

Let $\mathbf{y}(t) = [y_1(t), y_2(t), \cdots, y_D(t)]'$ be a log spectral vector of clean speech at time t with the prime denoting the matrix or vector transpose. In SLDM, the whole space of the log spectra of clean speech is divided into K disjoint clusters, and for each cluster k, $\mathbf{y}(t)$ is described in terms of a linear state space model given as follows:

$$\begin{aligned} \mathbf{x}(t+1) &= F_k\mathbf{x}(t) + \mathbf{w}_k(t) \\ \mathbf{y}(t) &= \mathbf{x}(t) + \mu_k + \mathbf{v}_k(t) \end{aligned} \quad (1)$$

where $\mathbf{x}(t)$ represents a state variable which has the same dimensionality as that of $\mathbf{y}(t)$. In (1), μ_k is a constant vector accounting for the steady state feature vector and $\mathbf{w}_k(t)$ and $\mathbf{v}_k(t)$ are respectively the process and measurement noises, which are assumed to be independent with each other and distributed according to

$$\begin{aligned} \mathbf{w}_k(t) &\sim \mathcal{N}(\mathbf{0}, Q_k) \\ \mathbf{v}_k(t) &\sim \mathcal{N}(\mathbf{0}, R_k) \end{aligned} \quad (2)$$

in which $\mathcal{N}(a, b)$ denotes a Gaussian probability density function (pdf) with mean vector a and covariance matrix b.

The parameters $\{F_k, \mu_k, Q_k, R_k\}$ associated with the specified SLDM can be estimated from a set of clean speech training data by means of the EM algorithm [9]. Given a sequence of

log spectra from clean speech $\mathbf{y}_1^T = \{\mathbf{y}(1), \mathbf{y}(2), \cdots, \mathbf{y}(T)\}$, the first step is to apply the conventional Kalman filtering operation. For simplicity, let us assume that $k(t)$ referring to the cluster index of $\mathbf{y}(t)$ is known for $t = 1, 2, \cdots, T$. Let us also define

$$\begin{aligned} \mathbf{x}_{t|\tau} &= E[\mathbf{x}(t)|\mathbf{y}_1^\tau] \\ \Sigma_{t|\tau} &= E[\mathbf{x}(t)\mathbf{x}'(t)|\mathbf{y}_1^\tau] - \mathbf{x}_{t|\tau}\mathbf{x}'_{t|\tau} \\ \Sigma_{t_1,t_2|\tau} &= E[\mathbf{x}(t_1)\mathbf{x}'(t_2)|\mathbf{y}_1^\tau] - \mathbf{x}_{t_1|\tau}\mathbf{x}'_{t_2|\tau} \end{aligned} \quad (3)$$

where $E[\cdot]$ indicates the expectation and

$$\mathbf{y}_{t_1}^{t_2} = \{\mathbf{y}(t_1), \mathbf{y}(t_1+1), \cdots, \mathbf{y}(t_2)\} \ . \quad (4)$$

The Kalman filtering algorithm processes the data in both the forward and backward directions to produce the final state estimates. The forward pass updates the predicted and filtered estimates of the state based on the current observation. The forward pass is summarized as follows [9]:

$$\begin{aligned} \mathbf{x}_{t|t} &= \mathbf{x}_{t|t-1} + K_t\epsilon_t \\ \epsilon_t &= \mathbf{y}(t) - \mathbf{x}_{t|t-1} - \mu_{k(t)} \\ \Sigma_{\epsilon_t} &= \Sigma_{t|t-1} + R_{k(t)} \\ K_t &= \Sigma_{t|t-1}\Sigma_{\epsilon_t}^{-1} \\ \Sigma_{t|t} &= \Sigma_{t|t-1} - K_t\Sigma_{\epsilon_t}K_t' \\ \Sigma_{t,t-1|t} &= (I - K_t)\, F_{k(t)}\Sigma_{t-1|t-1} \\ \mathbf{x}_{t+1|t} &= F_{k(t)}\mathbf{x}_{t|t} \\ \Sigma_{t+1|t} &= F_{k(t)}\Sigma_{t|t}F'_{k(t)} + Q_{k(t)} \ . \end{aligned} \quad (5)$$

After the forward pass is completed, the smoothed state estimates can be obtained via the backward recursions given as follows [9]:

$$\begin{aligned} J_t &= \Sigma_{t-1|t-1}F'_{k(t-1)}\Sigma_{t|t-1}^{-1} \\ \mathbf{x}_{t-1|T} &= \mathbf{x}_{t-1|t-1} + J_t\left[\mathbf{x}_{t|T} - \mathbf{x}_{t|t-1}\right] \\ \Sigma_{t-1|T} &= \Sigma_{t-1|t-1} + J_t\left[\Sigma_{t|T} - \Sigma_{t|t-1}\right]J_t' \\ \Sigma_{t,t-1|T} &= \Sigma_{t,t-1|t} + \left[\Sigma_{t|T} - \Sigma_{t|t}\right]\Sigma_{t|t}^{-1}\Sigma_{t,t-1|t} \ . \end{aligned} \quad (6)$$

The parameters $\{F_k, \mu_k, Q_k, R_k\}$ can now be estimated using the statistics computed by (5) and (6) based on the EM algorithm [9]. If $\left\{\tilde{F}_k, \tilde{\mu}_k, \tilde{Q}_k, \tilde{R}_k\right\}$ are the updated parameter values at an EM iteration, we have

$$\begin{aligned} \tilde{F}_k &= \left(\sum_{t=1:k(t)=k}^{T-1}\mathbf{x}(\widehat{t+1})\mathbf{x}'(t)\right) \\ &\quad \times \left(\sum_{t=1:k(t)=k}^{T-1}\mathbf{x}(\widehat{t)\mathbf{x}'}(t)\right)^{-1} \\ \tilde{\mu}_k &= \frac{1}{T(k)}\left[\sum_{t=1:k(t)=k}^{T}\left(\mathbf{y}(t) - \widehat{\mathbf{x}(t)}\right)\right] \\ \tilde{Q}_k &= \frac{1}{T(k)-1}\left[\left(\sum_{t=1:k(t)=k}^{T-1}\mathbf{x}(t+\widehat{1)\mathbf{x}'}(t+1)\right)\right. \\ &\quad \left. -\tilde{F}_k\left(\sum_{t=1:k(t)=k}^{T-1}\mathbf{x}(\widehat{t+1)\mathbf{x}'}(t)\right)'\right] \end{aligned}$$

$$\begin{aligned} \tilde{R}_k &= \frac{1}{T(k)}\left[\sum_{t=1:k(t)=k}^{T}\left(\mathbf{y}(t) - \widehat{\mathbf{x}(t)}\right)\left(\mathbf{y}(t) - \widehat{\mathbf{x}(t)}\right)'\right] \\ &\quad -\tilde{\mu}_k\tilde{\mu}_k' \end{aligned} \quad (7)$$

where $\widehat{a}$ indicates $E[a|\mathbf{y}_1^T]$ and $T(k)$ indicates the number of frames that have k as the cluster index.

Exact computation of the cluster indices $\{k(t)\}$ under the SLDM framework turns out to be practically impossible since it is exponentially proportional to the length of the given utterance [5]. To alleviate this difficulty, we employ a suboptimal approach based on the IMM algorithm which will be described in the following section [4]. First, the cluster indices, $\{k(t)\}$ are obtained by applying a conventional vector quantization (VQ) technique to each frame of feature vectors, and then the aforementioned EM algorithm is performed to obtain the initial estimate for the SLDM parameters in each speech cluster. Next, we apply the IMM algorithm which produces the probability of each cluster at each time conditioned on the given utterance. We treat the cluster index which shows the highest conditional probability at time t as $k(t)$, and carry out (5)-(7) to update the SLDM parameters.

3. Clean Feature Estimation with the IMM Algorithm

In this section, we propose a new technique to compensate the noisy features when the sequence of clean speech features is characterized by an SLDM which was originally proposed in [7]. Let $\mathbf{z}(t) = [z_1(t), z_2(t), \cdots, z_D(t)]'$ be a log spectrum of the noisy speech when the clean speech within the tth frame is corrupted by a background noise. Then, for $i = 1, 2, \cdots, D$,

$$z_i(t) = \log\left(\exp\left(y_i(t)\right) + \exp\left(n_i(t)\right)\right) \quad (8)$$

in which $\mathbf{y}(t) = [y_1(t), y_2(t), \cdots, y_D(t)]'$ and $\mathbf{n}(t) = [n_1(t), n_2(t), \cdots, n_D(t)]'$ represent the log spectra of the clean speech and the added noise, respectively. The purpose of feature compensation is to estimate the sequence of clean speech log spectra, $\{\mathbf{y}(1), \mathbf{y}(2), \cdots, \mathbf{y}(T)\}$ from a given sequence of noisy observations, $\{\mathbf{z}(1), \mathbf{z}(2), \cdots, \mathbf{z}(T)\}$. In order to make the non-linear relationship among $\mathbf{z}(t)$, $\mathbf{n}(t)$, and $\mathbf{y}(t)$ a more tractable one, (8) is approximated by a piecewise linear model given by

$$\mathbf{z}(t) = A_k\mathbf{y}(t) + B_k\mathbf{n}(t) + C_k \quad (9)$$

with the parameter k representing cluster index of $\mathbf{y}(t)$ is assumed to be k. The coefficient matrices $\{A_k, B_k, C_k\}$ shown in the above equation are obtained by the statistical linear approximation (SLA) algorithm that is based on the Taylor series expansion method [3].

In order to take into account the time-varying characteristics of the background noise, we adopt a simple noise process model described as

$$\mathbf{n}(t+1) = \mathbf{n}(t) + \xi(t) \quad (10)$$

where $\xi(t)$ is a zero-mean white Gaussian process with the covariance matrix Ψ which is a diagonal matrix with its diagonal elements being small positive numbers to account for the slow evolution of the noise characteristics. Combining (9) with (10) and (1) leads us to a linear state space model for each clean

speech cluster, which can be written as follows:

$$\begin{aligned}\begin{bmatrix}\mathbf{x}(t+1)\\ \mathbf{n}(t+1)\end{bmatrix} &= \begin{bmatrix}F_k & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{x}(t)\\ \mathbf{n}(t)\end{bmatrix}+\begin{bmatrix}\mathbf{w}_k(t)\\ \xi(t)\end{bmatrix}\\ \mathbf{z}(t) &= \begin{bmatrix}A_k & B_k\end{bmatrix}\begin{bmatrix}\mathbf{x}(t)\\ \mathbf{n}(t)\end{bmatrix}+\eta_k(t) \quad (11)\end{aligned}$$

in which

$$\eta_k(t) = A_k\mu_k + A_k\mathbf{v}_k(t) + C_k\ , \quad (12)$$

and it can be viewed as a Gaussian measurement noise process in the resulted model. From (11), it is noted that in contrast to the original IMM algorithm proposed in [2] where the state is solely defined in terms of the noise, now the state incorporates the clean speech components also. Let us define the augmented state vector, $\begin{bmatrix}\mathbf{x}'(t) & \mathbf{n}'(t)\end{bmatrix}'$ by $\tilde{\mathbf{x}}(t)$ and

$$\begin{aligned}\tilde{\mathbf{x}}_{t|\tau} &= E\left[\tilde{\mathbf{x}}(t)|\mathbf{z}_1^\tau\right]\\ \tilde{\Sigma}_{t|\tau} &= E\left[\tilde{\mathbf{x}}(t)\tilde{\mathbf{x}}'(t)|\mathbf{z}_1^\tau\right]-\tilde{\mathbf{x}}_{t|\tau}\tilde{\mathbf{x}}'_{t|\tau} \quad (13)\end{aligned}$$

in which $\mathbf{z}_1^\tau = \{\mathbf{z}(1), \mathbf{z}(2), \cdots, \mathbf{z}(\tau)\}$ denotes the sequence of noisy observation vectors up to time τ.

Once the linear state space model is specified as in (11) for each speech cluster, filtered state estimates, $\left\{\tilde{\mathbf{x}}_{t|t}, \tilde{\Sigma}_{t|t}\right\}$ can be sequentially computed by applying the IMM algorithm that consists of several steps summarized in the following [4]:

- *Mixing step*: the state estimates obtained from each cluster in the previous time are combined together to produce a single set of state estimates, which is provided to each Kalman filter as an initial statistic.
- *Kalman step*: the conventional Kalman update is carried out conditioned on the initial estimates computed from the *Mixing step*.
- *Probability computation step*: the a posteriori probability associated with each cluster is updated.
- *Output generation step*: the state estimates are generated by combining the estimates of all the clusters. In our case, this step is the same to the *Mixing step*.

After computing $\left\{\tilde{\mathbf{x}}_{t|t}, \tilde{\Sigma}_{t|t}\right\}$, the clean speech feature estimate, $\hat{\mathbf{y}}(t)$ at time t is obtained according to the minimum mean square error (MMSE) criterion such that

$$\begin{aligned}\hat{\mathbf{y}}(t) &= E\left[\mathbf{y}(t)|\mathbf{z}_1^t\right]\\ &= \sum_{k=1}^{K} p\left(k(t)=k|\mathbf{z}_1^t\right) E\left[\mathbf{y}(t)|k(t)=k, \mathbf{z}_1^t\right].\end{aligned} \quad (14)$$

Since the sequence of clean speech feature vectors is characterized by an SLDM, temporal correlations among adjacent frames can be efficiently utilized during the feature compensation procedure.

4. SS and Soft Decision SLDM-IMM

In this section, we briefly review the spectral subtraction (SS) technique used for feature compensation in noisy environments and propose a new feature compensation technique which combine the SLDM-IMM approach with the SS method through the soft decision.

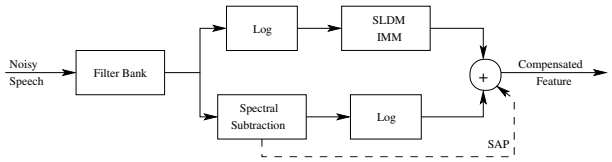

Figure 1: Block diagram of the soft decision SLDM-IMM algorithm.

4.1. Spectral Subtraction (SS)

In contrast to the IMM-based approach, the SS approach is performed in the linear spectral domain. Let $\mathbf{Z}(t)$=[$Z_1(t)$, $Z_2(t)$, $\cdots$, $Z_D(t)$], $\mathbf{Y}(t)$=[$Y_1(t), Y_2(t), \cdots, Y_D(t)$], $\mathbf{N}(t)$ = [$N_1(t)$, $N_2(t)$, $\cdots$, $N_D(t)$] denote the linear amplitude spectra of the noisy speech signal, clean speech signal, added noise signal, respectively. Two hypotheses H_0 and H_1, which indicate speech absence and presence, are described as follows:

$$H_0 : \mathbf{Z} = \mathbf{N}, \quad H_1 : \mathbf{Z} = \mathbf{Y} + \mathbf{N}. \quad (15)$$

The estimate for the clean-speech spectrum is obtained by

$$\hat{Y}_i = \tilde{G}_i \cdot Z_i, \quad i = 1, 2, ..., D \quad (16)$$

where $\hat{Y}_i$ represents the enhanced ith spectral component and $\tilde{G}_i$ indicates the applied gain.

For the spectral gains $\{\tilde{G}_i\}$, we use the noise suppression rule proposed by Ephraim and Malah [8] (which we call the EMSR method). According to the EMSR rule the gain appears as a function of two variables such that

$$\tilde{G}_i = G(\eta_i, \gamma_i) \quad (17)$$

in which η_i and γ_i are reffered to the *a priori* and *a posteriori* SNRs, respectively. In the EMSR method, η_i and γ_i play the key role, and their estimates are efficiently updated for each time by means of the decision-directed approach [8].

4.2. Soft Decision SLDM-IMM

The IMM-based approach has the drawback of generating many undesirable insertion errors in the recognition results. Because of inappropriate model approximation, non-speech regions are not compensated well, mostly less suppressed than expected. On the other hand, the spectral subtraction approach with the speech absence probability (SAP) more effectively suppress the background noise which exists in the non-speech regions. But the spectral subtraction may distort some active speech regions, which gives a number of substitution errors [6].

In order to reduce the number of insertion errors while maintaining the good performance of the SLDM-IMM approach during active speech regions, we import soft decision IMM technique [6] into the SLDM-IMM feature compensation algorithm. A block diagram of the soft decision SLDM-IMM algorithm is shown in Fig.1 where we implement both the SLDM-IMM and the SS modules in parallel. In this process, the two separate estimates provided by the SLDM-IMM and the SS modules are combined. Furthermore, contribution of each module is controlled in accordance with the SAP which is computed in the SS module. Let $\hat{\mathbf{y}}_{\mathbf{SLDM}}$ be the clean speech feature estimate provided by the SLDM-IMM algorithm, and $\hat{\mathbf{y}}_{\mathbf{SS}}$ be the estimate obtained based on the SS technique. Then, the soft

decision SLDM-IMM algorithm combines the two estimates as follows:

$$\hat{\mathbf{y}} = (1 - p(H_0|\mathbf{z}))\hat{\mathbf{y}}_{\mathbf{SLDM}} + p(H_0|\mathbf{z})\hat{\mathbf{y}}_{\mathbf{SS}} \quad (18)$$

where $p(H_0|\mathbf{z})$ indicates the SAP with $\mathbf{z}$ being the given noisy speech spectrum and H_0 indicate the hypothesis of the speech absence.

Given a noisy speech spectrum $\mathbf{Z}$, we can compute the SAP, $p(H_0|\mathbf{Z})$ such that

$$p(H_0|\mathbf{Z}) = \frac{p(\mathbf{Z}|H_0)p(H_0)}{p(\mathbf{Z}|H_0)p(H_0) + p(\mathbf{Z}|H_1)p(H_1)} \quad (19)$$

where $p(H_0)(= 1-p(H_1))$ is the *a priori* probability for speech absence. The likelihoods $\{p(\mathbf{Z}|H_0), p(\mathbf{Z}|H_1)\}$ are specified based on the assumed statistical models for the speech and noise spectra distributions. If $\mathbf{Y}$ and $\mathbf{N}$ are characterized by independent zero-mean complex Gaussian pdfs, we have

$$p(H_0|\mathbf{Z}) = \frac{1}{1 + \frac{p(H_1)}{p(H_0)} \prod_{i=1}^{D} \Lambda_i(Z_i)} \quad (20)$$

in which $\Lambda_i(Z_i)$ is the likelihood ratio computed for the ith spectral component as given by

$$\Lambda_i(Z_i) = \frac{1}{1+\eta_i} \exp(\frac{\gamma_i \eta_i}{1+\eta_i}). \quad (21)$$

The SAP is usually applied to modify the gain function shown in (17), and it also becomes an important parameter in the soft decision SLDM-IMM approach.

5. Experimental Results

Performance of the proposed soft decision SLDM-IMM algorithm was evaluated on the AURORA2 database [10]. Feature compensation was performed in the log spectral domain, and the compensated log spectra were converted to the cepstral coefficients through discrete cosine transform (DCT). In the SLDM-IMM algorithm, clean speech log spectra were modelled by a mixture of 128 Gaussian distributions with diagonal covariance matrices. The SLDM training algorithm described in Section 2 was applied to the divided clusters. In our experiments, all the parameters $\{F_k, Q_k, R_k\}$ specifying the structure of SLDM were maintained to be diagonal matrices. For the purpose of comparison, we also tested the previous IMM algorithm [4] where the clean speech feature vectors are assumed to be distributed according to a GMM without any consideration for the temporal correlation.

The recognition results obtained from the AURORA2 task in clean training condition are shown in Table 1. The term GMM-IMM refers to the IMM algorithm based on the GMM assumption for clean speech feature distribution while SLDM-IMM indicate compensation that is based on SLDM, respectively. The term SD means the algorithm performed with the soft decision. The SLDM-IMM approach shows better results compared with GMM-IMM. It is evident from these results that SLDM is a better way to characterize the clean speech feature distribution than GMM. We can also find the combination of the SLDM-IMM and the SS outperformed the other approaches.

	set A	set B	set C	Average
Baseline	61.34	55.75	66.14	60.06
GMM-IMM	81.09	81.82	76.17	80.40 (46.50)
SD GMM-IMM	83.05	83.37	78.00	82.17 (53.89)
SLDM-IMM	83.20	83.75	78.09	82.40 (57.36)
SD SLDM-IMM	83.98	84.20	81.13	83.50 (60.12)

Table 1: Word accuracies(%) over AURORA2 database for clean training condition (Relative improvements(%) compared to the baseline system).

6. Conclusions

In this paper, we have presented a new feature compensation technique based on the combination of the SLDM-IMM approach and the SS approach depending on a soft decision for speech activity. A major advantage of the SLDM is its capability to describe the temporal correlation of the original signal in a systematic way. The proposed soft decision SLDM-IMM approach has been found to improve the estimation accuracy in both non-speech regions and active speech regions, and this induces the improvements of the recognition results.

7. References

[1] P. J. Moreno, B. Raj and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," *Proc. of the ICASSP*, Atlanta, GA, May 1996.

[2] N. S. Kim, "IMM-based estimation for slowly evolving environments," *IEEE Signal Processing Letters*, vol. 5, no. 6, pp. 146-149, June 1998.

[3] N. S. Kim, "Statistical linear approximation for environment compensation," *IEEE Signal Processing Letters*, vol. 5, no. 1, pp. 8-10, Jan. 1998.

[4] N. S. Kim, "Feature domain compensation of nonstationary noise for robust speech recognition," *Speech Comm.*, Vol. 37, pp. 231-248, July 2002.

[5] J. Droppo and A. Acero, "Noise robust speech recognition with a switching linear dynamic model," *Proc. of the ICASSP*, Montreal, Canada, pp. 953-956, May 2004.

[6] N. S. Kim, Y. J. Kim, and H. W. Kim, "Feature compensation based on soft decision," *IEEE Signal Processing Letters*, vol. 11, no. 3, pp. 378-381, Mar. 2004.

[7] N. S. Kim, W. Lim, and R. M. Stern, "Feature compensation based on switching linear dynamic model," *IEEE Signal Processing Letters*, June 2005.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Processing, vol. 32, pp. 1109-1121, Dec. 1984.

[9] V. Digalakis, J. R. Rohlicek, M. Ostendorf, "ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition," *IEEE Trans. Speech, Audio Proc.*, pp. 431-442, Oct. 1993.

[10] H. -G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proc. of the ICSLP*, pp.16-20, Oct. 2000.