SPEECH FEATURE ESTIMATION UNDER THE PRESENCE OF NOISE WITH A SWITCHING LINEAR DYNAMIC MODEL

Jianping Deng, Martin Bouchard, and Tet Hin Yeap

School of Information Technology and Engineering, University of Ottawa
800 King Edward, Ottawa (Ontario), K1N 6N5, Canada
jdeng,bouchard,tet@site.uottawa.ca

ABSTRACT

This paper presents an approach to enhance speech feature estimation in the log spectral domain under noisy environments. A higher-order switching linear dynamic model (SLDM) is explored as a parametric model for the clean speech distribution, which enforces a state transition in the feature space and captures the smooth time evolution of speech conditioned on the state sequence. The clean speech components are estimated by means of an Interacting Multiple Model (IMM) algorithm. Our experimental results show that increasing the order of the linear dynamic model in the SLDM and the introduction of transition probabilities among the linear dynamic models can improve the performance of SLDM systems in feature compensation for robust speech recognition.

1. INTRODUCTION

The performance degradation of a speech recognizer in the presence of additive noise is one of the major problems that still remain unsolved in the real-field applications of speech recognition technology. Towards solving the noise robustness problem, in the past few years a variety of noise compensation techniques have been developed. One of the prevailing approaches is model based feature enhancement [1]-[7]. It is believed that enhancement in the cepstral or log domain is most desirable if the purpose of speech enhancement is for robust speech recognition, since this is the domain as close as possible to the back end of the recognizer [5]. Such systems include a model for speech, and also often a model for noise within the enhancement algorithm, where the speech is treated as the nonlinear obscuring influence that prevents us from observing the noise [2]-[4]. When the speech model is a Gaussian mixture model (GMM), each frame of speech is enhanced independently.

To take advantage of the temporal correlations among adjacent frames of a speech signal, a switching linear dynamic model (LDM) was proposed in [6]. In the switching LDM, as time progresses the signal passes through several distinct states. The linear dynamics of the model capture the smooth time evolution of speech, conditioned on the state sequence. Very recently, Kim et al [7] further improved this technique by incorporating the SLDM of speech with a noise model. Both the noise and clean speech components are estimated simultaneously during the feature compensation using an IMM algorithm.

This recent work by Kim et al [7] bears some similarity with our method described in this paper. However, our work was developed independently and there are three important differences between our method and the method presented in [7]. First, the SLDM in [7] uses a first-order AR to predict the t^{th} speech spectral vector. In our SLDM, we investigate the use of a second order AR model for spectral vector prediction. To the best of our knowledge, the explicit use of higher-order SLDM to model the spectral vector for feature compensation has not been reported so far.

Second, the SLDM used in [7] assumes time dependence among the continuous speech in one state, but not among the discrete states. To further take advantage of the temporal correlations among adjacent frames of the speech signal, our work includes the time-dependence among the discrete state variables by augmenting the SLDM with discrete state transition probabilities. As suggested by Droppo et al [6] in their future work, the state transition probabilities of HMM tend to eliminate single frame errors in the output.

Third, the estimation process of the SLDM parameters is different. In [7], the statistics to estimate the SLDM parameters were computed based on a cluster index that was obtained with a suboptimal approach. The current paper finds the maximum likelihood (ML) estimates of the parameters for the SLDM.

We present preliminary results demonstrating that, even with relatively small model sizes, substantial word error rate improvement can be obtained compared with a baseline recognizer. The rest of this paper is organized as follows: Section 2 describes how to model the sequence of clean speech features by a switching LDM. The method to compensate the noisy feature is presented in Section 3. Section 4 describes the experiments and presents the results. Conclusions are given in Section 5.

2. SLDM FOR CLEAN SPEECH FEATURES

To describe the clean speech features distribution, a switching linear dynamic model (LDM) obeys the system equation

$$\mathbf{y}_t = \boldsymbol{\mu}_{s_t=i} + \mathbf{B}_{1,s_t=i}\mathbf{y}_{t-1} + ... + \mathbf{B}_{p,s_t=i}\mathbf{y}_{t-p} + \mathbf{e}_t \quad (1)$$

Equation (1) could be rewritten in the short-hand form

$$\mathbf{y}_t = \boldsymbol{\mu}_{s_t=i} + \mathbf{B}_{s_t=i}\mathbf{x}_t + \mathbf{e}_t \quad (2)$$

where $\mathbf{x}_t$ is the column vector $[\mathbf{y}_{t-1}, \cdots \mathbf{y}_{t-p}]^T$ and $\mathbf{B}_{s_t=i}$ is matrix $\left(\mathbf{B}_{1,s_t=i} \quad \cdots \quad \mathbf{B}_{p,s_t=i}\right)$. $\mathbf{B}$ and $\boldsymbol{\mu}$ are dependent on a hidden variable s_t at each time *t*. The state-dependent residual $\mathbf{e}_t$ has a Gaussian distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}_{s_t}$. The graphical representation of the switching LDM used is shown in Figure 1, where shaded nodes are observed, and clear nodes are hidden. It depicts the special case where p=1.

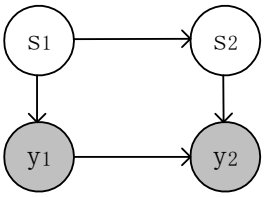

Fig. 1 Graphical representation of the switching LDM

It is seen from Fig.1 that unlike the SLDM presented in [6] and [7], the time-dependence among the discrete state variables and adjacent frames of speech are included. In this paper, the order p of the linear dynamic model is set to 2. Assume that the discrete hidden states follow a Markov chain process with N states and that the state transition matrix is defined as

$$z(i,j) = P(s_{t+1} = j \mid s_t = i), \quad i,j \in (1,\cdots,N)$$

Given the state sequence, the observation likelihood for $\mathbf{y}_t$ given that the LDM is at state i at time t is

$$\begin{aligned}&\Pr(\mathbf{y}_t \mid \mathbf{x}_t, s_t = i) \\ &= (2\pi)^{-d/2} \mid \boldsymbol{\Sigma}_i \mid^{-\frac{1}{2}} \exp(-\tfrac{1}{2}[\mathbf{y}_t - \mathbf{B}_i\mathbf{x}_t - \boldsymbol{\mu}_i]'\boldsymbol{\Sigma}_i^{-1}[\mathbf{y}_t - \mathbf{B}_i\mathbf{x}_t - \boldsymbol{\mu}_i])\end{aligned} \quad (3)$$

The parameters $\{\mathbf{B}_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ associated with the specified SLDM can be estimated from a set of clean speech training data using the standard EM algorithm [8]. The algorithm then iterates, using the current parameter estimate to compute the expected state occupancy

$$\gamma_t(i) = \Pr(s_t = i \mid \mathbf{y}_{1:T}) \quad (4)$$

where $\gamma_t(i)$ is the probability that the HMM is in state i at time t, and it is calculated by the forward-backward algorithm [9]. The EM algorithm requires us to maximize the following expected log-likelihood, $Q(M_0, M)$, by choosing the parameters of the new model M.

$$Q(M_0, M) = \sum_i \sum_t P_{M_0}(s_t = i \mid \mathbf{y}_{1:T}) \log P_M(\mathbf{y}_t \mid \mathbf{x}_t, s_t) \quad (5)$$

M_0 is the model corresponding to an initial estimate of the parameters, $P_{M_0}(s_t \mid \mathbf{y}_{1:T})$ stands for the probability of s_t conditioned on the observation sequence $\mathbf{y}_{1:T}$, calculated using the parameters of the model M_0. To present the result of this maximization, the following expected sufficient statistics are first introduced:

$$\begin{aligned}
S_{YY',i} &= \sum_t \gamma_t(i) E_{ti}[\mathbf{y}_t \mathbf{y}_t'] \\
S_{Y,i} &= \sum_t \gamma_t(i) E_{ti}[\mathbf{y}_t] \\
S_{XX',i} &= \sum_t \gamma_t(i) E_{ti}[\mathbf{x}_t \mathbf{x}_t'] \\
S_{X,i} &= \sum_t \gamma_t(i) E_{ti}[\mathbf{x}_t] \\
S_{XY',i} &= \sum_t \gamma_t(i) E_{ti}[\mathbf{x}_t \mathbf{y}_t'] \\
S_{YX',i} &= \sum_t \gamma_t(i) E_{ti}[\mathbf{y}_t \mathbf{x}_t']
\end{aligned} \quad (6)$$

Maximizing $Q(M_0, M)$ with respect to $\mathbf{B}_i$ and setting the derivative to zero, the following equation is obtained

$$\mathbf{B}_i = (S_{YX',i} - \boldsymbol{\mu}_i S_{X,i}')S_{XX',i}^{-1} \quad (7)$$

Likewise, setting the derivatives of the objective function with respect to $\boldsymbol{\mu}_i$ to zero, another vector equation is obtained

$$\boldsymbol{\mu}_i = \frac{S_{Y,i} - \mathbf{B}_i S_{X,i}}{\gamma_i} \quad (8)$$

where $\gamma_i = \sum_t \gamma_t(i)$. The new estimates for $\boldsymbol{\mu}_i$ and $\mathbf{B}_i$ can now be obtained by solving the pair of simultaneous equations. Likewise, the re-estimation formula for $\boldsymbol{\Sigma}_i$ is:

$$\boldsymbol{\Sigma}_i = \frac{1}{\gamma_i}(S_{YY',i} - S_{YX',i}\mathbf{B}_i' - \mathbf{B}_i S_{XY',i} + \mathbf{B}_i S_{XX',i}\mathbf{B}_i' - \gamma_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i') \quad (9)$$

The re-estimation formula for the state transition matrix is the same as an HMM [9]:

$$Z(i,j) = \frac{\sum_t \Pr(s_{t-1} = i, s_t = j \mid \mathbf{y}_{1:T})}{\gamma_i} \quad (10)$$

3. CLEAN FEATURES ESTIMATION

Assume that speech and noise are mixed linearly in the time domain. This corresponds to a nonlinear mixing in the log spectrum feature space as follows [1]:

$$\mathbf{o}_t = f(\mathbf{y}_t, \mathbf{n}_t) = \mathbf{y}_t + \log(\mathbf{I} + \exp(\mathbf{n}_t - \mathbf{y}_t)) \quad (11)$$

in which $\mathbf{o}_t$, $\mathbf{y}_t$ and $\mathbf{n}_t$ respectively represent the log spectrum of noisy speech, the hypothetical clean speech and noise at the t^{th} frame.

To take into account the time-varying characteristics of the background noise, we model the sequence of noise features as the output of a first-order auto-regressive (AR) system excited by a zero mean Gaussian process $\mathbf{v}$ with a covariance matrix $\mathbf{Q}_n$ as follows [2]:

$$\mathbf{n}_t = \mathbf{A}\mathbf{n}_{t-1} + \boldsymbol{\mu}_n + \mathbf{v}_t \quad (12)$$

In the experiments for this paper, a noise-type specific model was built from 30 seconds of training samples of the noise. Combining equation (2) with (11) and (12) leads to a nonlinear state space model as:

$$\mathbf{z}_t = \Phi_{s_t}\mathbf{z}_{t-1} + \mathbf{u}_{s_t} + G\bar{\mathbf{e}}_{s_t} \quad (13)$$

$$\mathbf{o}_t = F\bar{\mathbf{y}}_t + \log(\mathbf{1} + \exp(\mathbf{n}_t - F\bar{\mathbf{y}}_t)) \quad (14)$$

where

$$\mathbf{z}_t = \begin{bmatrix}\bar{\mathbf{y}}_t \\ \mathbf{n}_t\end{bmatrix} \quad \bar{\mathbf{y}}_t = \begin{bmatrix}\mathbf{y}_t \\ \mathbf{y}_{t-1}\end{bmatrix} \quad \bar{\mathbf{e}}_{s_t} = [\mathbf{e}_{s_t} \quad \mathbf{v}_t]^T$$

$$\Phi_{s_t} = \begin{bmatrix}C_{s_t} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}\end{bmatrix} \quad \mathbf{u}_{s_t} = [\boldsymbol{\mu}_{s_t} \quad \mathbf{0} \quad \boldsymbol{\mu}_n]^T$$

$$C_{s_t} = \begin{bmatrix}\mathbf{B}_{s_t} & \\ \mathbf{I} & \mathbf{0}\end{bmatrix} \quad G = \begin{bmatrix}G_Y & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\end{bmatrix} \quad \mathbf{Q}_j = \begin{bmatrix}\Sigma_j & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_n\end{bmatrix}$$

$$F = \begin{bmatrix}\mathbf{I} & \mathbf{0}\end{bmatrix} \quad G_Y = [\mathbf{I} \quad \mathbf{0}]^T$$

Both noise and clean speech components are incorporated in the state space, and are estimated simultaneously.

The major obstacle to use the switching LDM for enhancement is the computational burden that it brings. Optimal minimum mean squared error estimators involve a bank of filters tuned to all the possible parameter histories, which makes the cost in computations grow exponentially with data length [10]. To solve this problem, we show in this paper how the Interacting Multiple Model (IMM) algorithm [10] can be adapted to the nonlinear state-space model of feature dynamics presented above, to provide a sub-optimal approximation solution. A block diagram of the IMM algorithm is shown in Fig. 2.

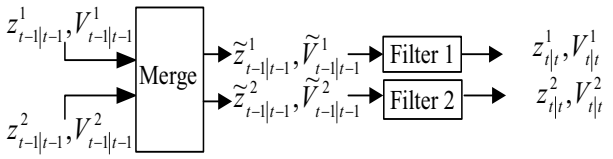

Fig. 2 IMM algorithm

There are *N* filters, each of which is supplied with a different input. Let us define

$$\tilde{\mathbf{z}}^j_{t-1|t-1} = E(\mathbf{z}_{t-1} \mid s_t = j, \mathbf{o}_{1:t-1})$$

$$\tilde{\mathbf{V}}^j_{t-1|t-1} = \mathrm{cov}(\mathbf{z}_{t-1} \mid s_t = j, \mathbf{o}_{1:t-1})$$

$$W^{i|j}_{t-1|t-1} = \Pr\left(s_{t-1} = i \middle| s_t = j, \mathbf{o}_{1:t-1}\right) \quad (15)$$

The algorithm mixes the estimates according to the Markov transition probability, in order to allow the system to react to changes in the model in force. In that way, the input to the j^{th} filter becomes the best estimate of the state at time instant *t*-1, conditioned on the event that model *j* is in force at time instant *t* (the new sample time). $\tilde{\mathbf{z}}^j_{t-1|t-1}$ and $\tilde{\mathbf{V}}^j_{t-1|t-1}$ are then obtained according to

$$\tilde{\mathbf{z}}^j_{t-1|t-1} = \sum_i W^{i|j}_{t-1|t-1}\mathbf{z}^i_{t-1|t-1} \quad (16)$$

$$\begin{aligned}\tilde{\mathbf{V}}^j_{t-1|t-1} &= \sum_i W^{i|j}_{t-1|t-1}\mathbf{V}^i_{t-1|t-1} \\ &+ \sum_i W^{i|j}_{t-1|t-1}\left(\mathbf{z}^i_{t-1|t-1} - \tilde{\mathbf{z}}^j_{t-1|t-1}\right)\left(\mathbf{z}^i_{t-1|t-1} - \tilde{\mathbf{z}}^j_{t-1|t-1}\right)^T\end{aligned} \quad (17)$$

The mixing probability $W^{i|j}_{t-1|t-1} = \Pr\left(s_{t-1} = i \middle| s_t = j, \mathbf{o}_{1:t-1}\right)$ is computed recursively with Bayes' rule

$$\begin{aligned}W^{i|j}_{t-1|t-1} &= \Pr\left(s_{t-1} = i \middle| s_t = j, \mathbf{o}_{1:t-1}\right) \\ &= \frac{\Pr(s_t = j \mid s_{t-1} = i, \mathbf{o}_{1:t-1})\Pr(s_{t-1} = i \mid \mathbf{o}_{1:t-1})}{\sum_i \Pr(s_t = j \mid s_{t-1} = i, \mathbf{o}_{1:t-1})\Pr(s_{t-1} = i \mid \mathbf{o}_{1:t-1})}\end{aligned} \quad (18)$$

The extended Kalman filter algorithm for each filter becomes:

$$\mathbf{z}^j_{t|t-1} = \boldsymbol{\Phi}_j\tilde{\mathbf{z}}^j_{t-1|t-1} + \mathbf{u}_j \quad (19)$$

$$\mathbf{V}^j_{t|t-1} = \boldsymbol{\Phi}_j\tilde{\mathbf{V}}^j_{t-1|t-1}\boldsymbol{\Phi}_j^T + G\mathbf{Q}_jG^T \quad (20)$$

$$\mathbf{e}^j_t = \mathbf{o}_t - f(\mathbf{z}^j_{t|t-1}) \quad (21)$$

$$\mathbf{K}^j_t = \mathbf{V}^j_{t|t-1}\mathbf{H}^T(\mathbf{H}\mathbf{V}^j_{t|t-1}\mathbf{H}^T)^{-1} \quad (22)$$

$$\mathbf{z}^j_{t|t} = \mathbf{z}^j_{t|t-1} + \mathbf{K}^j_t\mathbf{e}^j_t \quad (23)$$

$$\mathbf{V}^j_{t|t} = \left(\mathbf{I} - \mathbf{K}^j_t\mathbf{H}\right)\mathbf{V}^j_{t|t-1} \quad (24)$$

where $\mathbf{H} = [\mathbf{A}_0 \quad \mathbf{B}_0]$, $\mathbf{A}_0$ and $\mathbf{B}_0$ are the Jacobian matrices with respect to $\bar{\mathbf{y}}$ and $\mathbf{n}$. From our experiments, we found that the extended Kalman filter can bring better results than approximating the observation equation (11) by a piecewise linear model and then using the linear Kalman filter, as was done in previous work [3][7].

The model probabilities $\Pr(s_t = i \mid \mathbf{o}_{1:t})$ are updated according to

$$\begin{aligned}\Pr(s_t = j \mid \mathbf{o}_{1:t}) &= \tfrac{1}{\Omega}\sum_i \Pr(s_{t-1} = i, s_t = j, \mathbf{o}_t \mid \mathbf{o}_{1:t-1}) \\ &= \tfrac{1}{\Omega}\sum_i \Pr(\mathbf{o}_t \mid s_{t-1} = i, s_t = j, \mathbf{o}_{1:t-1})\Pr(s_{t-1} = i, s_t = j \mid \mathbf{o}_{1:t-1}) \\ &= \tfrac{1}{\Omega}\sum_i \Pr(\mathbf{o}_t \mid s_{t-1} = i, s_t = j, \mathbf{o}_{1:t-1})\Pr(s_t = j \mid s_{t-1} = i, \mathbf{o}_{1:t-1})\Pr(s_{t-1} = i \mid \mathbf{o}_{1:t-1})\end{aligned} \quad (25)$$

where Ω is a scale factor.

$$\begin{aligned}&\Pr(\mathbf{o}_t \mid \mathbf{o}_{1:t-1}, s_{t-1} = i, s_t = j) \\ &= N\left(\mathbf{e}^j_t; \mathbf{0}, \mathbf{H}\mathbf{V}^j_{t|t-1}\mathbf{H}\right)\end{aligned} \quad (26)$$

The estimated output at time *t* is calculated according to

$$\mathbf{z}_{t|t} = \sum_j \Pr(s_t = j \mid \mathbf{o}_{1:t})\mathbf{z}^j_{t|t} \quad (27)$$

4. EXPERIMENTS

The performance of the proposed method was evaluated with continuous digits recognition experiments. The speech data used for the experiments is from the Numbers v1.3 corpus provided by the Oregon Health & Science University (OGI). The corpus is a collection of 8kHz telephone speech data [11]. A 12th order cepstral coefficient vector was derived for each frame of 32ms. The derived cepstrum vectors and their first and second order derivatives were used for recognition. The recognition system used in the experiments was based on a continuous HMM model. Both the training and the

recognition phases were performed using the HTK toolbox [12]. In the testing stage, 100 utterances with fixed-length 5 connected digits from the database were used. Background noises (from the ITU-T Supplement P.23 database) were artificially added to the speech signals by a computer, with SNR varying from 0 dB to 15dB.

Speech features from 50 clean utterances were used to train the SLDM parameters, as described in section 2. The number of states used for the SLDM was 8. For each frame, the noisy 20^{th} order mel-scaled log spectrum was transformed to clean feature estimation. Figs. 3,4 show the average word error rate (WER) for white noise and babble noise mixtures. The proposed SLDM was tested with an AR model order of 1 and 2. The results are presented by 'proposed-p=1' and 'proposed-p=2' in the figures. For the purpose of comparison, we also tested with the previous IMM-GMM algorithm [4], where the clean speech feature vectors were modeled by a mixture of 18 GMM distributions, without any consideration for the temporal correlation. We also tested the IMM-SLDM algorithm proposed in [7] but with extended Kalman filter, and the SLDM was modeled using the method described in [6], as we found the structure of SLDM in [6] and [7] to be similar, the only difference being in the methods for finding the sufficient statistics to update the LDM parameters in the M step. The results are represented by 'IMM-SLDM1' in the figures. 'No NR' denotes a result without noise reduction.

The results show that the switching linear dynamic model approach with either the SLDM1 or the proposed SLDM is effective at improving the recognition performance and provides better results compared with the IMM-GMM. For both white noise and babble noise, the 'proposed-p=1' and 'proposed-p=2' show their superiority to their SLDM1 counterpart. At the same time, for all SNRs, the 'proposed-p=2' outperforms the 'proposed-p=1'. We can conclude that the second order SLDM is a better way to characterize the clean speech features distribution, and that the introduction of transition probabilities among linear dynamic models can further improve the performance of SLDM systems.

5. CONCLUSIONS

This paper explores the use of a switching linear dynamic model (SLDM) for speech features enhancement under the presence of noise in the log-spectral domain. The SLDM can capture the temporal correlations among adjacent frames of speech in a more accurate way compared with previous work. The simulation results show that this approach can reduce the word error rate, even with a simple structure specification for the SLDM.

6. REFERENCES

[1] P.J, Moreno, Speech recognition in noisy environments, *Ph.D Thesis*, Carnegie Mellon University, 1996

[2] B. Raj, R. Singh, and R. Stern "On tracking noise with linear dynamical system models", *IEEE ICASSP'04*, Vol. 1, pp. I-965-I-968, May 2004

[3] N.S. Kim, "IMM-based estimation for slowly evolving environments", *IEEE Signal Processing Letters*, Vol. 5, pp. 146–149, June 1998

[4] J. Deng, M. Bouchard, and T.H. Yeap, "Noise compensation using interacting multiple Kalman filters", *InterSpeech 2005*, pp. 949-952, Sept. 2005

[5] L. Deng, J. Droppo and A. Acero, "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features", *IEEE Trans. Speech and Audio Processing*, Vol. 12, pp. 218-233, May 2004

[6] J. Droppo and A. Acero, "Noise robust speech recognition with a switching linear dynamic model", *IEEE ICASSP'04*, Vol. 1, pp. I-953- I-956, May 2004

[7] N.S. Kim, W. Lim, and R. Stern, "Feature compensation based on switching linear dynamic model", *IEEE Signal Processing Letters*, Vol. 12, pp. 473-476, June, 2005

[8] P. Kenny, M. Lenning, and P. Mermelstein, "A linear predictive HMM for vector-valued observations with applications to speech recognition", *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 38, pp. 220-225, Feb. 1990

[9] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Application in Speech recognition", *Proc. IEEE*, Vol. 77, pp. 257-286, Feb. 1989

[10] Z. Ma, L. Deng, "Efficient decoding strategies for conversational speech recognition using a constrained nonlinear state space model", *IEEE Trans. Speech and Audio Processing*, Vol. 11, pp. 590-602, Nov. 2003

[11] http://www.cslu.ogi.edu/corpora/

[12] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland, *The HTK book – version 2.2*, *Entropic*, 1999

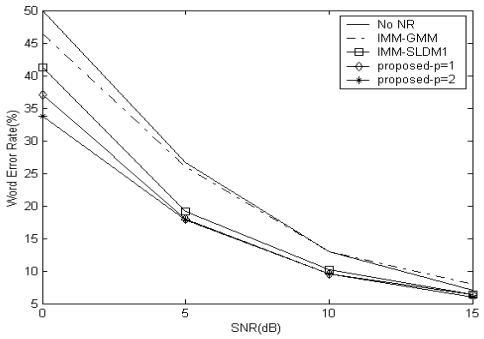

Fig.3 Results for speech mixed with white noise

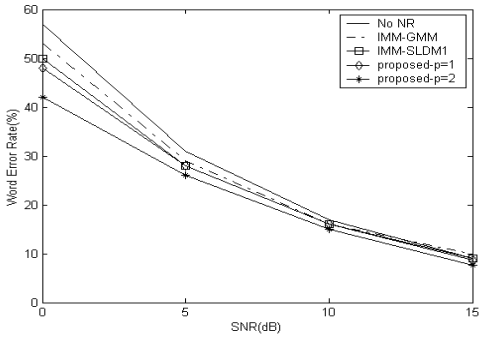

Fig.4 Results for speech mixed with babble noise