

Model-based Noisy Speech Recognition with Environment Parameters Estimated by Noise Adaptive Speech Recognition with Prior

Kaisheng Yao^{*}, Kuldip K. Paliwal[†] and Satoshi Nakamura[‡]

^{*}Institute for Neural Computation, University of California at San Diego

[†]School of Microelectronic Engineering, Griffith University, Australia

[‡]ATR Spoken Language Translation Research Laboratories, Kyoto, Japan

kyao@ucsd.edu k.paliwal@griffith.edu.au nakamura@slt.atr.co.jp

Abstract

We have proposed earlier a noise adaptive speech recognition approach for recognizing speech corrupted by nonstationary noise and channel distortion. In this paper, we extend this approach. Instead of maximum likelihood estimation of environment parameters (as done in our previous work), the present method estimates environment parameters within the Bayesian framework that is capable of incorporating prior knowledge of the environment. Experiments are conducted on a database that contains digit utterances contaminated by channel distortion and nonstationary noise. Results show that this method performs better than the previous methods.

1. Introduction

Speech recognition has to be carried out often under adverse environments. These environments cause distortions (mainly additive background noise and channel distortion) in the speech signal. Because of this distortion, there is a mismatch between the pre-trained models and the test speech signal to be recognized. This mismatch causes degradation in speech recognition performance (the amount of degradation depends on the type and amount of distortion caused by the environment). Among many approaches for handling this mismatch problem, one common approach is to assume explicit model for representing environmental effects on speech features [1] and use this model to construct a transformation which is applied either in the model space or feature space to decrease the mismatch. Though this model-based approach shows significant improvement, most of the research reported with this approach is focused on stationary noise distortion. Since the distortions introduced by adverse environments are nonstationary, it is necessary to devise methods that can cope up with nonstationary distortions and improve robustness of a speech recognition system under such conditions.

A number of speech recognition methods have been proposed in the literature to cope up with nonstationary environments. They can be categorized into two approaches. In the first approach, time-varying environment sources are modeled by hidden Markov models (HMMs) or Gaussian mixture models (GMMs) that are trained by prior measurement of environments, so that noise compensation is a task of identification of the underlying state/mixture sequences of the noise HMM/Mixtures, e.g., [1]. In the second approach, environment parameters are assumed to be time varying and need to be estimated. We have proposed earlier a noise-adaptive speech recognition approach [2] which uses a maximum

likelihood estimation method for estimating the time varying environment parameters and compensates for environment effects sequentially.

In this paper, we extend our work on noise-adaptive speech recognition (NASP) and estimate the environment parameters within the Bayesian framework. Compared to the previous work, which estimates environment parameters by maximum likelihood, the new method is capable of incorporating an appropriate prior knowledge of the environments. The outcome of this extension is a modified environment parameter updating formula, incurring weighting of the estimation by sequential maximum likelihood and that from the prior. Experiments are conducted on a specifically designed database in order to test algorithm performances in nonstationary noise and channel distortion. It is shown that the proposed algorithm provides consistent performance improvement over previous methods for robust speech recognition.

2. Model-based Noisy Speech Recognition

The speech recognition problem can be described as follows. Given a set of trained models $\Lambda_X = \{\lambda_{x_m}\}$ (where λ_{x_m} is the model of m -th speech unit trained from \mathbf{X}) and an observation vector sequence $\mathbf{Y}(T) = (\mathbf{y}(1), \mathbf{y}(2), \dots, \mathbf{y}(T))$, the aim is to recognize the word sequence $\mathbf{W} = (\mathbf{W}(1), \mathbf{W}(2), \dots, \mathbf{W}(L))$ embedded in $\mathbf{Y}(T)$. Each speech unit model λ_{x_m} is a Υ -state CDHMM with state transition probability a_{iq} ($0 \leq a_{iq} \leq 1$) and each state i is modeled by a mixture of Gaussian probability density functions $\{b_{ik}(\cdot)\}$ with parameter $\{w_{ik}, \mu_{ik}, \Sigma_{ik}\}_{k=1,2,\dots,M}$, where M denotes the number of Gaussian mixture components in each state. $\mu_{ik} \in R^{D \times 1}$ and $\Sigma_{ik} \in R^{D \times D}$ are the mean vector and covariance matrix, respectively, of each Gaussian mixture component. D is the dimensionality of feature space. w_{ik} is the mixture weight for state i and mixture k .

In speech recognition, the model Λ_X is used to decode $\mathbf{Y}(T)$ using the maximum a posteriori (MAP) decoder

$$\hat{\mathbf{W}} = \arg \max_{\mathbf{W}} P(\mathbf{Y}(T) | \Lambda_X, \mathbf{W}) P_{\Gamma}(\mathbf{W}) \quad (1)$$

where the first term is the likelihood of observation sequence $\mathbf{Y}(T)$ given that the word sequence is \mathbf{W} , and the second term is denoted as the language model.

2.1. Model-based Noisy Speech Recognition

In the model-based robust speech recognition methods [1], the effect of environment effects on speech feature vectors is represented

in terms of a model. Based on the assumption that the variances of speech, noise, and channel distortion are very small, the following non-linear transformation on the mean vector μ_{ik}^l in mixture k of state i in Λ_X can be used to represent environment effects on log-spectral speech features [1][2],

$$\hat{\mu}_{ik}^l(t) = \mu_{ik}^l + \mu_h^l(t) + \log(1 + \exp(\mu_n^l(t) - \mu_{ik}^l(t) - \mu_h^l(t))) \quad (2)$$

where $\mu_n^l(t) \in R^{J \times 1}$ and $\mu_h^l(t) \in R^{J \times 1}$ are respectively the (time-varying) mean vector for modeling statistics of the noise data $\{\mathbf{n}^l(t) : t = 1, \dots, T\}$ and channel distortion $\{\mathbf{h}^l(t) : t = 1, \dots, T\}$. Superscript l denotes log-spectral domain. We denote the parameters of the environment model, e.g., mean vector and variance of a GMM, of the noise $\{\mathbf{n}^l(t) : t = 1, \dots, T\}$ and channel distortion $\{\mathbf{h}^l(t) : t = 1, \dots, T\}$ by $\mathbf{\Lambda}_N$.

With the estimated $\mathbf{\Lambda}_N$ and certain transformation function (e.g., Eq. (2)), Eq. (1) can be carried out as

$$\hat{\mathbf{W}} = \arg \max_{\mathbf{W}} P(\mathbf{Y}(T) | \Lambda_X, \mathbf{\Lambda}_N, \mathbf{W}) P_{\Gamma}(\mathbf{W}) \quad (3)$$

This function defines the model-based noisy speech recognition approach in our paper. Note that the likelihood is obtained here given speech model Λ_X , word sequences \mathbf{W} , and $\mathbf{\Lambda}_N$. Compared to Eq. (1), this approach has an extra requirement on estimation of $\mathbf{\Lambda}_N$.

2.2. Environment Parameter Estimation

Estimation of $\mathbf{\Lambda}_N$ can be done in general using the following two approaches. The first approach, e.g., [1], assumes the distortion caused by the testing environment to be stationary and estimates HMMs/GMMs representing the testing environment. This requires environment data to train the recognition system. Another approach [2], which is followed in this paper, treats $\mathbf{\Lambda}_N$ as a time-varying model, e.g., with a time varying mean vector, to be estimated sequentially. Our previous work estimates environment parameters by maximum likelihood estimation [2]. In this paper, we extend it to environment parameter estimation within the Bayesian framework.

Denote the estimated environment parameter sequence till frame $t - 1$ as $\mathbf{\Lambda}_N(t - 1) = (\hat{\lambda}_N(1), \hat{\lambda}_N(2), \dots, \hat{\lambda}_N(t - 1))$, where $\hat{\lambda}_N(t - 1)$ is the parameter estimated in the previous frame. If $\lambda_N(t)$, which is assumed to be random vector taking values in R^J , is the parameter vector to be estimated from the sequence $\mathbf{Y}(t) = (\mathbf{y}(1), \mathbf{y}(2), \dots, \mathbf{y}(t))$ till frame t with probability density function (p.d.f.) $P(\mathbf{Y}(t) | \Lambda_X, (\mathbf{\Lambda}_N(t - 1), \lambda_N(t)))$, then the Bayesian estimation, in particular, the maximum a posterior probability (MAP) estimation, $\hat{\lambda}_N^{MAP}(t)$, is defined as the mode of the posterior p.d.f. of $\lambda_N(t)$ denoted as $P(\lambda_N(t) | \mathbf{Y}(t), \Lambda_X)$, i.e.,

$$\begin{aligned} \hat{\lambda}_N^{MAP}(t) &= \arg \max_{\lambda_N(t)} P(\lambda_N(t) | \mathbf{Y}(t), \Lambda_X) \\ &= \arg \max_{\lambda_N(t)} P(\mathbf{Y}(t) | \Lambda_X, (\mathbf{\Lambda}_N(t - 1), \lambda_N(t))) P(\lambda_N(t)) \end{aligned} \quad (4)$$

where the second term in the right side of equation is the prior density of $\lambda_N(t)$.

Note that there is hidden state sequence $S(t)$ in the above likelihood function $P(\mathbf{Y}(t) | \Lambda_X, (\mathbf{\Lambda}_N(t - 1), \lambda_N(t)))$. In fact, some previous works, e.g., [2], provide EM type recursive estimation procedures for maximizing this likelihood function. It follows that in the context of hidden state sequence $S(t)$ in HMM, the same iterative procedure can be used to estimate the mode of the posterior

density by appending a new cost function from prior density to that used for maximizing the likelihood function [3]. In particular, an objective function based on sequential Kullback proximal algorithm [2] is modified to

$$\begin{aligned} \hat{\lambda}_N^{MAP}(t) &= \arg \max_{\lambda_N(t)} Q_t(\hat{\lambda}_N(t - 1); \lambda_N(t)) \\ &\quad - (\beta_t - 1) I(\hat{\lambda}_N(t - 1); \lambda_N(t)) + \log P(\lambda_N(t)) \end{aligned} \quad (5)$$

where the auxiliary function $Q_t(\cdot)$ is defined as

$$\begin{aligned} Q_t(\hat{\lambda}_N(t - 1); \lambda_N(t)) &= \\ &\sum_{S(t)} P(S(t) | \mathbf{Y}(t), \Lambda_X, (\mathbf{\Lambda}_N(t - 1), \hat{\lambda}_N(t - 1))) \\ &\log\{P(\mathbf{Y}(t), S(t) | \Lambda_X, (\mathbf{\Lambda}_N(t - 1), \lambda_N(t)))\} \end{aligned} \quad (6)$$

In Eq. (5), $\beta_t \in R^+$ works as a relaxation factor, and the Kullback-Leibler (K-L) divergence, $I(\hat{\lambda}_N(t - 1); \lambda_N(t))$ is given as,

$$\begin{aligned} I(\hat{\lambda}_N(t - 1); \lambda_N(t)) &= \\ &\sum_{S(t)} P(S(t) | \mathbf{Y}(t), \Lambda_X, (\mathbf{\Lambda}_N(t - 1), \hat{\lambda}_N(t - 1))) \\ &\log \frac{P(S(t) | \mathbf{Y}(t), \Lambda_X, (\mathbf{\Lambda}_N(t - 1), \hat{\lambda}_N(t - 1)))}{P(S(t) | \mathbf{Y}(t), \Lambda_X, (\mathbf{\Lambda}_N(t - 1), \lambda_N(t)))} \end{aligned} \quad (7)$$

Note that the first two terms in Eq. (5) are for maximizing the likelihood. It is known that if $\lambda_N(t)$ does not have informative prior, Eq. (5) is the same as maximum likelihood estimation. If it is assumed that the prior density of $\lambda_N(t)$ is a Gaussian, i.e., $P(\lambda_N(t)) = N(\lambda_N(t); \lambda_N^0, \Sigma_N^0)$, then it can be seen that the parameter estimates are a weighted sum of the prior parameters and that from the observed data sequence. Now, by second order expansion of Eq. (5), parameter updating can be similarly devised as that in [2], i.e.,

$$\begin{aligned} \hat{\lambda}_N^{MAP}(t) &\leftarrow \hat{\lambda}_N(t - 1) - \left[\beta_t \frac{\partial^2 Q_t(\hat{\lambda}_N(t - 1); \lambda_N(t))}{\partial \lambda_N(t)^2} + \right. \\ &\quad \left. (1 - \beta_t) \frac{\partial^2 l_t(\lambda_N(t))}{\partial \lambda_N(t)^2} + \frac{\partial^2 \log P(\lambda_N(t))}{\partial \lambda_N(t)^2} \right]^{-1} \\ &\quad \left(\frac{\partial Q_t(\hat{\lambda}_N(t - 1); \lambda_N(t))}{\partial \lambda_N(t)} + \frac{\partial \log P(\lambda_N(t))}{\partial \lambda_N(t)} \right) \Big|_{\lambda_N(t) = \hat{\lambda}_N(t - 1)} \end{aligned} \quad (8)$$

where the first- and second-order derivations of the auxiliary function, $\frac{\partial Q_t(\hat{\lambda}_N(t - 1); \lambda_N(t))}{\partial \lambda_N(t)}$ and $\frac{\partial^2 Q_t(\hat{\lambda}_N(t - 1); \lambda_N(t))}{\partial \lambda_N(t)^2}$, are respectively given in Eq. (12) and Eq. (13) in [2]. The second-order derivation of the log-likelihood, $\frac{\partial^2 l_t(\lambda_N(t))}{\partial \lambda_N(t)^2}$, is given in Eq. (14) in [2]. The first- and second-order derivative of the log prior density function are respectively given as,

$$\frac{\partial \log P(\lambda_N(t))}{\partial \lambda_N(t)} = -(\Sigma_N^0)^{-1} (\lambda_N(t) - \lambda_N^0) \quad (9)$$

$$\frac{\partial^2 \log P(\lambda_N(t))}{\partial \lambda_N(t)^2} = -(\Sigma_N^0)^{-1} \quad (10)$$

Compared to Eq. (11) in [2], it is seen that, the cost from prior density adds a constant matrix of Eq. (10), and a weighted vector of Eq. (9) into the second- and the first-order derivative of the objective function, respectively. The Σ_N^0 is a hyper-parameter, which controls the relative importance of the prior to the estimate

from maximum likelihood. E.g., when the elements in the matrix Σ_N^0 approaches infinity, the prior does not have contribution to the estimation in Eq. (8). On the contrary, when the elements in Σ_N^0 approaches to 0, the estimation by Eq. (8) is in fact λ_N^0 .

Once the $\lambda_N^{MAP}(t)$ is obtained, it substitutes $\hat{\lambda}_N(t-1)$ for updating at the next frame by Eq. (8).

2.2.1. Derivation of Time-varying Channel Parameter Estimation - A Particular Case

Due to limited space, we only outline environment parameter estimation for channel distortions¹. In this case, the environment model is $\lambda_N(t) = \mu_h^l(t)$. The model of environment effects shown in Eq. (2) relates $\lambda_N(t)$ to the log-likelihood function of observation $\mathbf{y}(t)$ given state i , mixture k , and the model $\lambda_N(t)$ by

$$\begin{aligned} \log b_{ik}(\mathbf{y}(t)) &= -\frac{D}{2} \log(2\pi) \\ &\quad -\frac{1}{2} \log |\Sigma_{ik}| - \frac{1}{2} (\mathbf{y}(t) - \hat{\mu}_{ik}(t))^T \Sigma_{ik}^{-1} (\mathbf{y}(t) - \hat{\mu}_{ik}(t)) \end{aligned} \quad (11)$$

where superscript T denotes transpose operation.

By differentiation of the log-likelihood function w.r.t. the environment parameter, we see the ‘‘contribution’’ of the environment parameter to the change of the log-likelihood, i.e.,

$$\frac{\partial \log b_{ik}(\mathbf{y}(t))}{\partial \lambda_N(t)} = \mathbf{G}_{\hat{\lambda}_N} \frac{\partial \hat{\mu}_{ik}^l(t)}{\partial \lambda_N(t)} \quad (12)$$

where the jj th element in diagonal matrices $\mathbf{G}_{\hat{\lambda}_N} \in R^{J \times J}$ is given as $G_{\hat{\lambda}_N jj} = \sum_{d=1}^D [z_{dj} \frac{(y_t(d) - \hat{\mu}_{ikd}(t-1))}{\Sigma_{ikd}^2}]$. z_{dj} is the DCT coefficient.

The first-order differential term, $\frac{\partial \hat{\mu}_{ik}^l(t)}{\partial \lambda_N(t)}$, in Eq. (12) is obtained by differentiation Eq. (2) w.r.t. $\mu_h^l(t)$, and, for each element $\mu_{nj}^l(t)$ in the environment parameter $\lambda_N^l(t)$, it is given as

$$\frac{\partial \hat{\mu}_{ik}^l(t)}{\partial \mu_{nj}^l(t)} = 1 - \frac{\exp(\mu_{nj}^l(t) - \mu_{ikj}^l - \mu_{hj}^l(t))}{1 + \exp(\mu_{nj}^l(t) - \mu_{ikj}^l - \mu_{hj}^l(t))} \quad (13)$$

Using chain rule, this derivative of log-likelihood w.r.t. the channel distortion parameters contributes to $\frac{\partial Q_t(\hat{\lambda}_N(t-1); \lambda_N(t))}{\partial \lambda_N(t)}$ and $\frac{\partial^2 L_t(\lambda_N(t))}{\partial \lambda_N(t)^2}$ through Eq. (12) and Eq. (14) in [2], respectively.

Suppose that the informative prior on channel is available, the Bayesian updating in Eq. (8) then combines the updating from the above derivatives of log-likelihood w.r.t. to the channel distortion parameter $\mu_h^l(t)$ and that from the prior.² Furthermore, even a coarse choice of the hyper-parameter Σ_N^0 may still benefit the updating, since adding $(\Sigma_N^0)^{-1}$ into the inverse matrix in Eq. (8) reduces the condition number of this inverse matrix, which results in a more stable estimation than that without adding the $(\Sigma_N^0)^{-1}$.

¹Please refer to [4] for detailed updating formulae.

²Remark: In this paper, a model of environment effects in Eq. (2) is assumed for log-spectral speech features. However, the above environment model estimation procedure is not limited to this particular environment effects model. The procedure is general and, given a correct model reflecting environment effects on other speech features, e.g., LDA based features, the procedure can be possibly utilized to these speech features.

3. Experimental Results

A database was designed to evaluate system performances in non-stationary noise and channel distortion. Training set consisted of 8840 clean utterances from Aurora 2 database. Testing set included 1000 utterances in each of four SNR conditions. These testing utterances were generated from testing utterances in Aurora 2 database by convolution with a 50-tap FIR filter (simulating the channel distortion) and corruption by simulated non-stationary noise with white spectral characteristics. The spectral shape of the distortion filter can be seen in Fig. 1, and the spectral evolution of the non-stationary noise can be seen in Fig. 2. The signal-to-noise ratio (SNR) in the degraded speech was measured by $SNR = 10 \log_{10} \frac{\text{energy of filtered speech}}{\text{energy of additive noise}}$. The SNRs were 20.5 dB, 15.6 dB, 11.1 dB, and 6.8 dB.

The speech recognizer was based on whole-word HMM. Each digit was modeled by 18 states, and each state has 3 diagonal Gaussian mixture densities. A filter-bank with twenty-six filters was used in the binning stage. The window size was 25ms and time-shift was 10ms. Features were MFCC plus C0, and their Δ and $\Delta\Delta$ coefficients that, as a whole, had 39 dimension.

The prior environment parameters in Eq. (9) and Eq. (10) were chosen as $\lambda_N^0 = [\mu_n^l(0)^T \mathbf{0}^T]^T$. The variance-covariance matrix $\Sigma_N^0 = \mathbf{I}_{2J \times 2J}$. The relaxation factor β_t in Eq. (8) was set to 0.8. Initialization of the noise parameter $\mu_n^l(t)$ was made by setting it to be the mean vector of silence segments in the testing environments. $\mu_h^l(t)$ was initialized to be a zero vector.

3.1. Estimation of Channel and Noise Parameters

We show the performance of our method for the estimation of channel and noise parameters through an example. In this example, we use all the utterance in the testing set at 11.1 dB SNR. In Fig. 1, the estimated channel responses (log squared-magnitude) are shown at the initialization, the end of first utterance, the end of second utterance, the end of third utterance, and the end of the testing set, and are compared with the response of true channel. Note that the estimation is carried out in each Mel filter-bank bin. Compared to the true channel response, the initialization does not provide shape of the true channel response, which is bent in low- and high frequency. In the figure, ‘‘the 1st utterance’’ means the estimate of channel response at the end of the first utterance, ‘‘the 2nd utterance’’ means the estimate of channel response at the end of the second utterance, and so on. It can be seen from this figure that our noise-adaptive speech recognition approach provides updated channel response estimates that follow the general spectral slope of the channel. As the number of testing utterances increase, the slope of the estimated channel responses is closer to the true channel response. However, in the high frequency end, the estimates are not as sharply bent as the true channel response. This may be attributed to the lack of speech energy at the high frequency end.

In Fig. 2, the estimated noise spectrum is shown in the 7th Mel filter-bank bin as well as the true evolution of the noise spectrum along the time in the bin. The noise spectrum estimate is seen to evolve from poor initialization to the true noise spectrum. Note that the true noise power changes its value along the time with increasing frequency. Below a certain changing rate of the true noise power, the noise adaptive speech recognition provides the noise spectrum estimates that can follow the evolution of the true noise spectrum. Although rapid change of the true noise power spectrum makes the estimation more difficult to follow the trend, the estimated noise parameter is still within the range of noise

spectrum evolution.

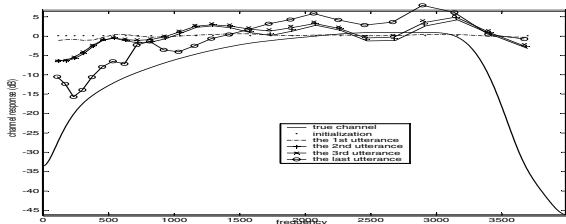


Figure 1: Estimated channel responses at 11.1dB.

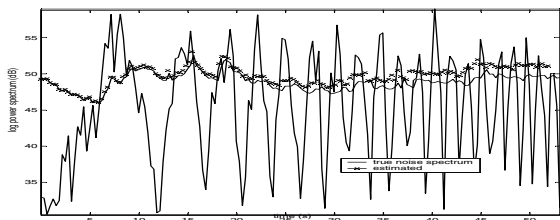


Figure 2: Estimated noise spectra at 7th Mel filter bank at 11.1dB.

3.2. Speech Recognition Results

We report the speech recognition results for the distorted testing utterances under different SNR conditions. The word recognition accuracy on the clean test set was 99.6%. A system with the proposed noise adaptive speech recognition, denoted as NASP, with a prior environment model defined before was compared to the following systems: 1) Baseline: Recognized degraded speech directly, 2) Parallel model combination assuming stationary noise, denoted as SNA: Obtained noise parameter estimation from whole testing utterances and applied Log-Add noise compensation [1] to adapt acoustic models, 3) Cepstral mean normalization, denoted as CMN, and 4) Speech enhancement using Wiener filtering³, denoted as ENH: A Wiener filter is applied for enhancement of testing utterances before speech feature extraction.

The recognition results are summarized in Table 1. The baseline performance dropped rapidly as the SNR decreased from 20.5 dB to 6.7 dB. It is found that the system ENH, which uses Wiener filter to enhance signals, had poor performances in this database. Also, the SNA system did not improve system performances over Baseline at all. Since both ENH and SNA systems estimated noise parameters from silence segments, in this particular task, the estimated noise parameters may not represent accurately the true noise statistics in speech utterances. The CMN system improved recognition accuracies over baseline when SNRs are 15.6 dB or below, but did not get improvement in higher SNR conditions. The NASP system jointly compensated channel distortion and additive time-varying noise, and its performance was consistently among the best in the evaluated systems.

³The Wiener filter was implemented according to proposal [5].

Table 1: Word Accuracy (in %) in the nonstationary environments achieved by the noise adaptive system (denoted as NASP) with $\beta_t = 0.8, \rho = 0.95$ in comparison with Baseline (recognized degraded speech directly), speech enhancement by Wiener filter (denoted as ENH), Log-Add [1] noise compensation assuming stationary noise (denoted as SNA), and the system employing cepstral mean normalization (denoted CMN).

SNR (dB)	6.7	11.1	15.6	20.5
Baseline	73.3	82.2	89.9	95.3
ENH	78.9	81.8	84.1	85.9
SNA	73.6	82.1	89.2	94.7
CMN	76.1	85.4	91.4	94.1
NASP	77.3	90.2	93.7	97.2

It is worth noting the comparison between baseline and CMN. For 15.6 dB SNR and below, compared to the baseline, CMN provided robustness in the sense that it compensated channel distortions to some extent. At 20.5 dB SNR, although the additive noise energy in average was relatively small (in this situation, the environment effects is usually considered as being dominated by channel distortions), the environment could not be considered as stationary due to fluctuation of additive noise powers as shown in Fig. 2. Thus, it is appropriate to consider compensating channel distortions and additive noise jointly and dynamically.

4. Summary

We have proposed a noise adaptive speech recognition approach for recognizing speech corrupted by nonstationary noise and channel distortion. Instead of maximum likelihood estimation of environment parameters (as done in our previous work), the present method estimates environment parameters within the Bayesian framework that is capable of incorporating prior knowledge of the environment. We have conducted on a database that contains digit utterances contaminated by channel distortion and nonstationary noise. Results show that this method performs better than the other methods. We plan to extend this research work by incorporating other schemes to construct informative priors and other models of environment effects.

5. References

- [1] M.J.F.Gales and S.J.Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Computer Speech and Language*, vol. 9, pp. 289–307, 1995.
- [2] K. Yao, K. Paliwal, and S. Nakamura, "Noise adaptive speech recognition in time-varying noise based on sequential Kullback proximal algorithm," in *ICASSP*, 2002, pp. 189–192.
- [3] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.
- [4] K. Yao, K. Paliwal, and S. Nakamura, "A noise adaptive speech recognition approach to robust speech recognition in time-varying environments," Tech. Rep. TR-SLT-0016, ATR SLT, 2002.
- [5] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," Tech. Rep. ETSI ES 202 050, ETSI, 2002.