NOISE ADAPTIVE SPEECH RECOGNITION IN TIME-VARYING NOISE BASED ON SEQUENTIAL KULLBACK PROXIMAL ALGORITHM

Kaisheng Yao⋆*, Kuldip K. Paliwal*† *and Satoshi Nakamura*⋆

⋆ATR Spoken Language Translation Research Laboratories, Kyoto, Japan
†School of Microelectronic Engineering, Griffith University, Australia
kaisheng.yao@slt.atr.co.jp k.paliwal@me.gu.edu.au nakamura@slt.atr.co.jp

ABSTRACT

We present a noise adaptive speech recognition approach, where time-varying noise parameter estimation and Viterbi process are combined together. The Viterbi process provides approximated joint likelihood of active partial paths and observation sequence given the noise parameter sequence estimated till previous frame. The joint likelihood after normalization provides approximation to the posterior probabilities of state sequences for an EM-type recursive process based on sequential Kullback proximal algorithm to estimate the current noise parameter. The combined process can easily be applied to perform continuous speech recognition in presence of non-stationary noise. Experiments were conducted in simulated and real non-stationary noises. Results showed that the noise adaptive system provides significant improvements in word accuracy as compared to the baseline system (without noise compensation) and the normal noise compensation system (which assumes the noise to be stationary).

1. INTRODUCTION

Speech recognition has to be carried out often in situations where there exists environment noise, which causes mismatches between pre-trained models and real testing data. Among many approaches for noisy speech recognition, model-based approach assumes explicit models representing noise effects on speech features. With the assumed models, transformations can be constructed in the model space or feature space to decrease the mismatch.

In the model-based approach, most researches are focused on stationary or slow-varying noise conditions. In this situation, parameters representing environments are often estimated prior to speech recognition, and then, plug into the transformations. However, it is known that the environment statistics may vary during recognition. As a result, the noise parameters estimated prior to speech recognition are no longer relevant to the subsequent inputs.

Recently, we have seen a number of techniques proposed to cope with time-varying noise. They can be categorized into two approaches. In the first approach, time-varying environment sources are modeled by Hidden Markov Models (HMM) or Gaussian mixtures that were 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. In the second approach, environment model parameters are assumed to be time varying and need to be estimated. Works in this approach are either based on (constrained) maximum likelihood estimation [1][2][3] or Bayesian approach [4][5].

In this paper, we apply our recent work on sequential Kullback proximal algorithm [3], which is an extension of the sequential EM algorithm, to the situations where the posterior probability of state sequence given observation sequence is approximated by Viterbi recognition process. With the approximation, the time-varying noise parameter estimation process is combined with the Viterbi process, where the noise parameter estimated in the current frame provides noise parameter for modification of the speech model parameters in the next frame.

2. NOISE ADAPTIVE SPEECH RECOGNITION

2.1. MAP Decision rule for automatic 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 mth subword HMM unit trained from X, and an observation sequence $Y(T) = (y(1), y(2), \cdots, y(T))$, the aim is to recognize the word sequence $W = (W(1), W(2), \cdots, W(L))$ embedded in $Y(T)$. Each speech unit model λ_{x_m} is a N-state CDHMM with state transition probability $a_{iq}(0 \leq a_{iq} \leq 1)$ and each state is modeled by a mixture of Gaussian probability density functions $\{b_{ik}(\cdot)\}$ with parameter $\{w_{ik}, \boldsymbol{\mu}_{\mathbf{ik}}, \boldsymbol{\Sigma}_{\mathbf{ik}}\}_{\mathbf{k=1,2,\cdots,M}}$, where M denotes the number of Gaussian mixture components in each state. $\boldsymbol{\mu}_{\mathbf{ik}}$ and $\boldsymbol{\Sigma}_{\mathbf{ik}}$ are the mean and variance vector of each Gaussian mixture component. w_{ik} is the mixture weight.

In speech recognition, the model Λ_X are used to decode $Y(T)$ using the maximum a posterior (MAP) decoder

$$\begin{aligned}\hat{W} &= \arg\max_W P(W|\Lambda_X, Y(T)) \\ &= \arg\max_W P(Y(T)|\Lambda_X, W)P_\Gamma(W) \end{aligned} \quad (1)$$

where the first term is the likelihood of observation sequence $Y(T)$ given that the word sequence is W, and the second term is denoted as the language model. However, in many situations, there exists mismatches due to environments, e.g., additive noise, and accordingly, there is a mismatch in the likelihood of $Y(T)$ given Λ_X evaluated by (1).

2.2. Time-varying noise parameter estimation

We consider mismatches due to additive noise. The model-based approach assumes explicit models representing environment effects on speech features. A commonly accepted model is

$$Y^l = X^l + \log(1 + \exp(N^l - X^l)) \quad (2)$$

where Y^l, X^l and N^l each denote the noisy observation, speech, and additive noise. Superscript l denotes that the observations are in log-spectral domain.

By explicit using the model in (2), (1) can be carried out as,

$$\hat{W} = \arg\max_{W} P(Y(T)|\Lambda_X, \Lambda_N, W)P_{\Gamma}(W) \tag{3}$$

where Λ_N is the noise model. In case that noise is stationary, Λ_N can be estimated prior to speech recognition.

Recent trend in the area of noisy speech recognition is to treat this noise to be non-stationary and estimate the noise parameters frame-by-frame. It considers that Λ_N (in (3)) is seldom available before speech recognition, or even it is available, the true noise environment may change during the recognition process, so that the Λ_N estimated prior to speech recognition is not related to the true noise parameter during the recognition process. Recent works are either based upon (constrained) maximum likelihood estimation, e.g., [1][2][3], or Bayesian approach [4][5]. In this paper, we consider methods based upon (constrained) maximum likelihood, since our method in this approach [3] shows less computational complexity than our work in [5].

In this approach, the noise parameter is recursively estimated. Denote the estimated noise parameter sequence till frame $t-1$ as $\Lambda_N(t-1) = (\lambda_N(1), \lambda_N(2), \cdots, \lambda_N(t-1))$. Given the current observation sequence $Y(t) = (y(1), y(2), \cdots, y(t))$ till frame t, the noise parameter estimation procedure will find $\lambda_N(t)$ as the current noise parameter estimate, which satisfies,

$$\sum_{S(t)} P(Y(t), S(t)|\Lambda_X, (\Lambda_N(t-1), \lambda_N(t))) \geq \tag{4}$$
$$\sum_{S(t)} P(Y(t), S(t)|\Lambda_X, (\Lambda_N(t-1), \lambda_N(t-1)))$$

where $S(t) = (s(1), s(2), \cdots, s(t))$ is the state sequence till frame t. The formula shows that the updated noise parameter sequence will not decrease the likelihood of observation sequence $Y(t)$, over that given by the previously estimated noise parameter sequence.

Since $S(t)$ is hidden, at each frame, we iteratively maximize the lower bound of the log-likelihood according to Jensen's inequality, i.e.,

$$\log P(Y(t)|\Lambda_X, (\Lambda_N(t-1), \hat{\lambda}_N(t))) \geq Q_t(\lambda_N^\star(t); \hat{\lambda}_N(t))$$
$$= \sum_{S(t)} P(S(t)|Y(t), \Lambda_X, (\Lambda_N(t-1), \lambda_N^\star(t)))$$
$$\log \frac{P(Y(t), S(t)|\Lambda_X, (\Lambda_N(t-1), \hat{\lambda}_N(t)))}{P(S(t)|Y(t), \Lambda_X, (\Lambda_N(t-1), \lambda_N^\star(t)))} \tag{5}$$

where $\lambda_N^\star(t)$ is initialized to be $\lambda_N(t-1)$. At each iteration, the procedure will calculate $P(S(t)|Y(t), \Lambda_X, (\Lambda_N(t-1), \lambda_N^\star(t)))$, and then maximize the lower bound to obtain $\hat{\lambda}_N(t)$. After one iteration, the estimated $\hat{\lambda}_N(t)$ will be set to $\lambda_N^\star(t)$, and a new iteration is carried out. Though, generally, several iterations are required to obtain the final $\hat{\lambda}_N(t)$ as the estimate of $\lambda_N(t)$, it can in fact be approximately estimated by only one iteration. The time recursive procedure is known as sequential EM algorithm.

Forgetting factor $\rho(0 < \rho \leq 1.0)$ can be adopted to improve convergence rate by reducing the effects of past observations relative to the new input, so that (5) is modified to,

$$Q_t(\lambda_N^\star(t); \hat{\lambda}_N(t)) =$$
$$\sum_{\tau=1}^{t} \rho^{t-\tau} \sum_{s(\tau)} P(s(\tau)|Y(\tau), \Lambda_X, (\Lambda_N(\tau-1), \lambda_N^\star(\tau)))$$
$$\log \frac{P(Y(\tau), s(\tau)|\Lambda_X, (\Lambda_N(\tau-1), \hat{\lambda}_N(\tau)))}{P(s(\tau)|Y(\tau), \Lambda_X, (\Lambda_N(\tau-1), \lambda_N^\star(\tau)))} \tag{6}$$

The estimate can be regularized by a Kullback-Leibler divergence between $(\Lambda_N(t-1), \lambda_N(t-1))$ and $(\Lambda_N(t-1), \hat{\lambda}_N(t))$,

$$Q_t(\lambda_N^\star(t); \hat{\lambda}_N(t)) - (\beta_t - 1) \tag{7}$$
$$\sum_{S(t)} \log \frac{P(S(t)|Y(t), \Lambda_X, (\Lambda_N(t-1), \lambda_N(t-1)))}{P(S(t)|Y(t), \Lambda_X, (\Lambda_N(t-1), \hat{\lambda}_N(t)))}$$
$$P(S(t)|Y(t), \Lambda_X, (\Lambda_N(t-1), \lambda_N(t-1)))$$

where $\beta_t \in R^+$ works as a relaxation factor. Recursive procedure by (7) is denoted as sequential Kullback proximal algorithm [3]. The sequential EM algorithm is a special case of this algorithm and corresponds to setting β_t equal to 1.0 in the algorithm. The algorithm can achieve faster parameter estimation than that by sequential EM algorithm.

2.3. Approximation of the posterior probability

Normally, time-varying noise parameter estimation is carried out separately from the recognition process, as that in [1][2], by sequential EM algorithm with summation over all state/mixture sequences of a separately trained speech model. In fact, the joint likelihood of observation sequence $Y(t)$ and state sequence $S(t)$ can be approximately obtained from the Viterbi process, i.e.,

$$P(Y(t), S(t)|\Lambda_X, \Lambda_N(t)) \approx a_{s^\star(t-1)s(t)} b_{s(t)}(y(t))$$
$$P(Y(t-1), S^\star(t-1)|\Lambda_X, \Lambda_N(t-1)) \tag{8}$$

where

$$S^\star(t-1) = \tag{9}$$
$$\arg\max_{S(t-1)} a_{s(t-1)s(t)} P(Y(t-1), S(t-1)|\Lambda_X, \Lambda_N(t-1))$$

By normalizing the joint likelihood w.r.t. sum of those from all active partial state sequences, an approximation of the posterior probability of state sequence can be obtained. Thus in (5) and (7), instead of summing over all state/mixture sequences, the summation is over all *active partial state sequence* (path) till frame t provided by Viterbi process. By Jensen's inequality (5), the summation still provides the lower bound of the log-likelihood. This approximation makes it easy to combine time-varying noise parameter estimation with the Viterbi process. We denote this scheme of time-varying noise parameter estimation as noise adaptive speech recognition.

3. IMPLEMENTATION

Time-varying noise parameter estimation is carried out in the log-spectral domain. The noise model $\lambda_N(t)$ is a single Gaussian with time-varying mean vector $\mu_{\mathbf{n}}^{\mathbf{l}}(\mathbf{t}) \in \mathbf{R^J}$, which needs to be estimated, and constant variance $\mathbf{\Sigma_N^l} \in \mathbf{R^J}$. J is the number of filter banks. At each frame, the pre-trained mean vector $\mu_{\mathbf{ik}}^{\mathbf{l}} \in \mathbf{R^J}$ in each mixture k of state i in speech models is transformed by a non-linear transformation in the log-spectral domain,

$$\mu_{\mathbf{ik}}^{\mathbf{l}}(\mathbf{t}) = \mu_{\mathbf{ik}}^{\mathbf{l}} + \log(\mathbf{1} + \exp(\mu_{\mathbf{n}}^{\mathbf{l}}(\mathbf{t}) - \mu_{\mathbf{ik}}^{\mathbf{l}})) \tag{10}$$

The covariance between components of the above mean vector is assumed to be zero. Cepstral mean vector $\mu_{\mathbf{ik}}(\mathbf{t}) \in \mathbf{R^D}$ is obtained by DCT on the above transformed mean vector $\mu^{\mathbf{l}}_{\mathbf{ik}}(\mathbf{t})$. D is the cepstral vector size.

The time-varying noise parameter $\mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t})$ is estimated by the sequential Kullback proximal algorithm [3]. Let $\lambda_N(t)$ denote the mean vector $\mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t})$, and $\lambda_N(0)$ be the initial parameter. Then given $Y(t)$, the recursive update of $\lambda_N(t)$ is given as,

$$\begin{aligned}\lambda_N(t) &= \lambda_N(t-1) \\ &- \frac{\frac{\partial Q_t(\lambda_N(t-1);\tilde{\lambda}_N)}{\partial \tilde{\lambda}_N}}{\beta_t \frac{\partial^2 Q_t(\lambda_N(t-1);\tilde{\lambda}_N)}{\partial \tilde{\lambda}_N^2} + (1-\beta_t)\frac{\partial^2 l_t(\tilde{\lambda}_N)}{\partial \tilde{\lambda}_N^2}}|_{\tilde{\lambda}_N=\lambda_N(t-1)}\end{aligned} \quad (11)$$

where

$$\frac{\partial Q_t(\lambda_N(t-1);\tilde{\lambda}_N)}{\partial \tilde{\lambda}_N} = \quad (12)$$
$$\sum_{s(t)}\sum_{k(t)} P(s(t)k(t)|Y(t),\Lambda_X,\Lambda_N(t-1))\frac{\partial \log b_{s(t)k(t)}(y(t))}{\partial \tilde{\lambda}_N}$$

$$\frac{\partial^2 Q_t(\lambda_N(t-1);\tilde{\lambda}_N)}{\partial \tilde{\lambda}_N^2} = \rho \cdot \frac{\partial^2 Q_{t-1}(\lambda_N(t-2);\tilde{\lambda}_N)}{\partial \tilde{\lambda}_N^2} +$$
$$\sum_{s(t)}\sum_{k(t)} P(s(t)k(t)|Y(t),\Lambda_X,\Lambda_N(t-1))\frac{\partial^2 \log b_{s(t)k(t)}(y(t))}{\partial \tilde{\lambda}_N^2} \quad (13)$$

$$\begin{aligned}\frac{\partial^2 l_t(\tilde{\lambda}_N)}{\partial \tilde{\lambda}_N^2} &= \sum_{s(t)}\sum_{k(t)} P(s(t)k(t)|Y(t),\Lambda_X,\Lambda_N(t-1)) \\ &[(\frac{\partial \log b_{s(t)k(t)}(y(t))}{\partial \tilde{\lambda}_N})^2 + \frac{\partial^2 \log b_{s(t)k(t)}(y(t))}{\partial \tilde{\lambda}_N^2}] \\ &- (\frac{\partial Q_t(\lambda_N(t-1);\tilde{\lambda}_N)}{\partial \tilde{\lambda}_N})^2\end{aligned} \quad (14)$$

and

$$\begin{aligned}\frac{\partial \log b_{s(t)k(t)}(y(t))}{\partial \tilde{\lambda}_N} &= \mathbf{G}_{\tilde{\lambda}_\mathbf{N}} \frac{\partial \mu^{\mathbf{l}}_{\mathbf{s(t)k(t)}}(\mathbf{t})}{\partial \tilde{\lambda}_\mathbf{N}} \\ \frac{\partial^2 \log b_{s(t)k(t)}(y(t))}{\partial \tilde{\lambda}_N^2} &= \mathbf{H}_{\tilde{\lambda}_\mathbf{N}} (\frac{\partial \mu^{\mathbf{l}}_{\mathbf{s(t)k(t)}}(\mathbf{t})}{\partial \tilde{\lambda}_\mathbf{N}})^\mathbf{2} \\ &+ \mathbf{G}_{\tilde{\lambda}_\mathbf{N}} \frac{\partial^\mathbf{2} \mu^{\mathbf{l}}_{\mathbf{s(t)k(t)}}(\mathbf{t})}{\partial \tilde{\lambda}^\mathbf{2}_\mathbf{N}}\end{aligned}$$

where the jjth element in diagonal matrices $\mathbf{G}_{\tilde{\lambda}_\mathbf{N}}$ and $\mathbf{H}_{\tilde{\lambda}_\mathbf{N}}$ are respectively given as $G_{\tilde{\lambda}_N jj} = \sum_{d=1}^{D}[z_{dj}\frac{(y_t(d)-\mu_{s(t)k(t)d}(t-1))}{\Sigma^2_{s(t)k(t)d}}]$ and $H_{\tilde{\lambda}_N jj} = \sum_{d=1}^{D}[-\frac{1}{\Sigma^2_{s(t)k(t)d}}z^2_{dj}]$. The posterior probability at state and mixture index $(s(t)k(t))$ given observation sequence $Y(t)$ and noise parameter sequence $\Lambda_N(t-1)$ is approximated by Viterbi process as in subsection 2.3. The jth element in $\frac{\partial \mu^{\mathbf{l}}_{\mathbf{s(t)k(t)}}(\mathbf{t})}{\partial \tilde{\lambda}_N}$ and $\frac{\partial^2 \mu^{\mathbf{l}}_{\mathbf{s(t)k(t)}}(\mathbf{t})}{\partial \tilde{\lambda}^2_N}$ are $\frac{\exp(\mu^l_{nj}(t)-\mu^l_{s(t)k(t)j})}{1+\exp(\mu^l_{nj}(t)-\mu^l_{s(t)k(t)j})}$ and $\frac{\exp(\mu^l_{nj}(t)-\mu^l_{s(t)k(t)j})}{(1+\exp(\mu^l_{nj}(t)-\mu^l_{s(t)k(t)j}))^2}$, respectively. z_{dj} is the DCT coefficient.

4. EXPERIMENTAL RESULTS

4.1. Experimental setup

Experiments were performed on TI-Digits database down-sampled to 16kHz. Five hundred clean speech utterances from 15 speakers were used for training and 111 utterances unseen in the training set were used for testing. Digits and silence were respectively modeled by 10-state and 3-state whole word HMMs with 4 diagonal Gaussian mixtures in each state. The window size was 25.0ms with a 10.0ms shift. A filter-bank of Twenty-six filters was used in the binning stage. The features were MFCC + C0.

We compared three systems. The first was the baseline trained on clean speech without noise compensation, denoted as Baseline, and the second was the system with noise compensation by (10) assuming stationary noise, i.e., $\mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t})$ was kept as constant once initialized, denoted as Normal. The third was the noise adaptive recognition system by (11). It is denoted according to the relaxation factor β_t set. Forgetting factor ρ in (6) and (13) was set to 0.995 empirically. Time-varying noise parameter estimation was unsupervised.

Four seconds of contaminating noise was used in each experiment to obtain noise mean vector for Normal. It was also for initialization of $\mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{0})$ in the third system. Baseline performance in clean condition was 97.89% word accuracy (WA).

4.2. Speech recognition in simulated non-stationary noise

White noise signals were multiplied by a Chirp signal, so that the noise power, e.g., in the 12th filter bank, changed continuously as the dash-dotted curve shown in Figure 1. The SNR ranged from 0dB to 20.4dB. We also plotted the estimated noise power versus time in the filter bank by the noise adaptive system.

Observations are as follows. First, the noise adaptive system can track the evolution of the true noise power. Second, the results show that the smaller relaxation factor β_t, the faster the convergence rate in estimation process. For example, estimation by $\beta_t = 0.5$ shows much better tracking performance than that by setting $\beta_t = 1.0$.

In terms of performance, "Baseline" without noise compensation attained 34.34% WA, and "Normal" with stationary noise compensation attained 58.73% WA. All the noise adaptive systems achieved 95.48% WA, higher than that by "Normal" assuming stationary noise.

4.3. Speech recognition in real noise

Speech signals were contaminated by non-stationary Babble noise in different SNRs. Recognition performances are shown in Table 1, together with "Baseline" and "Normal". It is observed that, in all SNR conditions, the noise adaptive system can further improve system performance, compared to that obtained by "Normal", over "Baseline". For example, in 21.5dB, the "Baseline" achieved 34.04% WA, and "Normal" attained 95.18%. The noise adaptive system with $\beta_t = 1.0$ achieved 96.69% WA. As a whole, the adaptive system with β_t set to 0.5, 0.9, and 1.0, achieved, respectively, 26.9%, 30.9%, and 30.9% relative error rate reduction (ERR) over that by "Normal".

We then increased the non-stationarity of the Babble noise by multiplying the noise signal with the Chirp signal as that in subsection 4.2. Results are shown in Table 2. It is observed that the

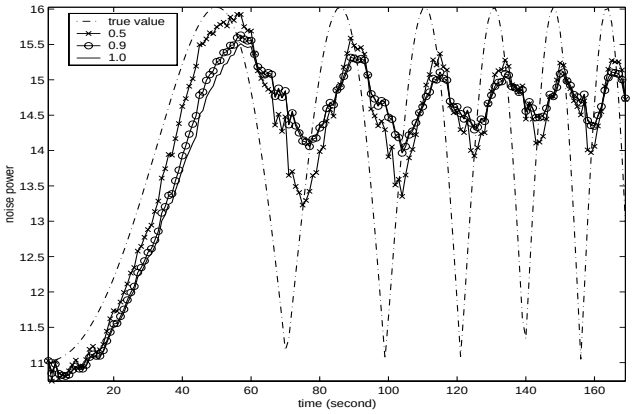

Fig. 1. Estimation of the time-varying parameter $\mu_{\mathbf{n}}^{\mathbf{l}}(\mathbf{t})$ by the noise adaptive systems in the 12th filter bank. Estimates are labeled according to the relaxation factor β_t. The dashed-dotted curve shows evolution of the true noise power in the filter bank.

Table 1. Word Accuracy (in %) in Babble noise, achieved by the noise adaptive system as a function of β_t in comparison with baseline without noise compensation (Baseline), and noise compensation assuming stationary noise (Normal). Relative error rate reduction (ERR) as a function of β_t over Normal are in the last row.

SNR (dB)	Baseline	Normal	0.5	0.9	1.0
29.5	96.69	96.69	97.59	97.89	97.89
21.5	34.04	95.18	96.39	96.69	96.69
13.6	25.30	83.13	90.96	91.27	91.27
7.6	16.27	73.19	75.60	75.30	75.30
ERR (in %)			26.9	30.9	30.9

Relative error rate reduction (ERR) of the noise adaptive system are larger than those in Table 1.

We also tested systems in highly non-stationary Machine-gun noise. Through results shown in Table 3, we observe that the noise adaptive system can improve recognition performance in the noise.

We have the following observations on the results: 1) Our derivation is based on the assumption that the noise is time varying. The assumption fits the real situations. In the non-stationary noises, we observed improvements over noise compensation assuming stationary noise. 2) As shown in Table 1, the highest ERR of the adaptive system over "Normal" was achieved at β_t equal to 1.0 and 0.9, whereas it achieved the highest ERR at $\beta_t = 0.5$, when the non-stationarity of the Babble noise was increased by multiplying it with a Chirp signal. Also, we observed that the highest ERR was achieved at $\beta_t = 0.5$ in Machine-gun noise, which is more non-stationary than Babble noise. It seems that the more non-stationary the noise is, the smaller the β_t to be set[1].

[1]The β_t cannot be too small, since, otherwise, the estimation error after convergence might be large [3].

Table 2. Word Accuracy (in %) in the Chirp-signal-multiplied Babble noise, achieved by the noise adaptive system as a function of β_t in comparison with baseline without noise compensation (Baseline), and noise compensation assuming stationary noise (Normal). Relative error rate reduction (ERR) as a function of β_t over Normal are in the last row.

SNR (dB)	Baseline	Normal	0.5	0.9	1.0
12.4	28.31	64.14	93.07	92.77	92.17
6.9	17.17	50.00	82.83	82.23	81.93
4.4	16.87	48.49	74.10	71.99	71.69
-1.6	14.76	37.65	47.59	50.0	51.51
ERR (in %)			53.0	52.4	52.3

Table 3. Word Accuracy (in %) in Machine-gun noise, achieved by the noise adaptive system as a function of β_t in comparison with baseline without noise compensation (Baseline), and noise compensation assuming stationary noise (Normal). Relative error rate reduction (ERR) as a function of β_t over Normal are in the last row.

SNR (dB)	Baseline	Normal	0.5	0.9	1.0
33.3	91.87	93.37	96.69	95.48	97.59
28.8	87.95	90.60	94.28	95.18	94.28
22.8	78.61	81.33	87.05	83.43	82.83
20.9	77.41	79.82	83.73	85.24	76.51
ERR (in %)			34.8	29.7	23.6

5. DISCUSSION AND SUMMARY

The above results have shown that the noise adaptive speech recognition improves system performances in time-varying noises. Results also show a possible relationship between the best relaxation factor β_t of the recursive noise parameter estimation and the contaminating noise. The noise parameter updating was derived for static MFCCs. It will be more beneficial by making it related to dynamic MFCCs as well. We are still investigating these issues and will report those results in future.

6. REFERENCES

[1] N.S. Kim, "Nonstationary environment compensation based on sequential estimation," *IEEE Signal Processing Letters*, vol. 5, no. 3, March 1998.

[2] Y. Zhao, S. Wang, and K-C. Yen, "Recursive estimation of time-varying environments for robust speech recognition," in *ICASSP*, 2001, pp. 225–228.

[3] K. Yao, K. K. Paliwal, and S. Nakamura, "Sequential noise compensation by a sequential kullback proximal algorithm," in *EUROSPEECH*, 2001, pp. 1139–1142, extended paper submitted for publication.

[4] B. Frey, L. Deng, A. Acero, and T. Kristjansson, "Algonquin: Iterating laplace's method to remove multiple types of acoustic distortion for robust speech recognition," in *EUROSPEECH*, 2001, pp. 901–904.

[5] K. Yao and S. Nakamura, "Sequential noise compensation by sequential monte carlo method," in *Advances in Neural Information Processing Systems 14*. MIT Press, 2002, some parts also to be appeared in IEEE ASRU, 2001.