TIME-VARYING NOISE COMPENSATION BY SEQUENTIAL MONTE CARLO METHOD

Kaisheng Yao and Satoshi Nakamura

ATR Spoken Language Translation Research Laboratories
2-2-2, Hikaridai Seika-cho, Souraku-gun, Kyoto, 619-0288, Japan
E-mail: {kyao, nakamura}@slt.atr.co.jp

ABSTRACT

We present a sequential Monte Carlo method applied to additive noise compensation for robust speech recognition in time-varying noise. At each frame, the method generates a set of samples, approximating the posterior distribution of speech and noise parameter given observation sequences till the current frame. Explicit model representing noise effects on speech features is used, so that an extended Kalman filter is constructed in each sample, generating updated continuous state as the estimation of the noise parameter, and prediction likelihood as the weight of each sample for minimum mean square error inference of the time-varying noise parameter over these samples. A selection step and a smoothing step are used to improve efficiency. Through experiments, we observed significant performance improvements, over that achieved by noise compensation with stationary noise assumption. It also performed better than the sequential EM algorithm in Machinegun noise.

1. INTRODUCTION

Speech recognition in noise has been considered to be essential for its real applications. There have been active research efforts in this area. Among many approaches, model-based approach assumes explicit models representing noise effects on speech features. In this approach, most researches are focused on stationary or slow-varying noise conditions. In this situation, environment noise parameters are often estimated before speech recognition, and then used to compensate noise effects on the subsequent frames of speech.

However, noise statistics may vary during recognition. For example, the contaminating additive noise to a recognizer may change due to the movement of recognizer. As a result, the noise parameters estimated prior to speech recognition of the utterances are no longer relevant to the subsequent frames of input speech.

A number of techniques have been proposed to compensate time-varying noise effects. 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 HMMs/Mixtures, e.g., [1][2]. In the second approach, environment model parameters are assumed to be time-varying, so it is not only an inference problem but also related to environment statistics estimation during speech recognition. The parameters can be estimated by, e.g., sequential EM algorithm [3][4][5]. They can also be estimated by Bayesian approach. In the Bayesian approach, all relevant information on the set of environment parameters and speech parameters is included in the posterior distribution given observation sequence. Except for a few cases including linear Gaussian state space model (Kalman filter), it is formidable to evaluate the distribution updating analytically. Approximation techniques have been applied. For example, Laplace transform is used to approximate the distribution by vector Taylor series [6].

We report an alternative approach for Bayesian estimation and compensation of noise effects on speech features. The method is based on sequential Monte Carlo method [7] for posterior distribution approximation. In the method, a set of samples is generated hierarchically. A state space model representing noise effects on speech features is used explicitly, to construct an extended Kalman filter (EKF) in each sample. The prediction likelihood of the EKF in each sample gives its weight for selection and smoothing of the samples, and inference of the time-varying noise parameter. Noise parameter estimation, noise compensation and speech recognition are carried out frame-by-frame.

2. SPEECH AND NOISE MODEL

The method works on speech features derived from Mel Frequency Cepstral Coefficients (MFCC), and it works in the log-spectral domain. Let t denote frame index.

Speech and noise are respectively modeled by HMMs and a time-varying Gaussian mixture. In case that speech recognition is carried out in stationary additive noise, the following formula [8] has been shown to be effective in compensating noise effects. For Gaussian mixture k_t at state s_t of speech HMM, it transforms mean vector $\mu^{\mathbf{l}}_{\mathbf{s_t k_t}}$ of the Gaussian mixture by,

$$\hat{\mu}^{\mathbf{l}}_{\mathbf{s_t k_t}} = \mu^{\mathbf{l}}_{\mathbf{s_t k_t}} + \log(\mathbf{1} + \exp(\mu^{\mathbf{l}}_{\mathbf{n}} - \mu^{\mathbf{l}}_{\mathbf{s_t k_t}})) \qquad (1)$$

where $\mu^{\mathbf{l}}_{\mathbf{n}}$ is the mean vector in the noise model. $s_t \in \{1, \cdots, S\}$ and $k_t \in \{1, \cdots, M\}$. S and M each denote the number of states in speech models and the number of mixtures at each state. Superscript l indicates that parameters are in log-spectral domain. After transformation, the mean vector $\hat{\mu}^{\mathbf{l}}_{\mathbf{s_t k_t}}$ is transformed by DCT, and then plugged into speech models for recognition of noisy speech.

In case of time-varying noise, the $\mu^{\mathbf{l}}_{\mathbf{n}}$ is a function of time, i.e., $\mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t})$. Accordingly, the compensated mean is $\hat{\mu}^{\mathbf{l}}_{\mathbf{s_t k_t}}(\mathbf{t})$.

We are interested in estimate of $\mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t})$. Denote joint speech and noise parameter as

$$\Theta(t) = (s_t, k_t, \mu^{\mathbf{l}}_{\mathbf{s_t k_t}}(\mathbf{t}), \mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t}), \mathbf{V}(\mathbf{t}), \mathbf{W}(\mathbf{t}))$$

, where $\mathbf{V}(\mathbf{t})$ and $\mathbf{W}(\mathbf{t})$ each denote the state driving noise variance and measurement variance, which will be respectively shown in (2) and (4). Dependences among speech parameter, observation and noise parameter can be viewed in Figure 1.

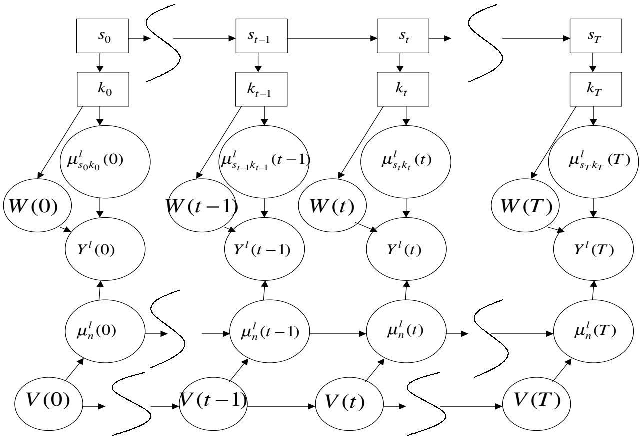

Fig. 1. The graphical model representation of the dependences of the speech and noise parameters. s_t and k_t each denote the state and Gaussian mixture in speech models. $\mu^{\mathbf{l}}_{\mathbf{s_t k_t}}(\mathbf{t})$ and $\mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t})$ each denote the speech and noise parameter. $\mathbf{Y^l}(\mathbf{t})$ is the noisy speech observation vector. $\mathbf{W}(\mathbf{t})$ and $\mathbf{V}(\mathbf{t})$ each denote the measurement and state driving noise variance.

In mixture k_t at state s_t of speech model, speech parameter $\mu^{\mathbf{l}}_{\mathbf{s_t k_t}}(\mathbf{t})$ is assumed to be distributed in Gaussian with mean $\mu^{\mathbf{l}}_{\mathbf{s_t k_t}}$ and variance $\mathbf{W_{s_t k_t}}$. On the other hand, since the environment parameter is assumed to be time varying, the evolution of the environment mean vector can be modeled by a random walk function, i.e.,

$$\mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t}) = \mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t-1}) + \mathbf{v}(\mathbf{t}) \tag{2}$$

where $\mathbf{v}(\mathbf{t})$ is the environment driving noise vector in Gaussian distribution with zero mean and variance $\mathbf{V}(\mathbf{t})$, i.e., $N(\cdot; 0, \mathbf{V}(\mathbf{t}))$. We thus write the prior of speech and noise parameter as

$$p(\Theta(t)|\Theta(t-1)) = a_{s_{t-1}s_t} p_{s_t k_t} N(\mu^{\mathbf{l}}_{\mathbf{s_t k_t}}(\mathbf{t}); \mu^{\mathbf{l}}_{\mathbf{s_t k_t}}, \mathbf{W_{s_t k_t}}) \tag{3}$$
$$N(\mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t}); \mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t-1}), \mathbf{V}(\mathbf{t}))\mathbf{p}(\mathbf{V}(\mathbf{t})|\mathbf{V}(\mathbf{t-1}))\mathbf{p}(\mathbf{W}(\mathbf{t})|\mathbf{s_t k_t})$$

where $a_{s_{t-1}s_t}$ and $p_{s_t k_t}$ are the state transition probability from s_{t-1} to s_t, and mixture weight of k_t at state s_t respectively.

Furthermore, assuming that there are modeling error in (1) and measurement error given observation vector $\mathbf{Y^l}(\mathbf{t})$ in each mixture k_t, we can write the following measurement function,

$$\mathbf{Y^l}(\mathbf{t}) = \mu^{\mathbf{l}}_{\mathbf{s_t k_t}}(\mathbf{t}) + \log\left(\mathbf{1} + \exp\left(\mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t}) - \mu^{\mathbf{l}}_{\mathbf{s_t k_t}}(\mathbf{t})\right)\right) + \mathbf{w}(\mathbf{t}) \tag{4}$$

where $\mathbf{w}(\mathbf{t})$ is Gaussian with zero mean and variance $\mathbf{W}(\mathbf{t})$. Accordingly, the likelihood of observation vector $\mathbf{Y^l}(\mathbf{t})$ at state s_t and mixture k_t is

$$p(\mathbf{Y^l}(\mathbf{t})|\mathbf{\Theta}(\mathbf{t})) = \mathbf{N}(\mathbf{Y^l}(\mathbf{t});$$
$$\mu^{\mathbf{l}}_{\mathbf{s_t k_t}}(\mathbf{t}) + \log\left(\mathbf{1} + \exp\left(\mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t}) - \mu^{\mathbf{l}}_{\mathbf{s_t k_t}}(\mathbf{t})\right)\right), \mathbf{W}(\mathbf{t})) \tag{5}$$

The accumulated joint distribution of speech and noise parameter sequence $\Theta(0:t)$ and observation vector sequence $\mathbf{Y^l}(\mathbf{0}:\mathbf{t})$ till frame t is given as,

$$p(\Theta(0:t), \mathbf{Y^l}(\mathbf{0}:\mathbf{t})) = \mathbf{p}(\mathbf{Y^l}(\mathbf{0})|\mathbf{\Theta}(\mathbf{0}))\mathbf{p}(\mathbf{\Theta}(\mathbf{0}))$$
$$\prod_{\tau=1}^{t} p(\mathbf{Y^l}(\tau)|\mathbf{\Theta}(\tau))\mathbf{p}(\mathbf{\Theta}(\tau)|\mathbf{\Theta}(\tau-\mathbf{1}) \tag{6}$$

The time-varying noise parameter is estimated by minimum mean square error (MMSE) estimation, where the posterior distribution $p(\mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t})|\mathbf{Y^l}(\mathbf{0}:\mathbf{t}))$ is given by marginalization of (6). MMSE estimation is given as

$$\hat{\mu}^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t}) = \int_{\mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t})} \mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t})\mathbf{p}(\mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t})|\mathbf{Y^l}(\mathbf{0}:\mathbf{t}))\mathbf{d}\mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t}) \tag{7}$$

Since $p(\mu^{\mathbf{l}}_{\mathbf{s_t k_t}}(\mathbf{t}), \mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t})|\mathbf{Y^l}(\mathbf{t}))$ is non-Gaussian in $\mu^{\mathbf{l}}_{\mathbf{s_t k_t}}(\mathbf{t})$ and $\mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t})$ due to the non-linearity in (4), it is difficult to assign conjugate prior of $\mu^{\mathbf{l}}_{\mathbf{n}}(\mathbf{t})$ to the likelihood function $p(\mathbf{Y^l}(\mathbf{t})|\mathbf{\Theta}(\mathbf{t}))$, to obtain analytical solution in (6). Another difficulty lies in the fact that there are, in fact, missing data of speech state and mixture sequences in the joint distribution updating in (6). We thus rely on the solution by sequential Monte Carlo method [7].

3. TIME-VARYING NOISE PARAMETER ESTIMATION BY SEQUENTIAL MONTE CARLO METHOD

We apply sequential Monte Carlo method [7] for joint distribution updating. At each frame t, a proposal importance distribution is sampled whose target is the posterior distribution in (6), and it is implemented by sampling from the proposal importance distribution in hierarchy. The method goes through the sampling, selection, and smoothing steps frame-by-frame. MMSE inference of the time-varying noise parameter is a by-product of the steps, carried out after the smoothing step.

In the sampling step, the proposal importance distribution is set as follows,

$$a_{s_{t-1}s_t} p_{s_t k_t} N(\mu^{\mathbf{l}}_{\mathbf{s_t k_t}}(\mathbf{t}); \mu^{\mathbf{l}}_{\mathbf{s_t k_t}}, \mathbf{W_{s_t k_t}})$$
$$p(\mathbf{V}(\mathbf{t})|\mathbf{V}(\mathbf{t-1}))\mathbf{p}(\mathbf{W}(\mathbf{t})|\mathbf{s_t k_t}) \tag{8}$$

It is sampled hierarchically as follows: set $i = 1$ and perform the following steps:

1. sample $s_t^{(i)} \sim a_{s_{t-1}^{(i)} s_t}$
2. sample $k_t^{(i)} \sim p_{s_t^{(i)} k_t}$
3. sample $\mu^{\mathbf{l(i)}}_{\mathbf{s_t^{(i)} k_t^{(i)}}}(\mathbf{t}) \sim \mathbf{N}(; \mu^{\mathbf{l}}_{\mathbf{s_t^{(i)} k_t^{(i)}}}, \mathbf{W}^{\mathbf{l}}_{\mathbf{s_t^{(i)} k_t^{(i)}}})$
4. sample $\log \mathbf{V}^{(\mathbf{i})}(\mathbf{t}) \sim \mathbf{N}(; \log \mathbf{V}, \mathbf{\Sigma_V})$ and set $\mathbf{W}^{(\mathbf{i})}(\mathbf{t}) = \mathbf{W}(\mathbf{t})$. Set $i = i + 1$
5. repeat step 1 to 4 until $i = N$

where superscript (i) denotes the index of samples and N denotes the number of samples. We have assigned $p(\mathbf{W}(\mathbf{t})|\mathbf{s_t k_t}) = \delta(\mathbf{W}(\mathbf{t}) - \mathbf{W_{s_t k_t}})$, where $\delta(0) = 1$, and $P(\log \mathbf{V}(\mathbf{t})|\log \mathbf{V}(\mathbf{t-1})) = \mathbf{N}(\log \mathbf{V}(\mathbf{t}); \log \mathbf{V}, \mathbf{\Sigma_V})$, where $\mathbf{V}$ and $\mathbf{\Sigma_V}$ are set in experiments. Each sample represents certain speech and noise parameter, denoted as

$$\Theta^{(i)}(t) = (s_t^{(i)}, k_t^{(i)}, \mu^{\mathbf{l(i)}}_{\mathbf{s_t^{(i)} k_t^{(i)}}}(\mathbf{t}), \mu^{\mathbf{l(i)}}_{\mathbf{n}}(\mathbf{t}), \mathbf{V}^{(\mathbf{i})}(\mathbf{t}), \mathbf{W}^{(\mathbf{i})}(\mathbf{t}))$$

Each sample has its weight as the remaining part in (6), given as,

$$\beta^{(i)}(t) = p(\mathbf{Y^l(t)}|\mathbf{\Theta^{(i)}(t)}) \\ N(\mu_{\mathbf{n}}^{\mathbf{l(i)}}(\mathbf{t}); \mu_{\mathbf{n}}^{\mathbf{l(i)}}(\mathbf{t-1}), \mathbf{V^{(i)}(t)})\breve{\beta}^{(\mathbf{i})}(\mathbf{t-1}) \quad (9)$$

where $\breve{\beta}^{(i)}(t-1)$ is the weight of sample i at previous frame.

Given $\breve{\beta}^{(i)}(t-1)$, the remaining of (9) can be calculated as the prediction likelihood of the state space model given by (2) and (4) for each sample (i). This likelihood can be obtained analytically since after linearization of (4) with respect to $\mu_{\mathbf{n}}^{\mathbf{l}}(\mathbf{t})$ at $\mu_{\mathbf{n}}^{\mathbf{l(i)}}(\mathbf{t-1})$, an extended Kalman filter can be obtained, where the prediction likelihood of the EKF gives the weight, and the updated continuous state of EKF gives $\mu_{\mathbf{n}}^{\mathbf{l(i)}}(\mathbf{t})$.

In practice, after the above sampling step, the weights of all but several samples may become insignificant. Given the fixed number of samples, this will result in degeneracy of the estimation. A selection step by residual resampling [7] is adopted after the sampling step. The method avoids the degeneracy by discarding those samples with insignificant weights, and in order to keep the number of the samples constant, samples with significant weights are duplicated. Accordingly, the weights after the selection step are also proportionally redistributed. Denote the set of samples after the selection step as $\tilde{\Theta}(t) = \{\tilde{\Theta}^{(i)}(t); i = 1 \cdots N\}$ with weights $\tilde{\beta}(t) = \{\tilde{\beta}^{(i)}(t); i = 1 \cdots N\}$.

After the selection step at frame t, these N samples are distributed approximately according to (6). However, the discrete nature of the approximation can lead to a skewed importance weights distribution, where the extreme case is all the samples have the same $\tilde{\Theta}(t)$ estimated. A Metropolis-Hastings smoothing [9] step is introduced in each sample where the step involves sampling a candidate $\Theta^{\star(i)}(t)$ given the current $\tilde{\Theta}^{(i)}(t)$ according to the proposal importance distribution. To simplify calculation, we assume that the importance distribution is symmetric, and after some mathematical manipulation, it is shown that an acceptance possibility is given by $\min\{1, \frac{\beta^{\star(i)}(t)}{\tilde{\beta}^{(i)}(t)}\}$. The Markov chain then moves towards $\Theta^{\star(i)}(t)$ with the acceptance possibility, otherwise it remains at $\tilde{\Theta}^{(i)}$. Denote the obtained samples as $\breve{\Theta}(t) = \{\breve{\Theta}^{(i)}(t); i = 1 \cdots N\}$ with weights $\breve{\beta}(t) = \{\breve{\beta}^{(i)}(t); i = 1 \cdots N\}$.

Noise parameter $\mu_{\mathbf{n}}^{\mathbf{l}}(\mathbf{t})$ is estimated via MMSE over the samples, i.e.,

$$\hat{\mu}_{\mathbf{n}}^{\mathbf{l}}(\mathbf{t}) = \sum_{\mathbf{i=1}}^{\mathbf{N}} \frac{\breve{\beta}^{(\mathbf{i})}(\mathbf{t})}{\sum_{\mathbf{j=1}}^{\mathbf{N}} \breve{\beta}^{(\mathbf{j})}(\mathbf{t})} \breve{\mu}_{\mathbf{n}}^{\mathbf{l(i)}}(\mathbf{t})$$

where $\breve{\mu}_{\mathbf{n}}^{\mathbf{l(i)}}(\mathbf{t})$ is the updated continuous state of the EKF in the sample after the smoothing step. Once the estimate $\hat{\mu}_{\mathbf{n}}^{\mathbf{l}}(\mathbf{t})$ has been obtained, it is plugged into (1) to do non-linear transformation of clean speech models.

4. EXPERIMENTAL RESULTS

4.1. Experimental setup

Experiments were performed on the TI-Digits database down-sampled to 16kHz. Five hundred clean speech utterances from 15 speakers and 111 utterances unseen in the training set were used for training and testing, respectively. 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. Twenty-six filter banks were used in the binning stage. The features were MFCC + Δ MFCC. The baseline system had 98.7% Word Accuracy under clean conditions.

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 (1) assuming stationary noise, denoted as Normal. The third was the proposed method, denoted according to the number of samples and variance of the environment driving noise $\mathbf{V(t)}$. In the experiments, $\mathbf{V(t)} = \mathbf{V}$ and $\mathbf{\Sigma_V} = \mathbf{0}$. Four seconds of contaminating noise was used in each experiment to obtain noise mean vector $\mu_{\mathbf{n}}^{\mathbf{l}}$ for Normal. It was also for initialization of $\mu_{\mathbf{n}}^{\mathbf{l}}(\mathbf{0})$ in the third system. The initial $\mu_{\mathbf{n}}^{\mathbf{l(i)}}(\mathbf{0})$ for each sample was sampled from $N(; \mu_{\mathbf{n}}^{\mathbf{l}}(\mathbf{0}), \mathbf{0.01}) + \mathbf{N}(; \mu_{\mathbf{n}}^{\mathbf{l}}(\mathbf{0}) + \zeta(\mathbf{0}), \mathbf{10.0})$, where $\zeta(0)$ was flat distribution in $[-1.0, 9.0]$.

4.2. Speech recognition in controlled environments

The noise power of the contaminating White noise was controlled so that, in the first experiment denoted as A, it changed continuously by a chirp signal. In the second experiment, denoted as B, the noise power was changed by a rectangular signal generator in the same changing frequency as chirp signal in experiment A. The signal-to-noise ratio (SNR) ranged from 0dB to 20.4dB. We plotted the noise power at 12th filter bank versus frames in Figure 2, together with the estimated noise power by the sequential method with number of samples N set to 120 and environment driving noise variance $\mathbf{V}$ set to 0.0001.

Observation from Figure 2 is that the method can track the evolution of the noise power. In terms of recognition performance, Table 1 shows that the method can effectively improve system robustness to the time-varying noise. For example, with 60 samples, and $\mathbf{V}$ set to 0.001, the method can improve word accuracy from 75.30% achieved by "Normal", to 94.28% in experiment A. The table also shows that, the word accuracies can be improved by increasing number of samples. For example, given environment driving noise variance $\mathbf{V}$ set to 0.0001, increasing number of samples from 60 to 120, can improve word accuracy from 77.11% to 85.84% in experiment B.

Table 1. Word Accuracy (in %) in controlled experiments, achieved by the sequential Monte Carlo method in comparison with baseline without noise compensation, denoted as Baseline, and noise compensation assuming stationary noise, denoted as Normal.

	Baseline	Normal	N = 60		N = 120	
			V		**V**	
			0.001	0.0001	0.001	0.0001
A	48.19	75.30	94.28	93.98	94.28	94.58
B	53.01	78.01	82.23	77.11	85.84	85.84

4.3. Speech recognition in real noise

In this experiment, speech signals were contaminated by highly non-stationary Machinegun noise in different SNRs. The number

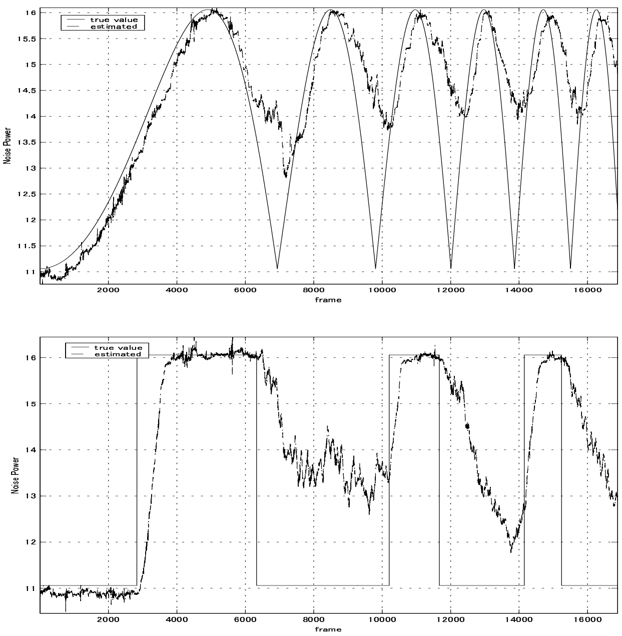

Fig. 2. Estimation of the time-varying parameter $\mu^{1}_{\mathbf{n}}(\mathbf{t})$ by the sequential Monte Carlo method at 12th filter bank in experiment A (upper) and B (lower). Number of samples is 120. Environment driving noise variance is 0.0001. Solid curve is the true noise power. Dash-dotted curve is the estimated noise power.

of samples was set to 120, and the environment driving noise variance $\mathbf{V}$ was set to 0.0001. Recognition performances are shown in Table 2, together with "Baseline" and "Normal". It is observed that, in all SNR conditions, the method can further improve system performance, compared to that obtained by "Normal", over "Baseline". For example, in 8.86dB SNR, the method can improve word accuracy from 75.60% by "Normal" to 83.13%. As a whole, the method can have 39.9% relative word error rate reduction compared to "Normal".

Table 2. Word Accuracy (in %) in Machinegun noise, achieved by the sequential Monte Carlo method in comparison with baseline without noise compensation, denoted as Baseline, and noise compensation assuming stationary noise, denoted as Normal.

SNR (dB)	Baseline	Normal	N = 120, V = 0.0001
28.86	90.36	92.77	97.59
14.88	64.46	76.81	88.25
8.86	56.02	75.60	83.13
1.63	50.0	68.98	72.89

Since the sequential Monte Carlo method approximates the posterior distribution of speech and noise parameters given observation sequences, it can, in principle, de-noise speech in the feature space by MMSE inference of clean speech from noisy observation sequences. Unfortunately, in our experiments carried out so far, the results by de-noising speech were not satisfactory, even a secondary training stage of the denoised speech, as proposed in [10], had been tried.

We also conducted experiments to compare the method with sequential noise compensation by sequential Kullback proximal algorithm [5], which is a generalization of the sequential EM algorithm for time-varying parameter estimation. In the experiments, we found that the sequential Monte Carlo method with the above setting performed better than the sequential Kullback proximal algorithm in Machinegun noise. However, in time-varying Factory noise, which is comparatively less non-stationary than the Machinegun noise, its performance was lower than the sequential Kullback proximal algorithm. It is possible to achieve the same performance as the sequential Kullback proximal algorithm by increasing sampling number N, though the computation load will be very expensive.

5. SUMMARY

We have presented a sequential Monte Carlo method for Bayesian estimation of time-varying noise parameter. The method uses samples to approximate the posterior distribution of the additive noise and speech parameters given observation sequences. A model representing noise effects on speech features has been utilized, and a state space with continuous state representing noise parameters has been constructed. Extended Kalman filter of the state space model provides the prediction likelihood to update sample weights. Estimate of the time-varying noise parameters is carried out by MMSE over the samples, given the weights. Experiments conducted on digits recognition in controlled experiments and Machinegun noise have shown that the method is very effective to improve system robustness to highly time-varying additive noise.

6. REFERENCES

[1] A. Varga and R.K. Moore, "Hidden markov model decomposition of speech and noise," in *ICASSP*, 1990, pp. 845–848.

[2] T. Takiguchi, S. Nakamura, and K. Shikano, "Speech recognition for a distant moving speaker based on hmm composition and seperation," in *ICASSP*, 2000, pp. 1403–1406.

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

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

[5] K. Yao, K. K. Paliwal, and S. Nakamura, "Sequential noise compensation by a sequential kullback proximal algorithm," in *EUROSPEECH*, 2001.

[6] 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.

[7] J. S. Liu and R. Chen, "Sequential monte carlo methods for dynamic systems," *J. Am. Stat. Assoc*, vol. 93, pp. 1032–1044, 1998.

[8] K. Yao, B. E. Shi, S. Nakamura, and Z. Cao, "Residual noise compensation by a sequential em algorithm for robust speech recognition in nonstationary noise," in *ICSLP*, 2000, vol. 1, pp. 770–773.

[9] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, pp. 97–109, 1970.

[10] L. Deng, A. Acero, M. Plumpe, and X. D. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *ICSLP*, 2000, pp. 806–809.