



Noise Robust Voice Activity Detection Based on Switching Kalman Filter

Masakiyo Fujimoto and Kentaro Ishizuka

NTT Communication Science Laboratories, NTT Corporation
 2-4, Hikari-dai, Seika-cho, Souraku-gun, Kyoto, 619-0288, Japan
 E-mail: {masakiyo, ishizuka}@cslab.kecl.ntt.co.jp

Abstract

This paper addresses the problem of voice activity detection (VAD) in noisy environments. The VAD method proposed in this paper is based on a statistical model approach, and estimates statistical models sequentially without *a priori* knowledge of noise. Namely, the proposed method constructs a clean speech / silence state transition model beforehand, and sequentially adapts the model to the noisy environment by using a switching Kalman filter when a signal is observed. The evaluation is carried out by using a VAD evaluation framework, CENSREC-1-C. The evaluation results revealed that the proposed method significantly outperforms the baseline results of CENSREC-1-C as regards VAD accuracy in real environments.
Index Terms: voice activity detection, statistical model, switching Kalman filter, real environment, CENSREC-1-C

1. Introduction

Voice activity detection (VAD) that automatically detects a period of target human speech from a continuously observed signal is one of the most important techniques for speech signal processing. VAD is widely used in various speech signal processing techniques, e.g., speech enhancement, speech coding for cellular or IP phones, and the front-end processing of automatic speech recognition.

Usually, VAD consists of two parts: a feature extraction part and a decision part. The feature extraction part extracts acoustic features for speech / non-speech discrimination, and the traditional features are the zero-crossing rate and the energy difference between speech and non-speech [1]. However, these parameters are not robust in the presence of interference noises, thus several noise robust features have been proposed [2, 3]. These parameters can improve the VAD accuracy. However, the improvement range decreases with degradation in the signal to noise ratio (SNR). When the SNR is low, the discriminative characteristics of the feature parameter unavoidably degrade due to the strong noise energy, even if a noise robust feature parameter is used. Consequently, differences between speech and non-speech become ambiguous, and it becomes difficult to achieve sufficient VAD accuracy with a low SNR. This problem indicates the difficulty of achieving robust VAD by feature extraction alone and the importance of a decision mechanism. If a robust decision mechanism is introduced into VAD, the VAD accuracy will improve, even if the discriminative characteristics of the feature parameter are ambiguous. In this paper, we focus on a decision mechanism for noise robust VAD.

A statistical model-based VAD technique has been proposed as a robust decision mechanism by Sohn *et al.* [4]. This method defines a speech / non-speech state transition model, and calculates the likelihood ratio of a speech state to a non-speech state by using forward probability estimation. Sohn's method provides robust performance in noisy environments;

however, this performance is restricted to specific environments. Namely, assumptions of stationary noise environments and *a priori* knowledge of noise are indispensable to Sohn's method. In the most cases, a noise observed in real world has non-stationary characteristics and knowledge of it is not provided in advance. Thus, the robustness in the presence of non-stationary noise and the lack of a need for *a priori* knowledge of noise are the most important factors for actual robust and useful VAD in the real world.

For VAD with such assumptions, we propose a VAD technique based on a switching Kalman filter. The proposed method constructs a clean speech / silence state transition model in advance, and the noise model is sequentially updated by using a Kalman filter when a signal is observed. After the noise model is updated, a noise adapted model, i.e., a model with a speech (clean speech + noise) state and a non-speech (silence + noise) state is composed by using probability density functions (PDFs) of a clean speech state or a silence state and updated noise model. In the method, the Kalman filter for noise updating is formulated by using PDF parameters of the clean speech state or the silence state. Consequently, two types of estimation (updating) results are given for the noise model by the selection of the clean speech state or the silence state. This means that the state-space representation of the Kalman filter depends on state selection, thus, the proposed method has the characteristic of the switching Kalman filter. After the model adaptation (composition) with switching Kalman filter, the likelihood ratio between a speech state and a non-speech state is calculated.

The proposed method was evaluated on the CENSREC-1-C database (Corpora and Environments for Noisy Speech RECOgnition-1 Concatenated) [5], which is concatenated Japanese noisy speech data for VAD evaluation. The evaluation results revealed that the proposed method significantly improves VAD accuracy compared with the CENSREC-1-C baseline. In addition, we confirmed that the proposed VAD improves the speech recognition accuracy of concatenated utterances.

2. Statistical model-based VAD

In this section, we briefly review the concept of the statistical VAD proposed by Sohn *et al.* [4]. Statistical VAD discriminates between speech and non-speech periods based on the likelihood ratio test (LRT) with a statistical model. The statistical model is constructed by using an ergodic state transition model with speech and non-speech states as shown in Figure 1.

In the figure, the symbols H_0 and H_1 denote the non-speech and speech states, respectively. $a_{i,j}$, $b_j(\mathbf{O}_t)$, and \mathbf{O}_t denote the state transition probability from state i to state j , the output probability at state j , and the L -dimensional vector of the observed signal at the t -th short time frame, respectively.

By using the state transition model, the discrimination of speech or non-speech periods is equivalent to the estimation of

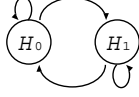


Figure 1: Speech / Non-speech state transition model (H_0 : Non-speech state, H_1 : Speech state)

the t -th frame state q_t when $\mathbf{O}_{0:t} = \{\mathbf{O}_0, \dots, \mathbf{O}_t\}$ is given. Thus, the observed signal assigned to speech state ($q_t = H_1$) is extracted as a speech signal. The state q_t is decided with respect to the conditional probability $p(q_t | \mathbf{O}_{0:t})$ as follows:

$$p(q_t | \mathbf{O}_{0:t}) = p(\mathbf{O}_{0:t}, q_t) / p(\mathbf{O}_{0:t}) \propto p(\mathbf{O}_{0:t}, q_t) \quad (1)$$

$$p(\mathbf{O}_{0:t}, q_t) = \sum_{q_{t-1}} p(q_t | q_{t-1}) p(\mathbf{O}_t | q_t) p(\mathbf{O}_{0:t-1}, q_{t-1}) \quad (2)$$

The joint probability $p(\mathbf{O}_{0:t}, q_t)$ can be represented by the recursive formula of Eq. (2) based on the first order Markov chain, and is usually called the forward probability $\alpha_{j,t}$. Thus, Eq. (2) is represented as the following equation:

$$\alpha_{j,t} = a_{0,j} b_j(\mathbf{O}_t) \alpha_{0,t-1} + a_{1,j} b_j(\mathbf{O}_t) \alpha_{1,t-1} \quad (3)$$

where $a_{i,j} = p(q_t = H_j | q_{t-1} = H_i)$ and $b_j(\mathbf{O}_t) = p(\mathbf{O}_t | q_t = H_j)$.

Finally, the state q_t is given by the LRT, namely, the thresholding likelihood ratio $R_t = \alpha_{1,t} / \alpha_{0,t}$ as

$$q_t = \begin{cases} H_0 & R_t < \text{Threshold} \\ H_1 & R_t \geq \text{Threshold} \end{cases} \quad (4)$$

In Eq. (3), the calculation of $b_j(\mathbf{O}_t)$ is a crucial factor as regards accurate VAD. In the original statistical VAD method proposed by Sohn *et al.*, output probability $b_j(\mathbf{O}_t)$ is given by using *a priori* and *a posteriori* SNRs [6].

3. VAD based on switching Kalman filter

3.1. Definition of state transition model

In the proposed method, we calculated $b_j(\mathbf{O}_t)$ using PDFs, because an LRT with PDFs is more flexible and applicable than the conventional *a priori* and *a posteriori* SNR-based approach. As the PDFs for the LRT, we chose a Gaussian mixture model (GMM) modeled in the log-Mel spectral domain as follows:

$$b_j(\mathbf{O}_t) = \sum_{k=1}^K w_{j,k} \prod_{l=0}^{L-1} \mathcal{N}(O_{t,l}; \mu_{O_{j,k,l}}, \sigma_{O_{j,k,l}}) \quad (5)$$

where $w_{j,k}$, $O_{t,l}$, $\mu_{O_{j,k,l}}$, and $\sigma_{O_{j,k,l}}^2$ denote the mixture weight of the k -th Gaussian distribution, the l -th element of \mathbf{O}_t , the mean of $O_{t,l}$, and the (diagonal) variance of $O_{t,l}$, respectively. With this approach, if a noise (non-speech state) GMM and a noisy speech (speech state) GMM are given in advance, we can easily calculate $b_j(\mathbf{O}_t)$. However, it is difficult and unrealistic to use these models, because they need *a priori* knowledge of noise. To cope with unknown noisy environments, it is necessary to construct environmentally matched model sets by using an on-line estimation. To deal with this problem, we first defined non-speech and speech periods as follows:

$$\begin{aligned} q_t = H_0 &: \text{Non-speech period} : \text{Silence} + \text{Noise} \\ q_t = H_1 &: \text{Speech period} : \text{Speech} + \text{Noise} \end{aligned}$$

With this definition, we modeled the speech state transition model by using an ergodic state transition model shown as Figure 1. The PDF of each state is given by GMM. This



Figure 2: State transition model of noise

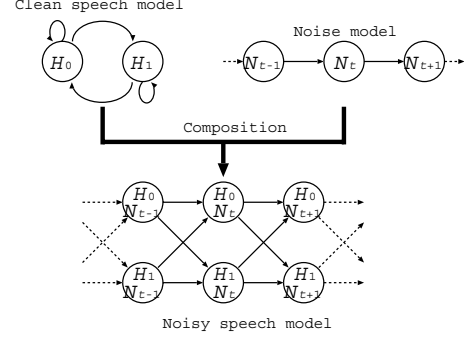


Figure 3: Speech / non-speech state transition model with noise dynamics

model is same as the model used in Sohn's method, however, the model used in the proposed method has clean speech and silence states. Next, we assume that noise has non-stationary characteristics, thus, the noise sequence is modeled by using a sequential state transition model as shown in Figure 2. Finally, by composing speech and noise models, we can construct the speech / non-speech state transition model with noise dynamics as shown in Figure 3. Namely, this model has state transition processes for both speech and noise. Speech has a discrete state transition process and noise has a sequential one.

With this approach, the silence and clean speech GMMs can be modeled in advance by using a clean speech corpus. On the other hand, the noise statistics are unknown. Thus, we estimate the noise statistics sequentially by using a Kalman filter.

3.2. Formulation of likelihood calculation

When $\mathbf{O}_{0:t}$ and noise sequence $\mathbf{N}_{0:t} = \{\mathbf{N}_0, \dots, \mathbf{N}_t\}$ are given, the state q_t is decided with respect to the conditional probability $p(q_t | \mathbf{O}_{0:t}, \mathbf{N}_{0:t})$ as follows:

$$\begin{aligned} p(q_t | \mathbf{O}_{0:t}, \mathbf{N}_{0:t}) &= p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t}) / p(\mathbf{O}_{0:t}, \mathbf{N}_{0:t}) \\ &\propto p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t}) \end{aligned} \quad (6)$$

The recursive formula of the joint probability $p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t})$ is given by

$$\begin{aligned} p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t}) &= \sum_{q_{t-1}} p(q_t, \mathbf{N}_t | q_{t-1}, \mathbf{N}_{t-1}) p(\mathbf{O}_t | q_t, \mathbf{N}_t) \\ &\quad \times p(\mathbf{O}_{0:t-1}, q_{t-1}, \mathbf{N}_{0:t-1}) \end{aligned} \quad (7)$$

Here, we assume that the state transition processes of q_t and \mathbf{N}_t are mutually independent, thus, the joint probability is given by

$$\begin{aligned} p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t}) &= \sum_{q_{t-1}} p(q_t | q_{t-1}) p(\mathbf{N}_t | \mathbf{N}_{t-1}) p(\mathbf{O}_t | q_t, \mathbf{N}_t) \\ &\quad \times p(\mathbf{O}_{0:t-1}, q_{t-1}, \mathbf{N}_{0:t-1}) \end{aligned} \quad (8)$$

By defining $p(\mathbf{N}_t | \mathbf{N}_{t-1}) = c_{t,t-1}$ and $p(\mathbf{O}_t | q_t = H_j, \mathbf{N}_t) = b_{j, \mathbf{N}_t}(\mathbf{O}_t)$, the forward probability $\alpha_{j,t} = p(\mathbf{O}_{0:t}, q_t = H_j, \mathbf{N}_{0:t})$ is represented as the following equation from Eq. (8).

$$\alpha_{j,t} = \sum_{i=0}^1 (a_{i,j} \alpha_{i,t-1}) b_{j, \mathbf{N}_t}(\mathbf{O}_t) c_{t,t-1} \quad (9)$$

In (8), $c_{t,t-1}$ is always set at 1, because we assume that the noise has a sequential state transition process. Thus, Eq. (9) is simplified as

$$\alpha_{j,t} = \sum_{i=0}^1 (a_{i,j} \alpha_{i,t-1}) b_{j, \mathbf{N}_t}(\mathbf{O}_t). \quad (10)$$

On the other hand, when we focus on the state transition model of noise shown in Figure 2, it is also given by the following equation. This equation is completely equivalent to statistical representation of a Kalman filter [7]D

$$\begin{aligned} p(\mathbf{O}_{0:t}, \mathbf{N}_{0:t}) \\ = p(\mathbf{N}_t | \mathbf{N}_{t-1}) p(\mathbf{O}_t | \mathbf{N}_t) p(\mathbf{O}_{0:t-1}, \mathbf{N}_{0:t-1}) \end{aligned} \quad (11)$$

If the probability (state) variable q_t is added to Eq. (11), the statistical process is equivalent to Eq. (8). This means that Eq. (8) is equivalent to a statistical representation of a switching Kalman filter that switches the state-space model of a Kalman filter based on a state variable.

3.3. Noise model updating based on switching Kalman filter
Sequential noise updating is carried out by Kalman filtering. The Kalman filter requires a definition of the signal model called a dynamical system (state-space model). Typically, a dynamical system can be defined by two equations: a state transition equation that represents the dynamics of the target signal, and an observation equation that represents the output system of the observed signal.

For the state transition process, a random walk process is applied to the state transition of $N_{t,l}$ as follows:

$$N_{t+1,l} = N_{t,l} + W_{t,l} \quad (12)$$

$$W_{t,l} \sim \mathcal{N}(0, \sigma_{W_l}^2) \quad (13)$$

where $W_{t,l}$ and $\sigma_{W_l}^2$ denote the driving noise for the state transition process and the variance of $W_{t,l}$, respectively.

On the other hand, the observation process is modeled by the following non-linear equation [8],

$$\begin{aligned} O_{t,l} &= S_{t,l} + \log(1 + \exp(N_{t,l} - S_{t,l})) \\ &= f(S_{t,l}, N_{t,l}) \end{aligned} \quad (14)$$

where $S_{t,l}$ denotes log-Mel spectra of silence or clean speech.

In Eq. (14), the parameter $S_{t,l}$ is usually unknown. Thus, the parameters of silence or clean speech GMMs are substituted for the parameter $S_{t,l}$ as follows:

$$O_{t,l} = f(\mu_{S_{j,k,l}}, N_{t,l}) + V_{t,j,k,l} \quad (15)$$

$$V_{t,j,k,l} \sim \mathcal{N}(0, \sigma_{S_{j,k,l}}^2) \quad (16)$$

where $\mu_{S_{j,k,l}}$ and $\sigma_{S_{j,k,l}}^2$ denote the mean and variance of silence ($j = 0$) and speech ($j = 1$) GMMs, respectively. $V_{t,j,k,l}$ denotes an error signal between $S_{t,l}$ and $\mu_{S_{j,k,l}}$.

Since a GMM consists of K Gaussian distributions, K types of observation processes are derived from Eq. (15). Using these observation processes, the (non-linear) Kalman filter is multiplied into K types and we can obtain K types of estimation results for each GMM. In addition, Eq. (15) switches the characteristic of the state-space representation according to the type of GMM (silence or clean speech), thus, the Kalman filter given by the state-space model of Eqs. (12) and (15) has the characteristic of a switching Kalman filter. The estimation formulas for each Kalman filter are described in [8, 9].

After the noise updating, the mean and variance of the observation are given by the following equations.

$$\mu_{O_{t,j,k,l}} = f(\mu_{S_{j,k,l}}, N_{t,j,k,l}) \quad (17)$$

$$\sigma_{O_{t,j,k,l}}^2 = F_{t,j,k,l} \sigma_{N_{t,j,k,l}}^2 F_{t,j,k,l} + \sigma_{S_{j,k,l}}^2 \quad (18)$$

$$F_{t,j,k,l} = \partial \mu_{O_{t,j,k,l}} / \partial N_{t,j,k,l} \quad (19)$$

where $N_{t,j,k,l}$ and $\sigma_{N_{t,j,k,l}}^2$ denote the noise mean and variance, respectively which are estimated by using the parameters of the k -th parameter included in GMM j . $\mu_{O_{t,j,k,l}}$ and $\sigma_{O_{t,j,k,l}}^2$ denote the composed mean and variance of the observation in the t -th frame. The output probability of each state $b_{j, \mathbf{N}_t}(\mathbf{O}_t)$ is given by a GMM that consists of $\mu_{O_{t,j,k,l}}$ and $\sigma_{O_{t,j,k,l}}^2$. Here, we use the mixture weight of a clean speech GMM and a silence GMM in place of those of the composed observation model.

4. Experiments

4.1. Experimental setup

The proposed method is evaluated by using CESNREC-1-C [5]. CENSREC-1-C was designed as an evaluation framework for VAD in noisy environments and has two types of evaluation data set, i.e., simulated data and real recorded data. In this paper, we chose the real recorded data set for the evaluation.

The data was recorded in two real noisy environments (a restaurant and a street) with two different sound pressure levels (avg. 60 dBA: High SNR and avg. 70 dBA: Low SNR). The data was originally recorded at a sampling rate of 48 kHz (with 16 bit quantization), and was down-sampled to 8 kHz. The number of speaker was ten person (five males and five females). The recorded speech consisted of four files per subject. A single file included 8-10 utterances of continuous numbers consisting of 1-12 digit numbers with two second intervals for each utterance in each noisy environment and each SNR condition. The correct segment labels were manually tagged.

The feature parameters for the proposed VAD were 24th order log-Mel spectra, which were extracted by using a Hamming window with a 25 msec frame length and a 10 msec frame shift length. We trained the silence and clean speech GMMs with 32 Gaussian distributions by using clean speech data for the HMM training of CENSREC-1 (AURORA-2J) [10]. The training data consisted of 8,440 utterances spoken by 110 speakers. The state transition probabilities of the clean speech model were set at $a_{i,j} = \{0.90, 0.10, 0.45, 0.55\}$. The variance of the driving noise $W_{t,l}$ was set at $\sigma_{W_l}^2 = 0.001$.

4.2. Experimental results of VAD

In the evaluation, we compare the VAD performance of the proposed method with the CENSREC-1-C baseline. The baseline VAD technique of CENSREC-1-C is energy-based VAD with adaptive thresholding.

The first evaluation is a frame-level evaluation. The evaluation criteria are the false rejection rate (FRR) and false acceptance rate (FAR) as shown by Eqs. (20) and (21).

$$\text{FRR} = N_{FR} / N_s \times 100 [\%] \quad (20)$$

$$\text{FAR} = N_{FA} / N_{ns} \times 100 [\%] \quad (21)$$

where N_s , N_{ns} , N_{FR} , and N_{FA} are the total number of speech frames, the total number of non-speech frames, the number of speech frames detected as non-speech frames, and the number of non-speech frames detected as speech frames, respectively. FRR and FAR are controlled by the threshold, and have a trade-off relationship. Thus, threshold was adjusted to a value that made the average FRR and FAR approximately equal.

Table 1: VAD results by frame-level evaluation

Baseline results			
	High SNR	Low SNR	Overall
FRR	23.05%	40.50%	31.78%
FAR	29.00%	27.35%	28.18%
Your results			
	High SNR	Low SNR	Overall
FRR	9.80%	18.05%	13.93%
FAR	8.75%	18.80%	13.78%

Table 2: VAD results by utterance-level evaluation

Baseline results			
	High SNR	Low SNR	Overall
Correct	56.81%	48.99%	52.90%
Accuracy	2.90%	-38.70%	-17.90%
Results			
	High SNR	Low SNR	Overall
Correct	93.05%	72.76%	82.90%
Accuracy	82.03%	37.68%	59.86%

Table 1 shows the results of a frame-level evaluation. As seen in the table, the proposed method significantly reduces both FRR and FAR from the baseline results¹.

The second evaluation is an utterance-level evaluation. The evaluation criteria are the utterance correct rate and utterance accuracy rate as shown by Eqs. (22) and (23).

$$\text{Correct} = N_c / N \times 100 \text{ [\%]} \quad (22)$$

$$\text{Accuracy} = (N_c - N_f) / N \times 100 \text{ [\%]} \quad (23)$$

where N , N_c , and N_f denote the total number of speech utterances, the number of correctly detected utterances, and the number of incorrectly detected utterances, respectively.

Table 2 shows results of an utterance-level evaluation. As seen in the table, the proposed method also significantly improves both Correct and Accuracy². In particular, the average improvement in Accuracy was approximately 78%.

4.3. Experimental results of speech recognition

We also carried out an evaluation of speech recognition with the proposed method. We used the HTK (HMM Tool Kit) [11] for speech recognition and acoustic model training. The acoustic model is trained as whole word (digit) HMMs (16 states, 20 Gaussian distributions per state) by using clean training data from CENSREC-1. The feature parameters used in this evaluation were composed of 39 MFCCs with 12 MFCCs, log-energy, and their first and second order derivatives. Cepstral mean normalization was not applied at the feature extraction. A more detailed evaluation scheme of CENSREC-1 is described in [10].

Table 3 shows the speech recognition results by word accuracy. In the table, "Baseline" (upper), "Results" (middle), and "Ideal VAD" (bottom) represent speech recognition results without VAD, with the proposed VAD, and with VAD using hand labeled utterance boundaries, respectively. The table shows that the proposed method improves speech recognition accuracy. In the speech recognition results obtained with the proposed method, there was an increase in the deleted word and substituted word errors caused by VAD errors. However, the insertion word error, especially in the silent periods between utterances, was significantly reduced. Therefore, we can confirm that the proposed method contributes to an improvement in speech recognition accuracy by reducing insertion word error.

¹The average (overall) FRR and FAR by Sohn's method were 21.80% and 23.70%, respectively.

²The average (overall) Correct and Accuracy by Sohn's method were 68.26% and 35.22%, respectively.

Table 3: Speech recognition results with VAD (%)

Baseline results (without VAD)			
	Restaurant	Street	Overall
High SNR	45.17	34.43	39.80
Low SNR	1.28	25.23	13.26
Overall	23.23	29.83	26.53
Results			
	Restaurant	Street	Overall
High SNR	49.64	44.73	47.19
Low SNR	10.34	31.61	20.98
Overall	29.99	38.17	34.08
Results with ideal VAD			
	Restaurant	Street	Overall
High SNR	52.67	41.25	46.96
Low SNR	29.17	29.50	29.34
Overall	40.92	35.38	38.15

5. Conclusion

This paper presented a noise robust VAD technique based on a switching Kalman filter. The evaluation results show that our proposed method significantly improves VAD accuracy compared with the baseline of CENSREC-1-C. In addition the proposed method also improves speech recognition accuracy. In the future, we are planning to investigate the combination of robust feature extraction and optimal threshold decision.

6. Acknowledgements

The present study was conducted using a CENSREC-1-C database developed by the IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group.

7. References

- [1] Rabiner, L. R. *et al.*, "An algorithm for determining the endpoints of isolated utterances," *The Bell System Technical Journal*, Vol. 54, No. 2, pp. 297–315, Feb. 1975.
- [2] Ramirez, J. *et al.*, "Efficient voice activity detection algorithm using long-term speech information," *Speech Communication*, Vol. 42, pp. 271–287, Apr. 2004.
- [3] Ishizuka, K. *et al.*, "A feature for voice activity detection derived from speech analysis with the exponential autoregressive model," *Proc. of ICASSP '06*, Vol. I, pp. 789–792, May 2006.
- [4] Sohn, J. *et al.*, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, Vol. 6, No. 1, pp. 1–3, Jan. 1999.
- [5] CENSREC-1-C Web site, <http://sp.shinshu-u.ac.jp/CENSREC/en/CENSREC/CENSREC-1-C/>
- [6] Ephraim, Y. *et al.*, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on Acoust., Speech, and Signal Processing*, Vol. ASSP-32, pp. 1109–1121, Dec. 1984.
- [7] Arulampalam, M. S. *et al.*, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. on Signal Processing*, Vol. 50, No. 2, pp. 174–188, Feb. 2002.
- [8] Fujimoto, M. *et al.*, "Noise robust voice activity detection based on statistical model and parallel non-linear Kalman filtering," *Proc. of ICASSP '07*, Vol. IV, pp. 797–800, Apr 2007.
- [9] Kim, N. S., "Time-varying noise compensation using multiple Kalman filters," *Proc. of ICASSP '99*, Vol. I, pp. 429–432, Mar. 1999.
- [10] Nakamura, S. *et al.*, "AURORA-2J, An evaluation framework for Japanese noisy speech recognition," *IEICE Trans. on Inf. & Syst.*, Vol. E88-D, No. 3, pp. 535–544, Mar. 2005.
- [11] HTK Web site, <http://htk.eng.cam.ac.uk/>