Noise Robust Voice Activity Detection Based on Switching Kalman Filter

Masakiyo FUJIMOTO^{†a)} and Kentaro ISHIZUKA^{†b)}, Members

SUMMARY 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. In this paper, we carried out two evaluations. In the first, we observed that the proposed method significantly outperforms conventional methods as regards voice activity detection accuracy in simulated noise environments. Second, we evaluated the proposed method on 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. In addition, we confirmed that the proposed method helps to improve the accuracy of concatenated speech recognition in real environments.

key words: voice activity detection, statistical model, switching Kalman filter, noisy 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.

VAD usually 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]–[5]. These parameters can improve 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, which makes it difficult to achieve suffi-

cient 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 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.* [6]. 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 a hidden Markov model (HMM)-based hang-over scheme that is equivalent to forward probability estimation. The speech and non-speech states of the observed signal are distinguished by thresholding the likelihood (forward probability) ratio, and the signals assigned to the speech state are extracted as the speech period. Sohn's method calculates the likelihood of each state by using a pseudo-method, i.e., *a priori* and *a posteriori* SNR-based approaches [7]. However, the estimation error of each SNR seriously affects the VAD accuracy. With respect to this problem, the positive utilization of suitable statistical models of each state will provide an accurate likelihood, because likelihood calculation with elaborate models is more flexible than the SNR-based approach.

Moreover, Sohn's method performs robustly in noisy environments; however, the 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 most cases, a noise observed in the real world has non-stationary characteristics and it is not known in advance. Thus, robustness in the presence of non-stationary noise and the absence 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 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 has been 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 an updated noise model. With the method, the Kalman filter for noise updating is formulated by using the PDF parameters of the clean speech

Manuscript received June 29, 2007.

Manuscript revised September 12, 2007.

[†]The authors are with NTT Communication Science Laboratories, NTT Corporation, Kyoto-fu, 619-0237 Japan.

a) E-mail: masakiyo@cslab.kecl.ntt.co.jp

b) E-mail: ishizuka@cslab.kecl.ntt.co.jp

DOI: 10.1093/ietisy/e91-d.3.467

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 we have adapted (composed) the model with the switching Kalman filter, we calculate the likelihood ratio between a speech state and a non-speech state.

The proposed method was first evaluated using Japanese speech data corrupted by real background noise. The evaluation results showed that the proposed method significantly improves VAD accuracy compared with conventional methods. In particular, we confirmed that the noise model estimation contributes greatly to the improvement of VAD accuracy. Second, the proposed method was evaluated with the CENSREC-1-C database (Corpora and Environments for Noisy Speech REcognition-1 Concatenated) [8], 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 helps to improve the speech recognition accuracy of concatenated utterances.

In this paper, Sec. 2 is a brief review of the statistical VAD approach proposed by Sohn *et al.*, Sec. 3 describes the formulation of switching Kalman filter-based VAD, Sec. 4 reports the VAD evaluation results, Sec. 5 reports the evaluation results obtained with the CENSREC-1-C framework and the speech recognition performance, and Sec. 6 summarizes the paper and describes future directions.

2. Statistical Model-Based VAD

In this section, we briefly review the concept of statistical VAD proposed by Sohn *et al.* [6]. 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 Fig. 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 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:

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

Fig. 1 Speech/non-speech state transition model with noise dynamics. (H_0 : Non-speech state, H_1 : Speech state)

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

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

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

The LRT with the first order Markov chain is called an HMM-based hang-over scheme [6].

In Eq. (3), the calculation of $b_j(\mathbf{O}_t)$ is crucial accurate VAD. In the original statistical VAD method proposed by Sohn *et al.*, the output probability $b_j(\mathbf{O}_t)$ is given by the following complex Gaussian distributions [6].

$$b_0(\mathbf{O}_t) = p(\mathbf{O}_t | q_t = H_0) = \prod_{m=0}^{M-1} \frac{1}{\pi \lambda_{N_m}} \exp \left\{ -\frac{|O_{t,m}|^2}{\lambda_{N_m}} \right\} \quad (5)$$

$$b_1(\mathbf{O}_t) = p(\mathbf{O}_t | q_t = H_1) = \prod_{m=0}^{M-1} \frac{1}{\pi (\lambda_{S_m} + \lambda_{N_m})} \exp \left\{ -\frac{|O_{t,m}|^2}{\lambda_{S_m} + \lambda_{N_m}} \right\}, \quad (6)$$

where $O_{t,m}$, M , λ_{S_m} , and λ_{N_m} denote the m -th bin complex spectrum of observed signal, the Fourier transform order, the variance of the complex spectrum of clean speech, and the variance of the complex spectrum of noise, respectively. The ratio of $b_j(\mathbf{O}_t)$ is given by

$$\frac{b_1(\mathbf{O}_t)}{b_0(\mathbf{O}_t)} = \prod_{m=0}^{M-1} \frac{1}{1 + \xi_m} \exp \left\{ \frac{\gamma_{t,m} \xi_m}{1 + \xi_m} \right\}, \quad (7)$$

where $\xi_m = \lambda_{S_m}/\lambda_{N_m}$ and $\gamma_{t,m} = |O_{t,m}|^2/\lambda_{N_m}$, and they correspond to using *a priori* and *a posteriori* SNRs [7], respectively. Sohn's method assumes that λ_{N_m} is already known and *a priori* SNR ξ_m is given by

$$\xi_m \approx \gamma_{t,m} - 1. \quad (8)$$

Consequently, Sohn's method defines the statistical model as a matter of form, and actually calculates ratio of $b_j(\mathbf{O}_t)$ by using *a priori* and *a posteriori* SNR's.

By using the ratio of $b_j(\mathbf{O}_t)$, the likelihood ratio R_t is given by

$$R_t = \frac{a_{0,1} + a_{1,1}R_{t-1}}{a_{0,0} + a_{1,0}R_{t-1}} \cdot \frac{b_1(\mathbf{O}_t)}{b_0(\mathbf{O}_t)}. \quad (9)$$

3. VAD Based on Switching Kalman Filter

3.1 Definition of State Transition Model

With 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:

$$b_j(\mathbf{O}_t) = \sum_{k=1}^K w_{j,k} \prod_{l=0}^{L-1} \frac{1}{\sqrt{2\pi}\sigma_{O_{j,k,l}}} \exp \left\{ -\frac{(O_{t,l} - \mu_{O_{j,k,l}})^2}{2\sigma_{O_{j,k,l}}^2} \right\}, \quad (10)$$

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, we must 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:

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

With this definition, we modeled the speech state transition model by using the ergodic state transition model shown in Fig. 1. The PDF of each state is given by GMM. This model is same as the ergodic 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 the sequential state transition model shown in Fig. 2. Finally, by composing speech and noise models, we can construct the speech / non-speech state transition model with noise dynamics shown in Fig. 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

Fig. 2 State transition model of noise.

Fig. 3 Speech / non-speech state transition model with noise dynamics.

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:

$$p(q_t | \mathbf{O}_{0:t}, \mathbf{N}_{0:t}) = \frac{p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t})}{p(\mathbf{O}_{0:t}, \mathbf{N}_{0:t})} \propto p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t}) \quad (11)$$

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

$$\begin{aligned} p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t}) &= \sum_{q_{t-1}} p(q_t, \mathbf{N}_t | q_{t-1}, \mathbf{N}_{t-1}) p(\mathbf{O}_t | q_t, \mathbf{N}_t) \\ &\quad \times p(\mathbf{O}_{0:t-1}, q_{t-1}, \mathbf{N}_{0:t-1}). \end{aligned} \quad (12)$$

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

$$\begin{aligned} p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t}) &= \sum_{q_{t-1}} p(q_t | q_{t-1}) p(\mathbf{N}_t | \mathbf{N}_{t-1}) p(\mathbf{O}_t | q_t, \mathbf{N}_t) \\ &\quad \times p(\mathbf{O}_{0:t-1}, q_{t-1}, \mathbf{N}_{0:t-1}). \end{aligned} \quad (13)$$

By defining $p(\mathbf{O}_t | q_t = H_j, \mathbf{N}_t) = b_{j, \mathbf{N}_t}(\mathbf{O}_t)$ and $p(\mathbf{N}_t | \mathbf{N}_{t-1}) = c_{t, t-1}$, 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. (13).

$$\alpha_{j,t} = \sum_{i=0}^1 (a_{i,j} \alpha_{i,t-1}) b_{j, \mathbf{N}_t}(\mathbf{O}_t) c_{t,t-1} \quad (14)$$

In (13), $c_{t,t-1}$ is always set at 1, because we assume that the noise has a sequential state transition process. Thus, Eq. (14) is simplified as

$$\alpha_{j,t} = \sum_{i=0}^1 (a_{i,j} \alpha_{i,t-1}) b_{j, \mathbf{N}_t}(\mathbf{O}_t). \quad (15)$$

On the other hand, when we focus on the state transition model of noise shown in Fig. 2, it is also given by the following equation. This equation is completely equivalent to a statistical representation of a Kalman filter [9], [10].

$$\begin{aligned} p(\mathbf{O}_{0:t}, \mathbf{N}_{0:t}) \\ = p(\mathbf{N}_t | \mathbf{N}_{t-1}) p(\mathbf{O}_t | \mathbf{N}_t) p(\mathbf{O}_{0:t-1}, \mathbf{N}_{0:t-1}) \end{aligned} \quad (16)$$

If the probability (state) variable q_t is added to Eq. (16), the statistical process is equivalent to Eq. (13). This means that Eq. (13) 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 noise as follows:

$$N_{t+1,l} = N_{t,l} + W_{t,l} \quad (17)$$

$$W_{t,l} \sim \mathcal{N}(0, \sigma_{W_l}^2), \quad (18)$$

where $N_{t,l}$, $W_{t,l}$ and $\sigma_{W_l}^2$ denote the l -th element of $\mathbf{N}_t$, 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 [11],

$$\begin{aligned} O_{t,l} &= S_{t,l} + \log(1 + \exp(N_{t,l} - S_{t,l})) \\ &= f(S_{t,l}, N_{t,l}), \end{aligned} \quad (19)$$

where $S_{t,l}$ denotes the log-Mel spectrum of silence or clean speech given by l -th Mel frequency filter at the t -th frame.

In Eq. (19), parameter $S_{t,l}$ is usually unknown. Thus, the parameters of silence or clean speech GMMs are substituted for parameter $S_{t,l}$ as follows:

$$O_{t,l} = f(\mu_{S_{j,k,l}}, N_{t,l}) + V_{t,j,k,l} \quad (20)$$

$$V_{t,j,k,l} \sim \mathcal{N}(0, \sigma_{S_{j,k,l}}^2), \quad (21)$$

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. (20). Using these observation processes, the non-linear Kalman filter is multiplied to K types and we can obtain K types of estimation results for each GMM. We call this method parallel non-linear Kalman filtering. The estimation formula of each non-linear Kalman filter is given by

[Prediction step]

$$N_{t|t-1,j,k,l} = \hat{N}_{t-1,l} \quad (22)$$

$$\sigma_{N_{t|t-1,j,k,l}}^2 = \hat{\sigma}_{N_{t-1,l}}^2 + \sigma_{W_l}^2 \quad (23)$$

$$\mu_{O_{t|t-1,j,k,l}} = f(\mu_{S_{j,k,l}}, N_{t|t-1,j,k,l}) \quad (24)$$

$$\begin{aligned} \sigma_{O_{t|t-1,j,k,l}}^2 \\ = F_{t|t-1,j,k,l} \sigma_{N_{t|t-1,j,k,l}}^2 F_{t|t-1,j,k,l} + \sigma_{S_{j,k,l}}^2 \end{aligned} \quad (25)$$

$$F_{t|t-1,j,k,l} = \frac{\partial f(\mu_{S_{j,k,l}}, N_{t|t-1,j,k,l})}{\partial N_{t|t-1,j,k,l}}. \quad (26)$$

[Estimation step]

$$\begin{aligned} N_{t,j,k,l} \\ = N_{t|t-1,j,k,l} + G_{t,j,k,l} (O_{t,l} - \mu_{O_{t|t-1,j,k,l}}) \end{aligned} \quad (27)$$

$$\sigma_{N_{t,j,k,l}}^2 = (1 - G_{t,j,k,l} F_{t|t-1,j,k,l}) \sigma_{N_{t|t-1,j,k,l}}^2 \quad (28)$$

$$G_{t,j,k,l} = \frac{\sigma_{N_{t|t-1,j,k,l}}^2 F_{t|t-1,j,k,l}}{\sigma_{O_{t|t-1,j,k,l}}^2} \quad (29)$$

$$\mu_{O_{t,j,k,l}} = f(\mu_{S_{j,k,l}}, N_{t,j,k,l}) \quad (30)$$

$$\sigma_{O_{t,j,k,l}}^2 = F_{t,j,k,l} \sigma_{N_{t,j,k,l}}^2 F_{t,j,k,l} + \sigma_{S_{j,k,l}}^2 \quad (31)$$

$$F_{t,j,k,l} = \frac{\partial f(\mu_{S_{j,k,l}}, N_{t,j,k,l})}{\partial N_{t,j,k,l}}, \quad (32)$$

where subscript $t|t-1$ denotes the predicted parameter from the $t-1$ -th frame. $N_{t,j,k,l}$ and $\sigma_{N_{t,j,k,l}}^2$ denote an $N_{t,l}$ candidate estimated using the parameters of the k -th Gaussian distribution contained in model j (silence or speech GMM) and the squared error variance of each candidate, respectively. $N_{t-1,l}$, and $\sigma_{N_{t-1,l}}^2$ denote the estimation results at the previous frame, respectively. $\mu_{O_{t|t-1,j,k,l}}$ and $\sigma_{O_{t|t-1,j,k,l}}^2$ denote the predicted mean and variance of the observed signal $O_{t,l}$ given by the previous frame $t-1$, respectively. $\mu_{O_{t,j,k,l}}$ and $\sigma_{O_{t,j,k,l}}^2$ denote the estimated mean and variance of $O_{t,l}$, respectively.

The estimated candidates are unified by weighted averaging as follows:

$$N_{t,j,l} = \sum_{k=1}^K w_{N_{t,j,k}} \cdot N_{t,j,k,l} \quad (33)$$

$$\sigma_{N_{t,j,l}}^2 = \sum_{k=1}^K w_{N_{t,j,k}} \cdot \sigma_{N_{t,j,k,l}}^2 \quad (34)$$

$$w_{N_{t,j,k}} = \frac{w_{S_{j,k}} \mathcal{N}(\mathbf{O}_t; \boldsymbol{\mu}_{O_{t,j,k}}, \boldsymbol{\Sigma}_{O_{t,j,k}})}{\sum_{k'=1}^K w_{S_{j,k'}} \mathcal{N}(\mathbf{O}_t; \boldsymbol{\mu}_{O_{t,j,k'}}, \boldsymbol{\Sigma}_{O_{t,j,k'}})} \quad (35)$$

where $N_{t,j,l}$ and $\sigma_{N_{t,j,l}}^2$ denote averaged results. $w_{N_{t,j,k}}$ and $w_{S_{j,k}}$ denote the weight for each candidate and the mixture weight of speech or silence GMM, respectively. $\boldsymbol{\mu}_{O_{t,j,k}}$ is a vector that has $\mu_{O_{t,j,k,l}}$ in each element and $\boldsymbol{\Sigma}_{O_{t,j,k}}$ is a matrix that has $\sigma_{O_{t,j,k,l}}^2$ in each diagonal element.

The probability $\sum_{k=1}^K w_{S_{j,k}} \mathcal{N}(\mathbf{O}_t; \boldsymbol{\mu}_{O_{t,j,k}}, \boldsymbol{\Sigma}_{O_{t,j,k}})$, a denominator of Eq. (35), is equivalent to the output probability $b_{j,N_t}(\mathbf{O}_t)$ of non-speech state ($j = 0$, silence + noise) or speech state ($j = 1$, speech + noise). Consequently, this means that the proposed method solves the composition of the state transition model and the likelihood calculation in the framework of parallel non-linear Kalman filtering. In addition, the proposed method has the characteristics of the switching Kalman filter, because filter equations and estimation results vary depending on the speech state q_t , i.e., type of GMM for parameter acquisition.

After the weighted averaging of Eqs. (33) and (34), the final estimation result of the noise used for estimation at the next frame is given by the weighted average with normalized $b_{j,N_t}(\mathbf{O}_t)$ as follows:

$$\hat{N}_{t,l} = \sum_{j=0}^1 \frac{b_{j,N_t}(\mathbf{O}_t)}{\sum_{j'=0}^1 b_{j',N_t}(\mathbf{O}_t)} \cdot N_{t,j,l} \quad (36)$$

$$\hat{\sigma}_{N_{t,l}}^2 = \sum_{j=0}^1 \frac{b_{j,N_t}(\mathbf{O}_t)}{\sum_{j'=0}^1 b_{j',N_t}(\mathbf{O}_t)} \cdot \sigma_{N_{t,j,l}}^2 \quad (37)$$

This method requires initial values of $\hat{N}_{t,l}$ and $\hat{\sigma}_{N_{t,l}}^2$. In this paper, we assumed that the first (0-th) frame of the observed signal is a noise frame. Thus, we use initial values of $\hat{N}_{0,l} = O_{0,l}$ and $\hat{\sigma}_{N_{0,l}}^2 = 0$.

4. Experiments

4.1 Experimental Setup

Speech signals mixed with real background noise were used in this experiment. We used Japanese speech data whose content consisted of travel arrangement dialogues [12]. The data consists of 2,292 utterances spoken by 178 speakers. The utterance duration is between 1.4 and 12.1 seconds. In terms of time the speech data accounts for approximately 2.98 hours. In the data, speech periods account for 65.5% and non-speech periods account for 34.5%. Although the data was originally recorded at a sampling rate of 48 kHz,

Table 1 Feature extraction conditions.

Sampling frequency	8 kHz (16 bit quantization)
Pre-emphasis	$1 - 0.97z^{-1}$
Feature parameters	24th order log-Mel spectra
Frame length	25 ms
Frame shift	10 ms
Window type	Hamming window

we down-sampled the data to 8 kHz. As noise data, we recorded real environmental sounds at an airport and on a street. The noise data were added to the clean speech data at SNRs of 0, 5, and 10 dB. Because environmental sounds are not stationary, we adjusted the SNR so that the power peaks of the speech and noise data within the period of an utterance were the same. Different noise intervals were added to different utterances. The feature extraction conditions are detailed in Table 1.

We trained the silence and clean speech GMMs with 32 mixture distributions by using phonetically balanced Japanese sentences. The amount of training data was 5,050 utterances spoken by 101 speakers. The feature parameters were the same as those shown in Table 1.

The state transition probabilities and the covariance matrix of driving noise were set at $a_{i,j} = \{0.90, 0.10, 0.45, 0.55\}$ and $W_{t,l} = 0.001$, respectively.

The evaluation criteria are the false rejection rate (FRR) and the false acceptance rate (FAR) as shown by Eqs. (38) and (39), respectively.

$$\text{FRR} = \frac{N_{FR}}{N_s} \times 100 [\%] \quad (38)$$

$$\text{FAR} = \frac{N_{FA}}{N_{ns}} \times 100 [\%], \quad (39)$$

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. Reference VAD labels were generated by employing the hand-labeled transcription, which includes temporal information about the speech-onset, speech-offset, and pause. FAR and FRR are controlled by a threshold or certain parameters, and have a trade-off relationship. Thus, we draw the receiver operating characteristics (ROC) curves by using several FARs and FRRs, which are obtained by changing the threshold from 0.1 to 10,000.0.

In the experiment, we compared the proposed method with following conventional methods. In each method, the frame length and the frame shift length were same as those given in Table 1.

- Sohn's method [6]: VAD, *a priori* and *a posteriori* SNR-based approach. Linear frequency (not Mel frequency) complex spectra were used for estimation of *a priori* and *a posteriori* SNR. The Fourier transform order M was 256. Noise variance λ_{N_m} was estimated by using the first ten frames of the observed signal.

- Long-term spectral divergence (LTSD) [3]: VAD based on robust feature parameters, i.e., long-term spectral divergence between the clean speech and the noise. Linear frequency power spectra were used for calculation of the spectral divergence. The Fourier transform order was 256.
- ITU-T G.729 Annex B. [13]: VAD based on the robust feature parameters, i.e., line spectra, all-band energy, low-band energy, and zero crossing ratio. This method is recommended by the ITU-T. Thus, the analysis conditions followed the ITU-T recommendation.
- ETSI ES 202 050 [14]: VAD based on robust feature parameters, i.e., all-band spectra, sub-band spectra of output signal of Wiener filter, and variance of spectra. This method is recommended by the ETSI. Thus, the analysis conditions followed the ETSI recommendation.

4.2 Experimental Results

Figure 4 shows the VAD accuracies of the proposed method and conventional methods, i.e., Sohn's statistical VAD, LTSD, ITU-T G.729 Annex B, and ETSI ES 202 050. Only one result each for ITU-T G.729 Annex B and ETSI ES 202 050 are shown in the figure, because their parameters are fixed. In the figure, the ROC curve closest to the origin shows the best performance.

The figure shows that the proposed method provides a considerable improvement compared with the conventional method as regards both airport and street noise environments. With the proposed method, the factor contributing most to the improvement was the noise estimation based on parallel non-linear Kalman filtering. The noises used in this experiment have non-stationary characteristics. Therefore, a sequential noise estimation scheme is required to improve the VAD performance. However, the conventional methods have no robust sequential noise estimation scheme. Through accurate estimation of a non-stationary noise sequence, we can obtain reliable distinctive information, e.g., energy difference between speech and noise, spectral divergence of each period, the LRT with noise adapted models, and so on. Thus, the sequential noise estimation is the most crucial factor of VAD.

Using GMMs of speech (clean speech and silence) as *a priori* knowledge also considerably contributes to the improvement, because the LRT with GMMs is more flexible and practical than the conventional *a priori* and *a posteriori* SNR-based approach. Moreover, by using the parameters of GMMs, we can robustly estimate non-stationary noise sequence.

Usually, in observed signal, both speech and noise have non-stationary characteristics, and noise varies independently. In this case, it is difficult to estimate the noise sequence without *a priori* knowledge of speech or noise. This is a typical ill-posed problem. In our approach, we assumed that the non-stationary characteristics of speech can

Table 2 CPU cycles and RTF of each method with the Intel Pentium 4 CPU (3.2 GHz). ITU-T G.729 Annex B involves VAD and speech coding. These processing are inseparable. Thus, CPU cycles and RTF of ITU-T G.729 Annex B includes processing of speech coding.

Method	CPU cycles	RTF
Sohn's method	3.6 M	0.07
LTSD	2.5 M	0.05
ITU-T G.729 Annex B	42.4 M	0.81
ETSI ES 202 050	3.3 M	0.06
Proposed	7.1 M	0.13

be described by using GMMs. With this assumption, we can approximately fix the parameters of speech for the estimation of unknown non-stationary noise by using the GMM parameters of speech. This makes robust noise estimation possible, even if both speech and noise have non-stationary characteristics.

These are conclusive factors underlying the improvement of VAD performance by the proposed method.

Table 2 shows the computational complexity of each method. The computational complexity is measured by CPU cycles and real time factor (RTF). CPU cycles is the number of executed CPU command per a second and RTF given by following equation is the required processing time per one second length signal.

$$\text{RTF} = \frac{\text{Required processing time (sec)}}{\text{Input data length (sec)}} \quad (40)$$

In the table, computation cost of the proposed method is nearly twice that of conventional methods. However, the RTF of the proposed method is sufficiently low for practical utility. Furthermore, VAD performance of the proposed method more than makes up for its computation cost.

The feature parameter used in the proposed method is the log-Mel spectrum, which is not generally a noise robust parameter. If we use robust feature parameters, i.e., those described in [4], [5], we will obtain more accurate VAD results. Thus, in further research, we plan to combine the proposed method and robust feature extraction.

In addition, finding the optimum threshold is a crucial factor of VAD techniques. The optimum threshold may be decided by an adaptive algorithm, e.g., the algorithm described in [15]. However, the optimum threshold depends on the application of the VAD technique. If VAD is applied to speech coding, the threshold is typically adjusted to a value which makes the FRR small, because to reject speech frames is undesirable. On the other hand, if VAD is applied to the speech recognizer used in a apoken dialogue system, the threshold is usually adjusted to a value which makes the FAR small, so as to restrict the incorrect action caused by insertion error of speech recognition. Thus, the threshold should be decided according to the target application. The ROC curve may be a reasonable evaluation criterion for VAD, because it helps choose a threshold resulting in the required VAD performance of a certain application.

Fig. 4 Experimental results of VAD.

5. Experiments with Using CENSREC-1-C

5.1 CENSREC-1-C

The proposed method was also evaluated by using CENSREC-1-C [8]. CENSREC-1-C was designed as an evaluation framework for VAD in noisy environments and has two types of evaluation data sets, i.e., simulated data and real recorded data. In this study, 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 (average 60 dBA: high SNR and average 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. There were ten speakers (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 under each SNR condition. The correct segment labels were manually tagged.

5.2 Experimental Results with Using CENSREC-1-C

In the evaluation, we compared 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 [8].

The feature extraction conditions of the proposed method are same as those in Table 1. The state transition probabilities and the variance of the driving noise are also same as the conditions in Sec. 4.1. The silence and clean speech GMMs with 32 Gaussian distributions were trained by using clean speech data for the HMM training of CENSREC-1 (AURORA-2J) [16]. The training data consisted of 8,440 utterances spoken by 110 speakers.

The first evaluation is a frame-level evaluation. The evaluation criteria are FRR and FAR given by Eqs. (38) and (39), respectively.

Table 3 shows the results of a frame-level evaluation. In the table, the top and bottom sections show the baseline results for CENSREC-1-C and the results obtained with the proposed method, respectively. The results in Table 3 were obtained by adjusting the threshold to make the average FRR and FAR approximately equal. As seen in the table, the proposed method significantly reduces both FRR

Table 3 VAD results for frame-level evaluation.

(a) Results of CENSREC-1-C baseline				
Real Data	False Rejection Rate [%]			
	Remote Microphone			
False Rejection Rate [%]		Restaurant	Street	Average
	High SNR	32.10	14.00	23.05
	Low SNR	48.10	32.90	40.50
	Average	40.10	23.45	31.78
Real Data	False Acceptance Rate [%]			
	Remote Microphone			
False Acceptance Rate [%]		Restaurant	Street	Average
	High SNR	3.70	54.30	29.00
	Low SNR	14.00	40.70	27.35
	Average	8.85	47.50	28.18
(b) Results of proposed method				
Real Data	False Rejection Rate [%]			
	Remote Microphone			
False Rejection Rate [%]		Restaurant	Street	Average
	High SNR	19.40	0.20	9.80
	Low SNR	36.10	0.00	18.05
	Average	27.75	0.10	13.93
Real Data	False Acceptance Rate [%]			
	Remote Microphone			
False Acceptance Rate [%]		Restaurant	Street	Average
	High SNR	6.30	11.20	8.75
	Low SNR	13.10	24.50	18.80
	Average	9.70	17.85	13.78

and FAR compared with the baseline results. Here, the average (overall) FRR and FAR obtained with Sohn's method were 21.80% and 23.70%, respectively.

The second evaluation is an utterance-level evaluation. The evaluation criteria are the utterance correct rate and the utterance accuracy rate as shown by Eqs. (41) and (42), respectively.

$$\text{Correct} = \frac{N_c}{N} \times 100 [\%] \quad (41)$$

$$\text{Accuracy} = \frac{(N_c - N_f)}{N} \times 100 [\%], \quad (42)$$

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 4 shows the results of an utterance-level evaluation. In the table, the top and bottom sections show the baseline results for CENSREC-1-C and the results obtained with the proposed method, respectively. In this evaluation, the threshold that gives the results of Table 3 was used.

As seen in the table, the proposed method also significantly improves both "Correct" and "Accuracy." In particular, the average improvement in "Accuracy" was approximately 78%. Here, the average (overall) "Correct" and "Accuracy" provided by Sohn's method were 68.26% and 35.22%, respectively.

In the restaurant noise environment with low SNR, although the proposed method significantly improves "Accuracy" from the baseline results, the absolute performance is not high. On the other hand, absolute performance of "Correct" is high. Thus, the reason why the proposed method cannot perform high absolute accuracy is mainly the large number of incorrectly detected utterances. The main source

Table 4 VAD results for utterance-level evaluation.

(a) Results of CENSREC-1-C baseline				
Real Data	Correct Rate [%]			
	Remote Microphone			
Correct Rate [%]		Restaurant	Street	Average
	High SNR	74.20	39.42	56.81
	Low SNR	56.52	41.45	48.99
	Average	65.36	40.44	52.90
Real Data	Accuracy [%]			
	Remote Microphone			
Accuracy [%]		Restaurant	Street	Average
	High SNR	21.45	-15.65	2.90
	Low SNR	-43.48	-33.91	-38.70
	Average	-11.02	-24.78	-17.90
(b) Results of proposed method				
Real Data	Correct Rate [%]			
	Remote Microphone			
Correct Rate [%]		Restaurant	Street	Average
	High SNR	88.41	97.68	93.05
	Low SNR	63.48	82.03	72.76
	Average	75.95	89.86	82.90
Real Data	Accuracy [%]			
	Remote Microphone			
Accuracy [%]		Restaurant	Street	Average
	High SNR	69.57	94.49	82.03
	Low SNR	14.49	60.87	37.68
	Average	42.03	77.68	59.86

Table 5 Utterance correct rate for the speech periods of isolated digit (%).

Noise environment	Baseline	Sohn	Proposed
Restaurant – High SNR	80.81	91.59	96.26
Restaurant – Low SNR	72.73	55.14	75.70
Street – High – SNR	66.67	93.46	98.13
Street – Low SNR	72.73	62.62	90.65
Average	73.23	75.70	90.19

Table 6 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

of restaurant noise is utterances of other speakers. In this environment, the noise estimation and the LRT-based decision confuse target speaker's utterance with those from other speakers. Since most of other speakers' utterances come from further away than the target speaker, the low frequency component of other speaker's utterances may be attenuated. Thus, the confusion can be improved by focusing the low frequency component of observed signal.

Table 7 Confusion matrices of speech recognition results. “Z” and “0” are pronounced “zero” and “maru,” respectively. Each column shows recognized words and each row shows input speech words. The diagonal cells and others show the number of correct words and the number of substitution error words, respectively. “Del.” and “Ins.” show the number of deletion errors and the number of insertion errors, respectively.

(a) Confusion matrices of baseline

Restaurant – High SNR													
Digit	1	2	3	4	5	6	7	8	9	Z	0	Del.	
1	7	8	37	17	0	0	0	0	0	1	0	33	
2	0	74	10	5	0	0	0	0	0	0	0	13	
3	0	0	94	1	0	0	0	0	0	0	0	0	
4	0	0	0	100	0	0	0	0	0	0	0	0	
5	0	0	10	6	72	0	0	0	0	0	0	14	
6	0	1	20	37	0	17	2	0	0	0	0	23	
7	0	0	7	2	0	0	88	0	0	0	0	7	
8	0	2	37	10	0	0	1	22	0	0	3	26	
9	0	0	23	11	0	0	2	0	32	0	0	28	
Z	0	0	4	8	0	0	4	0	0	76	0	7	
0	0	0	9	3	0	0	0	0	0	0	70	14	
Ins.	0	7	107	34	0	0	7	0	0	1	0		

Street – High SNR													
Digit	1	2	3	4	5	6	7	8	9	Z	0	Del.	
1	0	4	29	8	0	0	1	0	0	0	0	61	
2	0	45	15	4	1	0	0	0	0	0	0	37	
3	0	0	88	0	0	0	2	0	0	0	0	5	
4	0	0	2	95	0	0	0	0	0	0	0	3	
5	0	1	10	5	53	0	2	0	0	0	0	31	
6	0	1	17	19	0	10	4	0	0	0	0	49	
7	0	0	4	3	0	0	83	0	0	0	1	13	
8	0	2	21	9	0	0	1	12	0	0	2	54	
9	0	0	18	6	0	0	1	0	10	0	0	61	
Z	0	0	6	8	0	0	6	0	0	45	0	34	
0	0	0	8	4	0	0	1	0	0	0	56	27	
Ins.	0	2	70	37	1	0	5	0	0	0	4		

Restaurant – Low SNR													
Digit	1	2	3	4	5	6	7	8	9	Z	0	Del.	
1	0	3	14	46	0	0	14	0	0	0	0	26	
2	0	18	11	31	0	0	12	0	0	0	0	30	
3	0	0	68	15	0	0	6	0	0	0	1	5	
4	0	0	0	98	0	0	0	0	0	0	0	2	
5	0	0	8	41	28	0	8	0	0	1	1	15	
6	0	0	8	56	1	2	9	0	0	0	5	19	
7	0	0	2	6	0	0	92	0	0	0	0	4	
8	0	1	14	45	0	1	4	2	0	2	9	23	
9	0	0	14	48	0	0	11	0	3	0	1	19	
Z	0	0	4	25	0	0	11	0	0	47	1	11	
0	0	0	5	14	0	0	8	0	0	0	49	20	
Ins.	0	0	30	244	0	0	107	0	0	2	10		

Street – Low SNR													
Digit	1	2	3	4	5	6	7	8	9	Z	0	Del.	
1	0	0	10	9	0	0	2	0	0	0	0	82	
2	0	13	8	6	0	0	1	0	0	0	0	74	
3	0	0	77	1	0	0	4	0	0	0	1	12	
4	0	0	0	86	0	0	0	0	0	0	0	14	
5	0	0	10	12	23	0	0	0	0	0	0	57	
6	0	0	8	20	0	2	3	0	0	1	0	66	
7	0	0	3	3	0	0	61	0	0	0	0	37	
8	0	1	12	10	0	0	0	3	0	0	1	74	
9	0	0	13	9	0	0	1	0	0	0	0	73	
Z	0	0	8	11	0	0	0	0	0	25	0	55	
0	0	0	7	3	0	0	3	0	0	0	30	53	
Ins.	0	0	19	16	0	0	7	0	0	0	1		

(b) Confusion matrices of proposed method

Restaurant – High SNR													
Digit	1	2	3	4	5	6	7	8	9	Z	0	Del.	
1	9	13	15	33	0	0	4	0	0	1	1	27	
2	0	86	5	6	0	0	0	0	0	0	0	5	
3	0	0	82	7	0	0	4	0	0	0	0	2	
4	0	0	0	99	0	0	0	0	0	0	0	1	
5	0	0	3	11	82	0	0	0	0	0	0	6	
6	0	3	13	43	0	23	3	0	0	1	0	14	
7	0	0	3	3	0	0	96	0	0	0	0	2	
8	0	3	22	13	0	0	6	30	0	1	8	18	
9	0	4	15	19	0	0	3	0	43	0	0	12	
Z	0	0	2	7	0	0	4	0	0	83	0	3	
0	0	0	2	7	0	0	1	0	0	0	82	4	
Ins.	0	15	46	82	0	0	23	0	0	0	4		

Street – High SNR													
Digit	1	2	3	4	5	6	7	8	9	Z	0	Del.	
1	0	4	5	52	0	0	24	0	0	0	1	20	
2	0	23	3	38	0	0	16	0	0	0	3	29	
3	0	0	41	27	1	0	23	0	0	0	2	5	
4	0	0	0	94	0	0	1	0	0	0	0	10	
5	0	0	3	35	33	0	14	0	0	0	0	26	
6	0	0	4	48	0	1	20	0	0	1	2	28	
7	0	0	0	2	0	0	98	0	0	0	0	12	
8	0	2	5	47	1	0	16	6	0	1	10	19	
9	0	0	5	54	0	0	19	0	2	0	3	18	
Z	0	0	2	24	1	0	12	0	0	52	4	9	
0	0	0	2	15	0	0	13	0	0	0	50	19	
Ins.	0	2	4	161	0	0	105	0	0	4	4		

Restaurant – Low SNR													
Digit	1	2	3	4	5	6	7	8	9	Z	0	Del.	
1	2	15	3	34	0	0	3	0	0	1	1	52	
2	0	86	1	6	0	0	1	0	0	0	2	14	
3	0	0	59	13	1	0	9	0	0	0	1	16	
4	0	0	1	98	0	0	0	0	0	0	0	5	
5	0	0	1	7	72	0	4	0	0	0	1	22	
6	0	3	2	41	0	14	8	0	0	2	0	36	
7	0	0	0	2	0	0	92	0	0	0	1	19	
8	0	5	8	22	0	0	7	22	0	0	1	42	
9	0	2	9	29	0	0	5	0	13	0	0	41	
Z	0	1	2	16	0	0	11	0	0	59	0	14	
0	0	0	2	5	0	0	5	0	0	0	77	10	
Ins.	0	9	1	43	1	0	14	0	1	0	7		

Street – Low SNR													
Digit	1	2	3	4	5	6	7	8	9	Z	0	Del.	
1	0	5	3	24	0	0	7	0	0	0	1	67	
2	0	40	3	24	0	0	4	0	0	0	0	40	
3	0	0	31	14	1	0	14	0	0	0	2	36	
4	0	0	0	91	0	0	2	0	0	1	0	9	
5	0	0	5	17	38	0	12	0	0	0	0	40	
6	0	1	2	35	0	3	10	0	0	1	0	54	
7	0	0	0	4	0	0	96	0	0	0	0	9	
8	0	3	4	20	0	0	12	7	0	0	0	65	
9	0	1	2	26	0	0	12	0	2	0	1	54	
Z	0	1	0	11	0	1	10	0	0	36	0	44	
0	0	0	1	10	1	0	6	0	0	0	60	25	
Ins.	0	0	1	26	0	0	9	0	0	0	1		

Next, to evaluate convergence performance of the proposed method, we carried out an evaluation by using a speech data with extremely short time duration. In this evaluation, we used only speech periods of isolated digit contained in the real data of the CENSREC-1-C. The number of isolated digit in a noise environment is 99, and the average utterance duration is approximately 0.37 seconds. To

evaluate correctness of short time speech detection, utterance correct rate is used for the evaluation criterion.

Table 5 shows the evaluation results of isolated digit periods. In the table, the proposed method shows the significant improvements compared with the conventional method, even if the input speech length is extremely short. This means that the proposed has rapid convergence perfor-

mance.

5.3 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) [17] 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. The other evaluation scheme is also the same as the baseline evaluation of CENSREC-1 [16].

Table 6 shows the speech recognition results for word accuracy. In the table, "Baseline" (top), "Results" (middle), and "Ideal VAD" (bottom) represent speech recognition results without VAD technique, with the proposed VAD, and with VAD using hand-labeled utterance boundaries, respectively. For this evaluation, we used the threshold that gives the results in Tables 3 and 4.

The table shows that the proposed method improves speech recognition accuracy. The speech recognition accuracy of Table 6 is not high, because no noise suppression method or noise adaptation method was used for speech recognition. However, the proposed method contributes to the improvement of concatenated speech recognition in real noise environments without noise compensation methods.

From the results of Table 4, we can see a tendency that if the improvement of "Correct" and "Accuracy" is large, the improvement of speech recognition accuracy is large high[†]. This also shows the importance of correctness of utterance detection and reduction of the incorrect utterance detection for concatenated speech recognition.

Table 7 shows the confusion matrices of each noise environment. In the table, we can see that the error tends to concentrate on some words, especially word "4". In addition, the number correct for word "1" is small in all environments. In further research, we plan to examine the reason for the error concentration and inaccuracy of word "1".

In the confusion matrices of the speech recognition results obtained with the proposed method, there was an increase in the deletion word or substitution word errors caused by VAD errors. However, insertions, especially in the silent periods between utterances, were reduced and the number of correct words also increased. This may be caused by decision of utterance boundaries. Therefore, we can confirm that the proposed method contributes to an improvement in speech recognition accuracy by increasing the number of correct hits and reducing insertion errors.

6. Conclusion

This paper presented a noise robust VAD technique based-

[†]Note that the improvement in speech recognition accuracy depends on the type of noise.

on a switching Kalman filter. The evaluation results show that our proposed method significantly improves VAD accuracy compared with the conventional method or the baseline of CENSREC-1-C. In addition the proposed method also contributes to an improvement in speech recognition accuracy. In the future, we are planning to investigate the combination of robust feature extraction and the optimum threshold decision.

In addition, if we can use a noise suppression method, we will obtain a more significant improvement in speech recognition. With the noise suppression methods, VAD is usually used to design the noise suppression filter by using the data of a speech period. Thus, integration of VAD and noise suppression is important factor for concatenated speech recognition in noise. We are also planning to investigate this integration.

Acknowledgements

The present study was conducted using the CENSREC-1-C database and the AURORA-2J database developed by the IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group.

References

- [1] L.R. Rabiner and M.R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol.54, no.2, pp.297-315, Feb. 1975.
- [2] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Speech Audio Process.*, vol.9, no.3, pp.217-231, March 2001.
- [3] J. Ramirez, J.C. Segura, C. Benitex, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithm using long-term speech information," *Speech Commun.*, vol.42, pp.271-287, April 2004.
- [4] K. Ishizuka and H. Kato, "A feature for voice activity detection derived from speech analysis with the exponential autoregressive model," *Proc. ICASSP '06*, vol.1, pp.789-792, Toulouse, France, May 2006.
- [5] K. Ishizuka and T. Nakatani, "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio," *Proc. SAPA '06*, pp.65-70, Pittsburgh, PA, USA, Sept. 2006.
- [6] J. Sohn, N.S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol.6, no.1, pp.1-3, Jan. 1999.
- [7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol.32, no.12, pp.1109-1121, Dec. 1984.
- [8] CENSREC-1-C Web site, <http://sp.shinshu-u.ac.jp/CENSREC/en/CENSREC/CENSREC-1-C/>
- [9] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online non-linear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol.50, no.2, pp.174-188, Feb. 2002.
- [10] A.V. Balakrishnan, *Kalman filtering theory*, Springer-Verlag, 1984.
- [11] J.C. Segura, A. de la Torre, M.C. Benitez, and A.M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. experiments using AURORA II database and tasks," *Proc. EuroSpeech '01*, vol.1, pp.221-224, Aalborg, Denmark, Sept. 2001.

- [12] A. Nakamura, S. Matsunaga, T. Shimizu, M. Tonomura, and Y. Sagisaka, "Japanese speech database for robust speech recognition," Proc. ICSLP '96, vol.IV, pp.2199–2202, Philadelphia, USA, Oct. 1996.
- [13] ITU-T Recommendation G.729 Annex B., "A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70," Nov. 1996.
- [14] ETSI standard document, "Speech processing, transmission and quality aspects (STQ), advanced distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms," ETSI ES 202 050 v.1.1.4, Nov. 2005.
- [15] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Trans. Syst. Man Cybern., vol.9, no.1, pp.62–66, Jan. 1979.
- [16] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto, and T. Endo, "AURORA-2J, An evaluation framework for Japanese noisy speech recognition," IEICE Trans. Inf. & Syst., vol.E88-D, no.3, pp.535–544, March 2005.
- [17] HTK Web site, <http://htk.eng.cam.ac.uk/>

Masakiyo Fujimoto received B.E., M.E., and Doctor of Engineering degrees all from Ryukoku University in 1997, 2001, and 2005, respectively. From 2004–2006, he worked with ATR Spoken Language Communication Research Laboratories. He is currently a researcher at NTT Communication Science Laboratories, NTT Corporation, Japan. He received the Awaya Award from ASJ in 2003. His current research interests include noise-robust speech recognition. He is a member of the IEEE, ISCA,

IPSJ, and ASJ.

Kentaro Ishizuka received the B.E. and M.E from Kyoto University in 1997 and 2000, respectively. Since 2000, he has been a member of NTT Communication Science Laboratories, NTT Corporation, Japan. He received the Awaya Award from ASJ in 2007. His research interests include acoustic, speech, signal processing, hearing psychology and physiology, speech perception, automatic speech recognition, and auditory scene analysis. He is a member of IEEE, ISCA, ASA, and ASJ.