NOISE ROBUST VOICE ACTIVITY DETECTION BASED ON STATISTICAL MODEL AND PARALLEL NON-LINEAR KALMAN FILTERING

Masakiyo Fujimoto, Kentaro Ishizuka, and Hiroko Kato

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

ABSTRACT

This paper addresses the problem of voice activity detection in noise environments. The proposed voice activity detection technique described in this paper is based on a statistical model approach, and estimates the statistical models sequentially without *a prior* knowledge of noise. The crucial factor as regards the statistical model-based approach is noise parameter estimation, especially non-stationary noise. To deal with this problem, a parallel non-linear Kalman filter, that is a multiplied estimator, is used for sequential noise estimation. Also, a backward estimation is used for noise estimation and likelihood calculation for speech / non-speech discrimination. In the evaluation results, we observed that the proposed method significantly outperforms conventional methods as regards voice activity detection accuracy in noisy environments.

***Index Terms*—** Speech processing, State space methods, Kalman filtering, Multiplied estimator, Forward-backward estimation

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, 4]. These parameters can improve the VAD accuracy. However, 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 noise 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 has been proposed as a robust decision mechanism by Sohn *et al.* [5]. 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 [6]. 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. Thus, we positively utilize statistical models, i.e., Gaussian mixture models (GMMs) of noise (noise + silence) and noisy speech (noise + speech), and calculate the likelihood of speech and non-speech states.

With the proposed method, the GMMs of noise and noisy speech are composed by Log-Add composition [7] using a noise mean vector and GMMs of silence and clean speech. The GMMs of silence and clean speech can be trained in advance using a clean speech corpus. On the other hand, to cope with non-stationary noise, the noise mean vector is sequentially estimated by using a parallel non-linear Kalman filter that is a multiplied parameter estimator. In addition, a backward techniques, i.e., a parallel Kalman smoother and a backward probability estimation, are used to estimate the noise mean vector and for the likelihood calculation.

The proposed method was evaluated using Japanese speech data corrupted by real background noise. In the evaluation results, we observed that the proposed method significantly improves VAD accuracy compared with conventional methods. In particular, we confirmed that the noise mean vector estimation contributes greatly to the improvement of VAD accuracy.

2. STATISTICAL MODEL-BASED VAD

In this section, we briefly review the concept of the statistical VAD proposed by Sohn *et al.* [5]. The 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, 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

1-4244-0728-1/07/$20.00 ©2007 IEEE

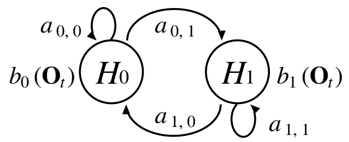

Fig. 1. Speech / non-speech state transition model

or non-speech periods is equivalent to the estimation of the t-th frame state q_t when $\mathbf{O}_{0:t} = \{\mathbf{O}_0, \cdots, \mathbf{O}_t\}$ 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:

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

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

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:

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

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

$$q_t = \begin{cases} H_0 & R_t < \text{Threshold} \\ H_1 & R_t \geq \text{Threshold} \end{cases} . \tag{4}$$

The LRT with the first order Markov chain is called an HMM-based hang-over scheme [5].

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 [5, 6]. The details of the $b_j(\mathbf{O}_t)$ calculation are described in [5].

3. LIKELIHOOD CALCULATION BASED ON SILENCE AND SPEECH MODELS

3.1. Definition of probability density functions

It is obvious that the framework of the HMM-based hang-over scheme is based on a statistical approach, however, the statistical VAD proposed by Sohn *et al.* did not strictly use a statistical model. Although they calculated output probability $b_j(\mathbf{O}_t)$ using *a priori* and *a posteriori* SNRs, it is a pseudo-method, and $b_j(\mathbf{O}_t)$ is not directly calculated by using any kind of probability density function (PDF). Furthermore, the estimation error of each SNR seriously affects the estimation accuracy of $b_j(\mathbf{O}_t)$.

On the other hand, if we can choose suitable PDFs, a more accurate estimate of $b_j(\mathbf{O}_t)$ will be obtained, because the likelihood calculation with PDFs is more flexible and applicable than the conventional *a priori* and *a posteriori* SNR-based approach. Thus, we looked at how to calculate $b_j(\mathbf{O}_t)$ directly using PDFs. As the PDFs for likelihood calculation, we chose a GMM modeled in the log-Mel spectral domain as follows:

$$b_j(\mathbf{O}_t) = \sum_{k=1}^{K} w_{j,k} \prod_{l=0}^{L-1} \frac{1}{\sqrt{2\pi}\sigma_{j,k,l}} \exp\left\{ -\frac{(O_{t,l} - \mu_{j,k,l})^2}{2\sigma_{j,k,l}^2} \right\} \tag{5}$$

where $w_{j,k}$, $O_{t,l}$, $\mu_{j,k,l}$, and $\sigma_{j,k,l}^2$ 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. In 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 prior* knowledge of noise. To cope with unknown noise environments, it is necessary to construct environmentally matched model sets by using an on-line estimation. In this problem, we first defined non-speech and speech periods as follows:

$q_t = H_0$:	Non-speech period :	Silence + Noise
$q_t = H_1$:	Speech period :	Speech + Noise

Using this definition, if a silence GMM, a clean speech GMM, and a noise mean vector are given, the noise GMM and the noisy speech GMM can be composed by Log-Add composition [7]. The Log-Add composition is given by

$$\mu_{j,k,l} = \mu_{S,j,k,l} + \log(1 + \exp(\mu_{N,l} - \mu_{S,j,k,l})) \tag{6}$$

where $\mu_{S,j,k,l}$ and $\mu_{N,l}$ denote a mean of silence ($j = 0$) or speech ($j = 1$) GMM and a noise mean in the log-Mel spectral domain, respectively. Usually, The Log-Add composition is applied only to mean parameters, thus, the variance parameters are given by

$$\sigma_{j,k,l}^2 \simeq \sigma_{S,j,k,l}^2 \tag{7}$$

where $\sigma_{S,j,k,l}^2$ denotes the variance of a silence or speech GMM in the log-Mel spectral domain.

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 mean $\mu_{N,l}$ is unknown. Therefore, we estimate the noise mean sequentially by using a parallel non-linear Kalman filter described in the next section. In addition, the noise GMM and noisy speech GMM means are also sequentially updated with the noise mean updating technique.

3.2. Sequential noise estimation by parallel non-linear Kalman filtering

Sohn's statistical VAD assumed that noise has stationary characteristics. However, most of the noise observed in real environments has non-stationary characteristics. To improve VAD accuracy in non-stationary noise environments, it is necessary to estimate the noise sequence as accurately as possible. To overcome this problem, we introduce Kalman filtering into sequential noise estimation, and the noise mean is estimated as $\mu_{N,l} = N_{t,l}$ where $N_{t,l}$ denotes the estimated noise.

Kalman filtering 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:

$$N_{t+1,l} = N_{t,l} + W_{t,l} \tag{8}$$

$$W_{t,l} \sim \mathcal{N}\left(0, \sigma_{W_l}^2\right) \tag{9}$$

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,

$$\begin{aligned} O_{t,l} &= S_{t,l} + \log(1 + \exp(N_{t,l} - S_{t,l})) \\ &= f(S_{t,l}, N_{t,l}) \end{aligned} \tag{10}$$

where $S_{t,l}$ denotes log-Mel spectra of silence or clean speech. Note that the Kalman filter derived from a dynamical system with a non-linear scheme becomes a non-linear Kalman filter.

In Eq. (10), 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:

$$O_{t,l} = f\left(\mu_{S,j,k,l}, N_{t,l}\right) + V_{t,j,k,l} \tag{11}$$

$$V_{t,j,k,l} \sim \mathcal{N}\left(0, \sigma^2_{S_{j,k,l}}\right) \tag{12}$$

where $V_{t,j,k,l}$ 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. (11). 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

$$N_{t|t-1,j,k,l} = N_{t-1,j,l} \tag{13}$$

$$\sigma^2_{N_{t|t-1,j,k,l}} = \sigma^2_{N_{t-1,j,l}} + \sigma^2_{W_l} \tag{14}$$

$$G_{t,j,k,l} = \frac{\sigma^2_{N_{t|t-1,j,k,l}} F_{t,j,k,l}}{F_{t,j,k,l}\sigma^2_{N_{t|t-1,j,k,l}} F_{t,j,k,l} + \sigma^2_{S_{j,k,l}}} \tag{15}$$

$$F_{t,j,k,l} = \partial f\left(\mu_{S,j,k,l}, N_{t|t-1,j,k,l}\right) / \partial N_{t|t-1,j,k,l} \tag{16}$$

$$\begin{aligned} N_{t,j,k,l} &= N_{t|t-1,j,k,l} \\ &\quad + G_{t,j,k,l}\left(O_{t,l} - f\left(\mu_{S,j,k,l}, N_{t|t-1,j,k,l}\right)\right) \end{aligned} \tag{17}$$

$$\sigma^2_{N_{t,j,k,l}} = \left(1 - G_{t,j,k,l}F_{t,j,k,l}\right)\sigma^2_{N_{t|t-1,j,k,l}} \tag{18}$$

where subscript $t|t-1$ denotes the predicted parameter from the $t-1$-th frame. $N_{t,j,k,l}$ and $\sigma^2_{N_{t,j,k,l}}$ denote a candidate for $N_{t,l}$ estimated using the parameters of the k-th Gaussian distribution contained in model j (silence or speech GMM) and a the squared error variance of each candidate, respectively. $N_{t-1,j,l}$, and $\sigma^2_{N_{t-1,j,l}}$ denote the corresponding estimation results in a previous frame.

The estimated candidates are unified by weighted averaging as follows:

$$N_{t,j,l} = \sum_{k=1}^{K} w_{N_{t,j,k}} \cdot N_{t,j,k,l} \tag{19}$$

$$\sigma^2_{N_{t,j,l}} = \sum_{k=1}^{K} w^2_{N_{t,j,k}} \cdot \sigma^2_{N_{t,j,k,l}} \tag{20}$$

$$w_{N_{t,j,k}} = \frac{w_{j,k}\mathcal{N}\left(\mathbf{O}_t; \boldsymbol{\mu}_{O_{t,j,k}}, \boldsymbol{\Sigma}_{O_{t,j,k}}\right)}{\sum_{k'=1}^{K} w_{j,k'}\mathcal{N}\left(\mathbf{O}_t; \boldsymbol{\mu}_{O_{t,j,k'}}, \boldsymbol{\Sigma}_{O_{t,j,k'}}\right)} \tag{21}$$

where $N_{t,j,l}$ and $\sigma^2_{N_{t,j,l}}$ denote averaged results. $\boldsymbol{\mu}_{O_{t,j,k}}$ is a vector that has $f\left(\mu_{S,j,k,l}, N_{t,j,k,l}\right)$ in each element and $\boldsymbol{\Sigma}_{O_{t,j,k}} = diag\left(\sigma^2_{S_{j,k,l}}\right)$. The parallel non-linear Kalman filtering is carried out in accordance with the silence and speech GMMs. Thus, weighted averaging is also applied according to each GMM. The unified noise means, $N_{t,0,l}$ and $N_{t,1,l}$, are substituted to Eq. (6) as $\mu_{N,l} = N_{t,0,l}$ (for silence) or $\mu_{N,l} = N_{t,1,l}$ (for clean speech).

The denominator of Eq. (21) can also be regarded as the output probability $b_j\left(\mathbf{O}_t\right)$. However, the probability given by the denominator of Eq. (21) may be inaccurate, because the noise mean candidates $N_{t,j,k,l}$ include some outliers. To reduce the influence of these outliers, we applied weighted averaging to the candidates, and substituted $N_{t,j,l}$ to Eq. (6).

3.3. Backward estimation

The likelihood estimation described in sections 2 and 3.1 is simply carried out by employing forward estimation with the parameters of the present and preceding frames. However, the effects of future frames $t+1, \cdots, T$ are also an important factor for time series estimation. Likelihood estimation using future frames is called backward estimation, and we introduce backward estimation into LRT as follows:

$$p\left(\mathbf{O}_{0:T}, q_t\right) = p\left(\mathbf{O}_{0:t}, q_t\right) p\left(\mathbf{O}_{t+1:T}|q_t\right) . \tag{22}$$

Based on a recursive formula, the conditional probability $p\left(\mathbf{O}_{t+1:T}|q_t\right)$ is represented as

$$p\left(\mathbf{O}_{t+1:T}|q_t\right) = \sum_{q_{t+1}} p\left(q_{t+1}|q_t\right) p\left(\mathbf{O}_{t+1}|q_{t+1}\right) p\left(\mathbf{O}_{t+2:T}|q_{t+1}\right) , \tag{23}$$

and is equivalent to backward probability $\beta_{j,t}$. Using the same notation as Eq. (3), Eq. (23) is rewritten as

$$\beta_{i,t} = a_{i,0} b_0\left(\mathbf{O}_{t+1}\right)\beta_{0,t+1} + a_{i,1} b_1\left(\mathbf{O}_{t+1}\right)\beta_{1,t+1} . \tag{24}$$

Thus, the likelihood with forward-backward estimation is derived from $p\left(\mathbf{O}_{0:T}, q_t = H_j\right) = \alpha_{j,t} \cdot \beta_{j,t}$, and the likelihood ratio R_t used for LRT is given by

$$R_t = \frac{p\left(\mathbf{O}_{0:T}, q_t = H_1\right)}{p\left(\mathbf{O}_{0:T}, q_t = H_0\right)} = \frac{\alpha_{1,t} \cdot \beta_{1,t}}{\alpha_{0,t} \cdot \beta_{0,t}} . \tag{25}$$

In addition, Kalman smoother [8], which is a backward estimator of Kalman filter, is applied to noise estimation. Kalman smoother is also multiplied in the same way as the parallel non-linear Kalman filter described in section 3.2, and we call this method parallel Kalman smoothing. The smoothing formula is given by

$$J_{t,j,k,l} = \sigma^2_{N_{t,j,k,l}} / \sigma^2_{N_{t+1|t,j,k,l}} \tag{26}$$

$$\tilde{N}_{t,j,k,l} = N_{t,j,k,l} + J_{t,j,k,l}\left(\tilde{N}_{t+1,j,k,l} - N_{t+1|t,j,k,l}\right) \tag{27}$$

$$\begin{aligned} \tilde{\sigma}^2_{N_{t,j,k,l}} &= \sigma^2_{N_{t,j,k,l}} \\ &\quad + J_{t,j,k,l}\left(\tilde{\sigma}^2_{N_{t+1,j,k,l}} - \sigma^2_{N_{t+1|t,j,k,l}}\right) J_{t,j,k,l} \end{aligned} \tag{28}$$

where $\tilde{N}_{t,j,k,l}$ and $\tilde{\sigma}^2_{N_{t,j,k,l}}$ denote a smoothed candidate of $N_{t,l}$ and a smoothed squared error variance, respectively. The smoothed candidates are unified by weighted averaging in the same way as Eqs. (19) to (21).

Usually, both backward likelihood estimation and Kalman smoothing are carried out from the end of the observed signal. However, the end of the observed signal is unknown in VAD, therefore, a block-wise backward estimation is introduced, namely, backward estimations are carried out from $T = t + tb$ with a constant time length tb. Backward estimations are not performed for $tb = 0$.

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. The data consists of 2,292 utterances spoken by 178 speakers. The utterance duration is from 1.4 to 12.1 seconds. Although the data was originally recorded at a sampling rate of 48 kHz, we down-sampled the data to 8 kHz. As noise data, we recorded real environmental sounds at an airport and on

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	20 ms
Frame shift	10 ms
Window type	Hamming window

a street. The noise data were added to the clean speech data at an SNR of 0 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 256 distributions by using phonetically balanced Japanese sentences. The training data consisted of 5,050 utterances spoken by 101 speakers. The feature parameters were the same as those shown in Table 1.

State transition probabilities and the number of frames for backward estimation were set at $a_{i,j} = \{0.8, 0.2, 0.1, 0.9\}$ and $tb = \{0, 5, 10\}$, respectively. The variance of the driving noise $W_{t,l}$ was set at $\sigma^2_{W_l} = 0.001$.

The VAD evaluation criteria used in this paper are the false acceptance rate (FAR) and false rejection rate (FRR) as shown by Eqs. (29) and (30). Reference VAD labels were generated by using hand-labeled transcription, which includes temporal information on speech-onset, speech-offset, and pause. FAR and FRR are controlled by the 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.

$$\mathrm{FAR} = \frac{\text{Number of falsely detected speech frame}}{\text{Number of non-speech frame}} \qquad (29)$$

$$\mathrm{FRR} = \frac{\text{Number of falsely detected non-speech frame}}{\text{Number of speech frame}} \qquad (30)$$

4.2. Experimental results

Figure 2 shows the VAD accuracies of the proposed and conventional methods, i.e., Sohn's statistical VAD [5], the long-term spectral divergence (LTSD) [3], ITU-T G.729 Annex B [9], and ETSI ES 202 050 [10]. Only one result each of 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 origin shows the best performance.

The figure shows that the proposed method is considerably better than the conventional method in both airport and street noise environments. The proposed method with $tb = 10$ slightly improves the VAD accuracy compared with $tb = 0$, namely, noise estimation and likelihood estimation by forward estimation alone. With the proposed method, the factor that contributed most to the improvement was the noise estimation based on parallel non-linear Kalman filtering. This suggests that the noise estimation is the most crucial factor as regards statistical model-based VAD. Moreover, 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, we will obtain more accurate VAD results. Thus, in further research, we plan to combine the proposed method and robust feature extraction.

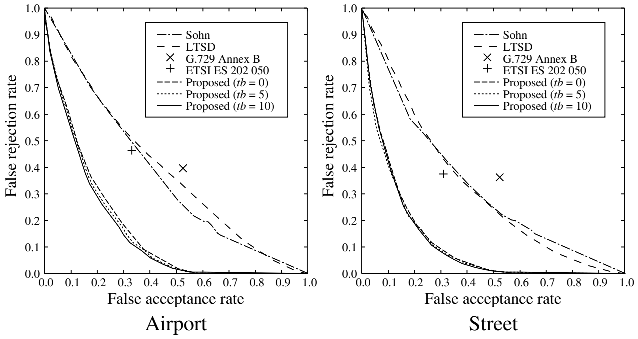

Fig. 2. Experimental results of VAD

5. CONCLUSION

This paper presented a noise robust VAD based-on statistical model. In the proposed method, we introduced a likelihood estimator that used silence and clean speech GMMs, sequential noise estimation by parallel non-linear Kalman filtering, and backward estimations into the statistical model-based approach. The evaluation results show that our proposed method significantly improves VAD accuracy compared with conventional methods. In the future, we are planning to investigate the combination of noise robust feature extraction and the optimal threshold decision.

6. REFERENCES

[1] Rabiner, L. R. and Sambur, M. R., "An algorithm for determining the endpoints of isolated utterances," *The Bell System Technical Journal*, Vol. 54, No. 2, pp. 297–315, Feb. 1975.

[2] Nemer, E., Goubran, R., and Mahmoud, S., "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. on Speech and Audio Processing*, Vol. 9, No. 3, pp. 217–231, March 2001.

[3] Ramirez, J., Segura, J.C., Benitex, C., de la Torre, A., and Rubio, A., "Efficient voice activity detection algorithm using long-term speech information," *Speech Communication*, Vol. 42, pp. 271–287, Apr. 2004.

[4] Ishizuka, K. and Kato H., "A feature for voice activity detection derived from speech analysis with the exponential autoregressive model," *Proc. of ICASSP '06*, Toulouse, France, Vol. I, pp. 789–792, May 2006.

[5] Sohn, J. , Kim, N. S., and Sung, W., "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, Vol. 6, No. 1, pp. 1–3, Jan. 1999.

[6] Ephraim, Y. and Malah, D., "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *Trans. on Acoust., Speech, Signal Processing*, Vol. ASSP–32, pp. 1109–1121, Dec. 1984.

[7] Gales, M. J. F., "Model-Based Techniques for Noise Robust Speech Recognition", *Ph.D Thesis, Cambridge University*, Sept. 1995.

[8] Balakrishnan, A.V., "Kalman Filtering Theory," *Optimization Software*, 1987.

[9] ITU-T Recommendation G.729 Annex B., "A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70," Nov. 1996.

[10] 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.