A VOICE ACTIVITY DETECTION BASED ON THE ADAPTIVE INTEGRATION OF MULTIPLE SPEECH FEATURES AND A SIGNAL DECISION SCHEME

Masakiyo Fujimoto, Kentaro Ishizuka, and Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikari-dai, Seika-cho, Souraku-gun, Kyoto, 619-0237, Japan
E-mail: {masakiyo, ishizuka, nak}@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 integrates multiple speech features and a signal decision scheme, namely the speech periodic to aperiodic component ratio and a switching Kalman filter. The integration is carried out by using the weighted sum of likelihoods outputted from each VAD (stream). The stream weight is decided adaptively each short time frame. 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. In addition, we carried out speech recognition evaluations by using detected speech signals, and confirmed that the proposed method contributes to an improvement in speech recognition accuracy.

Index Terms— voice activity detection, periodic to aperiodic component ratio, switching Kalman filter, adaptive integration

1. INTRODUCTION

Voice activity detection (VAD), which 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 noise, thus several noise robust features have been proposed [2, 3]. These parameters can improve the VAD accuracy. On the other hand, 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.

The applicable noise environment for VAD differs depending on the feature parameter and decision scheme. It is difficult to cope with all noises observed in the real world by using only one method. Thus, we investigate the VAD method with wide noise coverage by combining multiple VAD methods. For the combination methods, in this paper, we adopt a speech periodic to aperiodic component ratio (PAR)-based VAD [3] and the switching Kalman filter (SKF)-based VAD [5]. The combination is carried out by employing the adaptive weighting sum of likelihood outputted from each method independently. By using this approach, we can improve the VAD performance compared with when using each method alone.

Table 1. Comparison of VAD techniques

	VAD performance				RTF
		Noise environments			
	Clean	Station-ary	Non-stationary	Burst	
ITU-T G. 729B	○	×	×	×	0.06
ETSI ES 202 050	△	○	△	×	0.06
Sohn	○	○	×	×	0.07
Ramirez	○	○	△	×	0.05
PAR	○	○	△	○	0.06
SKF	○	○	○	△	0.10
PAR + SKF	○	○	○	○	0.13

The proposed method was evaluated on the CENSREC-1-C (Corpora and Environments for Noisy Speech RECognition-1 Concatenated) [6], 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. COMPARISON OF EACH METHOD

This section compares several VAD methods, and confirms the applicable and inapplicable noise environments for each method. Table 1 shows expected performance in several noise environments and the real time factor (RTF) of each method. The RTF was measured by using an Intel Pentium 4 3.6 GHz CPU.

Several VAD methods have been proposed, e.g., ITU-T recommendation G. 729 Annex B. [7], ETSI recommendation ES 202 050 [8], the statistical model-based approach proposed by Sohn *et al.* [4], and the spectral divergence proposed by Ramirez *et al.* [2]. However, although these methods can be applied to stationary noise environments, it is difficult to apply them to non-stationary ot burst noise environments.

Here, our proposed methods, PAR-based VAD [3] and SKF-based VAD [5], has cover a wide range of noise environments. As shown in Table 1, the PAR-based VAD is robust for not only stationary noise but also burst noise, because almost burst noises have no periodic characteristics. On the other hand, the SKF-based VAD is robust for stationary and non-stationary noises, because the method can estimate the time varying noise sequentially. However, each method has inapplicable noise environments, i.e., PAR-based VAD

1-4244-1484-9/08/$25.00 ©2008 IEEE

and SKF-based VAD are not robust for the noise with periodic characteristics and the burst noise, respectively. Consequently, by integrating each method effectively, the resulting method can cope with all noise environments with a practical RTF by compensating each inapplicable noise environment.

3. PERIODIC TO APERIODIC COMPONENT RATIO

We first provide a short explanation of the PAR calculation (see [3] for details). With this calculation, the dominant harmonic component in the observed signal is referred to as the periodic component, which is not always the target signal, and the other sound components are referred to as aperiodic components and include both environmental noise and the aperiodic components of target speech. Although the estimated power of the periodic component is affected by the changes in the aperiodic components, this effect can be mitigated in the PAR. Therefore, PAR is expected to be insensitive to dynamic changes in noise power. We define the power of the observed signal o_τ within a t-th short-time frame of length T as:

$$\rho_t = \sum_{\tau=0}^{T-1} |\hat{o}_\tau|^2 = \frac{1}{M}\sum_{m=1}^{M} \left|O_{t,m}^{STFS}\right|^2 , \qquad (1)$$

where ρ_t is the power of the observed signal, and $O_{t,m}^{STFS}$ is a short-time Fourier spectrum (STFS) of $\hat{o}_\tau = o_\tau g_\tau$. g_τ is a symmetric short-time analysis window, and M is the number of STFS bins. Let us suppose that the fundamental frequency (F0) at the t-th frame is already obtained as f_{0_t}. Then, we assume the following equations to decompose the power ρ_t into the powers of its periodic and aperiodic components, ρ_{p_t} and ρ_{a_t}:

$$\rho_t = \rho_{p_t} + \rho_{a_t} \qquad (2)$$

$$\rho_{p_t} = \eta \cdot \sum_{h=1}^{H_t} \left| O^{STFS}_{p_{t,[hf_{0_t}]}} \right|^2 \qquad (3)$$

$$\eta = \left(2\sum_{\tau=1}^{T} g_\tau^2\right) \Big/ \left(\sum_{\tau=1}^{T} g_\tau\right)^2 \qquad (4)$$

$$\rho_{a_t} = \frac{1}{M}\sum_{m=1}^{M}\left|O^{STFS}_{a_{t,m}}\right|^2 = \frac{1}{H_t}\sum_{h=1}^{H_t}\left|O^{STFS}_{a_{t,[hf_{0_t}]}}\right|^2 , \qquad (5)$$

where $O^{STFS}_{p_{t,m}}$ and $O^{STFS}_{a_{t,m}}$ are the STFSs of unknown periodic and aperiodic components in $\hat{o}_\tau$, η is a normalization constant that represents the ratio of the power of a sinusoidal component to its power spectrum, and $[hf_{0_t}]$ and H_t are the function that outputs an STFS bin index corresponding to the h-th harmonic frequency and the number of harmonics defined based on f_{0_t}, respectively. In the above decomposition, we estimate F0 by the autocorrelation method widely used for estimating F0 [9]. Note that Eq. (3) can be viewed as a kind of comb filtering, and that Eq. (5) represents the assumption that the power of the aperiodic components is widely distributed over the entire frequency range in a manner that is independent from the frequencies of the periodic components. After certain mathematical manipulations described in [3] based on the assumptions given by Eqs. (2) to (5), we can obtain the following:

$$\hat{\rho}_{p_t} = \eta \frac{\sum_{h=1}^{H_t}\left|O^{STFS}_{p_{t,[hf_{0_t}]}}\right|^2 - H_t\rho_t}{1-\eta H_t} \qquad (6)$$

$$\hat{\rho}_{a_t} = \frac{\rho_t - \eta\cdot\sum_{h=1}^{H_t}\left|O^{STFS}_{p_{t,[hf_{0_t}]}}\right|^2}{1-\eta H_t} , \qquad (7)$$

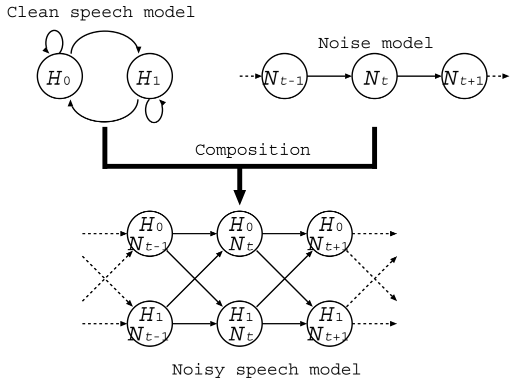

Fig. 1. Speech/non-speech state transition model with noise dynamics. The symbols H_0 and H_1 denote the non-speech and speech states, respectively. The symbol N_t denotes a noise state sequence.

where $\hat{\rho}_{p_t}$ and $\hat{\rho}_{a_t}$ denote estimated values of ρ_{p_t} and ρ_{a_t}. PAR can be calculated by using the ratios of the estimated $\hat{\rho}_{p_t}$ and $\hat{\rho}_{a_t}$.

4. VAD BASED ON STATISTICAL MODEL

4.1. Speech / non-speech state transition model

The proposed method discriminates between speech and non-speech periods based on the likelihood ratio test (LRT) with a statistical model.

As shown by the clean speech model in Fig. 1, the proposed method trains Gaussian mixture models (GMMs) of clean speech and silence in advance by using a clean speech corpus.

Next, we assume that noise has non-stationary characteristics, thus, the noise sequence is modeled by using a sequential state transition model as shown by the noise model in Fig. 1. With this method, we assume that the noise statistics are not known in advance. Thus, we estimate the parameters of the noise model sequentially by using a Kalman filter.

Finally, by composing speech and noise models, we can construct the speech / non-speech state transition model with noise dynamics as shown by the noisy speech model in Fig. 1. Namely, this model has state transition processes for both speech and noise. Speech has a discrete state transition process and noise has a sequential process. By using this model, we can construct VAD that is robust as regards a variety of speech and time varying noise.

4.2. Formulation of likelihood ratio calculation

This section describes the speech / non-speech discrimination method based on the state transition model shown in Fig. 1.

In the proposed method, $\mathbf{O}_{0:t}$, $\mathbf{N}_{0:t}$, and q_t denote the L-dimensional vector of the log Mel spectra of the observed signal and noise at the t-th short time frame, and the speech or the non-speech state at the t-th frame. When $\mathbf{O}_{0:t} = \{\mathbf{O}_0, \cdots, \mathbf{O}_t\}$ and $\mathbf{N}_{0:t} = \{\mathbf{N}_0, \cdots, \mathbf{N}_t\}$ are given, the state is decided with respect to the conditional probability $p\left(q_t|\mathbf{O}_{0:t}, \mathbf{N}_{0:t}\right)$ as follows:

$$\begin{aligned} p\left(q_t|\mathbf{O}_{0:t},\mathbf{N}_{0:t}\right) &= p\left(\mathbf{O}_{0:t},q_t,\mathbf{N}_{0:t}\right)/p\left(\mathbf{O}_{0:t},\mathbf{N}_{0:t}\right) \\ &\propto p\left(\mathbf{O}_{0:t},q_t,\mathbf{N}_{0:t}\right) \qquad (8) \end{aligned}$$

Here, we assume that q_t and $\mathbf{N}_t$ are mutually independent, thus, the recursive formula of $p\left(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t}\right)$ is given by

$$\begin{aligned} p\left(\mathbf{O}_{0:t},q_t,\mathbf{N}_{0:t}\right) &= \sum_{q_{t-1}} p\left(q_t|q_{t-1}\right)p\left(\mathbf{N}_t|\mathbf{N}_{t-1}\right)p\left(\mathbf{O}_t|q_t,\mathbf{N}_t\right) \\ &\quad \times p\left(\mathbf{O}_{0:t-1},q_{t-1},\mathbf{N}_{0:t-1}\right) . \qquad (9) \end{aligned}$$

By defining $p(q_t = H_j | q_{t-1} = H_i) = a_{i,j}$ (the state transition probability of speech), $p(\mathbf{O}_t | q_t = H_j, \mathbf{N}_t) = b_{j,\mathbf{N}_t}(\mathbf{O}_t)$ (the output probability), and $p(\mathbf{N}_t | \mathbf{N}_{t-1}) = c_{t,t-1}$ (the state transition probability of noise), $\alpha_{j,t} = p(\mathbf{O}_{0:t}, q_t = H_j, \mathbf{N}_{0:t})$ (the forward probability) is represented as the following equation from Eq. (9).

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

In Eq. (10), $c_{t,t-1}$ is set at 1, because we assume that the noise has a continuous state transition process. Thus, Eq. (10) is simplified as

$$\alpha_{j,t} = \sum_{i=0}^{1} (a_{i,j}\alpha_{i,t-1})\, b_{j,\mathbf{N}_t}(\mathbf{O}_t)\ . \tag{11}$$

In Eq. (11), when $t = 0$, the current frame is assumed to be a non-speech frame. Thus, the initial values $\alpha_{0,0} = 1$ and $\alpha_{1,0} = 0$ are given.

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

$$q_t = \begin{cases} H_0 & R_t < \text{Threshold} \\ H_1 & R_t \geq \text{Threshold} \end{cases} . \tag{12}$$

On the other hand, when we focus on the state transition noise model shown in Fig. 1, it is also given by the following equation. This equation is completely equivalent to a statistical representation of a Kalman filter [10].

$$p(\mathbf{O}_{0:t}, \mathbf{N}_{0:t}) = p(\mathbf{N}_t|\mathbf{N}_{t-1})\, p(\mathbf{O}_t|\mathbf{N}_t)\, p(\mathbf{O}_{0:t-1}, \mathbf{N}_{0:t-1}) \tag{13}$$

If the probability (state) variable q_t is added to Eq. (13), the statistical process is equivalent to Eq. (9). This means that Eq. (9) 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. The details of the Kalman filter-based noise state updating are provided in [5].

5. INTEGRATION OF VAD METHODS

This section describes the combination of PAR and SKF mentioned in sections 3 and 4. The combination is carried out by employing the adaptive weighted sum of likelihoods outputted from each method independently. Thus, the GMMs of clean speech and silence for PAR are trained in advance, and the likelihood of PAR $b_{j,PAR}(PAR_t)$ is given by the following equation.

$$\begin{aligned} & b_{j,PAR}(PAR_t) \\ = \ & \sum_{k=1}^{K} w_{PAR_{j,k}} \mathcal{N}\left(PAR_t; \mu_{PAR_{t,j,k}}, \sigma^2_{PAR_{t,j,k}}\right) \end{aligned} \tag{14}$$

When $b_{j,PAR}(PAR_t)$ is given, the likelihood of SKF $b_{j,\mathbf{N}_t}(\mathbf{O}_t)$ is added with weight γ_t as follows:

$$b_j(\mathbf{O}_t, PAR_t) = \gamma_t b_{j,\mathbf{N}_t}(\mathbf{O}_t) + (1-\gamma_t) b_{j,PAR}(PAR_t)\ , \tag{15}$$

where $0 \leq \gamma_t \leq 1$. The forward probabilities for the LRT are calculated by using combined likelihood $b_j(\mathbf{O}_t, PAR_t)$ as follows:

$$\alpha_{j,t} = \sum_{i=0}^{1} (a_{i,j}\alpha_{i,t-1})\, b_j(\mathbf{O}_t, PAR_t) \tag{16}$$

When $\gamma_t = 0$, the discrimination is carried out by PAR alone, and for $\gamma_t = 1$, the discrimination is carried out by SKF alone.

The adaptive frame by frame decision of the weight γ_t is given by the following method.

First, $b_{j,\mathbf{N}_t}(\mathbf{O}_t)$ and $b_{j,PAR}(PAR_t)$ are normalized as the total likelihood of non-speech state ($j = 0$) and speech state ($j = 1$) equal to 1. Next, the absolute differences of the likelihoods of the non-speech state and the speech state $D_{SKF,t}$ and $D_{PAR,t}$ are calculated by the following equations.

$$D_{SKF,t} = |b_{0,\mathbf{N}_t}(\mathbf{O}_t) - b_{1,\mathbf{N}_t}(\mathbf{O}_t)| \tag{17}$$

$$D_{PAR,t} = |b_{0,PAR}(PAR_t) - b_{1,PAR}(PAR_t)| \tag{18}$$

When $D_{SKF,t}$ or $D_{PAR,t}$ has a large value, it shows that the decision confidence is high regarding whether the current frame belongs to a non-speech state or a speech state. When the value are low, it shows that the decision confidence is low. The weight γ_t given by using $D_{SKF,t}$ and $D_{PAR,t}$ as follows:

$$\gamma_t = D_{SKF,t} / (D_{SKF,t} + D_{PAR,t}) \tag{19}$$

Equation (19) shows that the weight of the SKF increases or decreases according to the confidence of the SKF. Thus, by using this weight, we can select the most suitable method for the current frame, frame by frame.

6. EXPERIMENTS

6.1. Experimental setup

The proposed method was evaluated by using the CENSREC-1-C database [6]. 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 (Rest.) and a street (St.)) with two different sound pressure levels (avg. 60 dBA: high SNR (Hi.) and avg. 70 dBA: low SNR (Lo.)). The data were originally recorded at a sampling rate of 48 kHz (with 16 bit quantization), and were 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 between each utterance in each noisy environment and for each SNR condition. The correct segment labels were tagged manually.

The feature parameters for the PAR-based VAD and SKF-based VAD were 1st order PAR and 24th order log-Mel spectra, respectively, 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 for PAR-based VAD and SKF-based VAD by using clean speech data for the HMM training of CENSREC-1 (AURORA-2J) [11]. Each GMM has 32 Gaussian distributions. The training data consisted of 8,440 utterances spoken by 110 speakers. The state transition probabilities of the clean speech model were set at $a_{i,j} = \{0.90, 0.10, 0.45, 0.55\}$.

6.2. Experimental results of VAD

In the evaluation, we compare the VAD performance of the proposed method with the CENSREC-1-C baseline, Sohn's method [4], PAR alone ($\gamma_t = 0$), and SKF alone ($\gamma_t = 1$). The baseline VAD technique of CENSREC-1-C is energy-based VAD with adaptive thresholding.

The evaluation criteria are the utterance correct rate and utterance accuracy rate as shown by Eqs. (20) and (21).

$$Corr = N_c/N \times 100\ [\%] \tag{20}$$

$$Acc = (N_c - N_f)/N \times 100\ [\%]\ , \tag{21}$$

Table 2. VAD results (%)

	$Corr(\%)$					$Acc(\%)$				
	Rest. Hi.	Rest. Lo.	St. Hi.	St. Lo.	Ave.	Rest. Hi.	Rest. Lo.	St. Hi.	St. Lo.	Ave.
Baseline	74.20	56.52	39.42	41.45	52.90	21.45	-43.48	-15.65	-33.91	-17.90
Sohn	72.75	57.10	97.39	78.55	76.45	45.51	-6.38	94.49	57.39	47.75
PAR	70.72	57.10	87.25	80.58	73.91	24.35	-6.67	64.35	54.49	34.13
SKF	89.57	66.96	**100.00**	**97.97**	88.63	68.41	12.46	97.68	93.62	68.04
PAR+SKF	**93.04**	**70.72**	**100.00**	**97.97**	**90.43**	**72.75**	**19.71**	**99.13**	**94.78**	**71.60**

Table 3. Speech recognition results after VAD (%)

	Word accuracy (%)					Error reduction rate from w/o VAD (%)				
	Rest. Hi.	Rest. Lo.	St. Hi.	St. Lo.	Ave.	Rest. Hi.	Rest. Lo.	St. Hi.	St. Lo.	Ave.
w/o VAD	45.17	1.28	34.43	25.23	26.53	0.00	0.00	0.00	0.00	0.00
Baseline	44.16	18.12	29.96	21.62	28.47	-1.84	17.06	-6.82	-4.83	2.64
Ideal VAD	*52.67*	*29.17*	*41.25*	*29.50*	*38.15*	*13.68*	*28.25*	*10.40*	*5.71*	*15.82*
Sohn	37.45	-3.81	33.41	29.58	24.16	-13.83	-6.22	-1.46	5.55	-3.31
PAR	39.76	8.89	39.16	24.08	27.97	-9.87	7.71	7.21	-1.54	1.97
SKF	43.75	12.50	46.99	33.15	34.10	-2.59	11.37	19.16	10.59	10.30
PAR+SKF	**46.85**	**18.67**	**47.27**	**33.52**	**36.58**	**3.06**	**17.62**	**19.58**	**11.09**	**13.68**

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 the results of an utterance-level evaluation. As seen in the table, the proposed method "PAR+SKF" significantly improves both $Corr$ and Acc compared with the baseline. In particular, the average improvement in Acc when compared with the baseline was approximately 83 %. In addition, the proposed method improves both $Corr$ and Acc compared with PAR and SKF. This means that the proposed likelihood combination works effectively.

6.3. 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) [12] 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 consisted 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 for CENSREC-1 is described in [11].

Table 3 shows the speech recognition results in terms of word accuracy. In the table, "w/o VAD" and "Ideal VAD" represent speech recognition results without VAD and with VAD using hand labeled utterance boundaries, respectively. The table shows that the proposed method "PAR+SKF" improves speech recognition accuracy. As regards 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 errors, 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.

7. CONCLUSION

This paper presented a noise robust VAD technique based on the integration of multiple speech features and a signal decision scheme. The proposed method combined a speech periodic to aperiodic component ratio and a switching Kalman filter with adaptive weighting. The evaluation results show that our proposed method improves VAD accuracy compared with the periodic to aperiodic component ratio or the switching Kalman filter alone. In addition the proposed method also improves speech recognition accuracy. In the future, we are planning to investigate the optimal threshold decision.

8. ACKNOWLEDGEMENTS

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

9. REFERENCES

[1] Rabiner, L. R. *et al.*, "An algorithm for determining the end-points 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.*, "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio," *Proc. of SAPA '06*, pp.65–70, Sept. 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] Fujimoto, M. *et al.*, "Noise robust voice activity detection based on switching Kalman filter," *Proc. of Intespeech '07*, pp. 2933-2936, Aug. 2007.

[6] CENSREC-1-C Web site, http://sp.shinshu-u.ac.jp/CENSREC/en/CENSREC/CENSREC-1-C/

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

[8] 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.

[9] Hess, W., "Pitch Determination of Speech Signals," Springer–Verlag, New York, 1983.

[10] 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.

[11] Nakamura, S. *et al.*, "Data collection and evaluation of AURORA2-J Japanese corpus," *Proc. of ASRU '03*, pp. 619–623, Dec. 2003.

[12] HTK Web site, http://htk.eng.cam.ac.uk/