SPEAKER INDEXING AND SPEECH ENHANCEMENT IN REAL MEETINGS / CONVERSATIONS

Shoko Araki, Masakiyo Fujimoto, Kentaro Ishizuka, Hiroshi Sawada, Shoji Makino

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

ABSTRACT

This paper presents a speaker indexing method that uses a small number of microphones to estimate who spoke when. Our proposed speaker indexing is realized by using a noise robust voice activity detector (VAD), a GCC-PHAT based direction of arrival (DOA) estimator, and a DOA classifier. Using the estimated speaker indexing information, we can also enhance the utterances of each speaker with a maximum signal-to-noise-ratio (MaxSNR) beamformer. This paper applies our system to real recorded meetings / conversations recorded in a room with a reverberation time of 350 ms, and evaluates the performance by a standard measure: the diarization error rate (DER). Even for the real conversations, which have many speaker turn-takings and overlaps, the speaker error time was very small with our proposed system. We are planning to demonstrate a real-time speaker indexing system at ICASSP2008.

***Index Terms*—** Speaker indexing, diarization, voice activity detector, maximum SNR beamformer

1. INTRODUCTION

Meeting recognition has been studied [1–5] and it has been pointed out that speaker indexing (sometimes called "diarization"), i.e., estimating who spoke when, is an important topic. The speaker indexing information should be useful for such applications as speech recognition during minute taking and speech enhancement.

Let us formulate the task. Suppose that $N \geq 2$ speech sources $\mathrm{s}_1, \dots, \mathrm{s}_N$ are convolutively mixed and observed at M microphones,

$$\mathrm{x}_j(t) = \sum_{k=1}^{N} \sum_l \mathrm{h}_{jk}(l)\, \mathrm{s}_k(t-l) + \mathrm{n}_j(t),\ j=1,\dots,M, \quad (1)$$

where $\mathrm{h}_{jk}(l)$ represents the impulse response from source k to microphone j, and $\mathrm{n}_j(t)$ is the observed stationary background noise at microphone j. The speech $\mathrm{s}_k(t)$ are intermittent signals. In this paper we assume that the speakers do not change their seats during one meeting / conversation. Our goals are (i) to give speaker indices to each time point t, and (ii) to enhance each speaker utterance, without knowing the number of speakers N, the speech sources s_k or the mixing process h_{jk}.

Recently, some papers (e.g., [1] and related papers) employ several microphones and utilize the time-difference of arrival information between microphones to improve speaker indexing. However, most previous work uses an undesigned microphone array. However, in the meeting, we can employ a small and precisely arranged microphone array. By using such an array, we can utilize the speaker position information more accurately, and easily relate the estimated speaker position to the meeting's seating arrangement. Such an array is also portable, which is important for a minuting system.

As regards the meeting recognition task, we have proposed a speech enhancement method for a meeting situation [6]. The paper employed a maximum signal-to-noise-ratio (MaxSNR) beamformer, which requires "target speaker speaking period" and "target silent (interferences-and-noise) period" information. That is, our enhancement method includes a speaker indexing system.

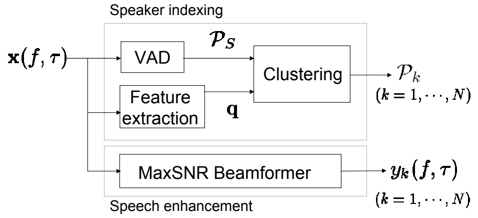

Fig. 1. Flow for proposed method.

In this paper, first, we detail our speaker indexing method. Our speaker indexing system consists of a noise robust voice activity detector (VAD), a GCC-PHAT based direction of arrival (DOA) estimator, and a DOA classifier. That is, our system estimates the speaker indices by relying on the speaker seat locations. Our system is characterized by a VAD that is robust with respect to all types of noise (i.e., stationary, non-stationary, and burst noise). In addition, we explain a speech enhancement method with the MaxSNR beamformer. Although the authors of [2] employ a beamformer as a frontend for the speaker indexing, this paper proposes to utilize the speaker indexing result for designing the beamformer coefficients.

Then, we report the speaker indexing and enhancement performance of our proposed system. We also evaluate the noise robust VAD. We utilized real recordings of meetings / conversations in a room whose reverberation time was 350 ms. We obtained encouraging results even for the recorded conversations, which usually have more speaker turn-takings and overlaps than meetings.

2. PROPOSED METHOD

Figure 1 shows the system flow of our proposed method. This section explains each step in Fig. 1 closely.

Our system works in the time-frequency domain. That is, it utilizes the time-frequency representation $x_j(f,\tau)$ of the observations $\mathrm{x}_j(t)$ (1), which is obtained by a short-time Fourier transform (STFT). Here f is a frequency and τ is a time-frame index.

2.1. Speaker indexing

2.1.1. Voice activity detector (VAD)

First, we detect automatically the periods of target human speeches from a continuously observed signal by using VAD.

A block diagram of the VAD employed in this paper is shown in Fig. 2. In the figure, the VAD is constructed by using two stream

1-4244-1484-9/08/$25.00 ©2008 IEEE

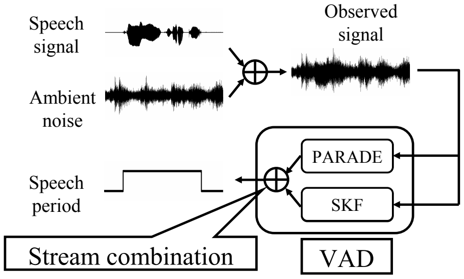

Fig. 2. Block diagram of VAD. PARADE: a Periodic to Aperiodic component RAtio-based DEtection, SKF: a switching Kalman filter.

speech / non-speech discriminators, i.e., periodic to aperiodic component ratio-based detection (PARADE) [7] and a switching Kalman filter (SKF)-based approach [8]. Each stream outputs the likelihood of speech / non-speech discrimination frame by frame, thus the speech period $\mathcal{P}_S$ is decided by using the adaptively weighted sum of each likelihood [9].

PARADE is robust for burst noise and the SKF is robust for stationary and non-stationary noises. Therefore, by integrating them, we can obtain a VAD that is robust for all types of noises, i.e., stationary noise, non-stationary noise, and burst noise.

Here, the SKF constructs a clean speech / silence state transition model (GMM: Gaussian mixture model) in advance, and sequentially updates the noise model without *a priori* knowledge of ambient noise by using a Kalman filter when a signal is observed. After the noise model updating, 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 the clean speech state or the silence state and updated noise model. With the method, the Kalman filter for noise updating is formulated by using PDF parameters of the clean speech state or the silence state. Consequently, two types of estimation (updating) results are given for the noise model by the selection of the clean speech state or the silence state. This means that the state-space representation of the Kalman filter depends on the state selection, thus the method has the characteristic of a switching Kalman filter. By using the adapted model, we can construct the VAD that is robust to a variety of speech and time varying noise.

In this paper, the VAD results are given by binary labeling, i.e., the non-speech and the speech frames are labeled as 0 and 1, respectively. As we employ an array of multiple microphones, first we apply the VAD to each channel independently. Then, each outputted binary label is unified with a frame by frame logical sum operation. The speech period $\mathcal{P}_S$ is determined by $\mathcal{P}_S = \{\tau | \text{frames labeled as } 1\}$.

2.1.2. Speaker indexing

Then, we partition the speech period $\mathcal{P}_S$ into each speaker period $\mathcal{P}_k$ ($k = 1, \cdots, N$). The estimated speaker periods $\mathcal{P}_k$ gives us the speaker indexing result.

Feature extraction: In this paper, we classify the direction of arrival (DOA) $\mathbf{q}(\tau)$ in the speech period $\mathcal{P}_S$. To estimate the DOA, first we estimate the time differences of arrival (TDOA) $q'_{jj'}(\tau)$ for all microphone pairs j and j' by using the generalized cross correlation method with the phase transform (GCC-PHAT) [10]

$$q'_{jj'}(\tau) = \text{argmax}_{q'} \sum_f \frac{x_j(f,\tau)x^*_{j'}(f,\tau)}{|x_j(f,\tau)x^*_{j'}(f,\tau)|} e^{\jmath 2\pi f q'}. \qquad (2)$$

We can use a TDOA (column) vector $\mathbf{q}'(\tau)$, which consists of the $q'_{jj'}(\tau)$ of all the microphone pairs, however, in this paper we use the DOA vector $\mathbf{q}(\tau)$ as a feature for simplicity of implementation.

```
for $\tau = 1$ : end {  /* for all frame $\tau$, do online clustering */
  if exist(centroid) & $\tau \in \mathcal{P}_S$ {
    if min $||\mathbf{c}_k - \mathbf{q}(\tau)|| < th1$ {  /* close enough to existing centroid */
      then $k \leftarrow$ argmin$_k ||\mathbf{c}_k - \mathbf{q}(\tau)||$  /* find nearest centroid $\mathbf{c}_k$ */
           $C_k \leftarrow C_k \cup \mathbf{q}(\tau)$  /* add cluster member to cluster $C_k$*/
           $\mathbf{c}_k \leftarrow \mathbf{c}_k + \mu\mathbf{q}(\tau)$  /* update centroid $\mathbf{c}_k$*/
    }
  }
  if mod($\tau/F1$)==0 {  /* check new centroid(s) for every $F1$ frames */
    if (the number of $\tau \in \mathcal{P}_S$ in bin$_i$) $> th2$ in the last $F2$ frames
      & min $||\mathbf{c}_k - \mathbf{b}_i|| > th3$ { /* if recent features concentrate in a DOA
                                             & it is far enough to existing centroids */
      then create new centroid $\mathbf{c}_{K+1} = \mathbf{b}_i$ /* $K$: # of existing clusters */
    }
  }
}
```

Fig. 3. Pseudo code of online clustering. bin$_i$ is the pre-defined feature range, $\mathbf{b}_i$ is a representing vector for bin$_i$. In our implementation, bin$_i$ = $[10°(i-1), 10°i)$ is defined in azimuth θ (see section 2.1.2), and $\mathbf{b}_i$ is the center value of bin$_i$. The thresholds in our paper are $th1 = 15°$ and $th3 = 30°$ in azimuth θ, $F1 = 20$, $F2 = 500$, and $th2 = 16$ frames.

The DOA vector $\mathbf{q}(\tau)$ is calculated by the TDOA information $\mathbf{q}'(\tau)$ and the given microphone coordinate information $\mathbf{D}$ [11]:

$$\mathbf{q}(\tau) = c\mathbf{D}^{-}\mathbf{q}'(\tau). \qquad (3)$$

where c is the propagation velocity of the signals and $^{-}$ denotes the Moore-Penrose pseudo-inverse. When the source azimuth is $\theta(\tau)$ and the elevation is $\phi(\tau)$, the DOA vector can be written as $\mathbf{q}(\tau) = [\cos\theta(\tau)\cos\phi(\tau), \sin\theta(\tau)\cos\phi(\tau), \sin\phi(\tau)]^T$.

Because we employed the GCC-PHAT, the feature $\mathbf{q}(\tau)$ is estimated frame-wise (not time-frequency slot-wise).

Clustering: To divide $\mathcal{P}_S$ into individual speaker periods $\mathcal{P}_k$, we then perform clustering for the extracted features $\mathbf{q}(\tau)$ of all time-frames $\tau \in \mathcal{P}_S$. In order to apply our method even when the number of speakers N is unknown, we employ an online clustering algorithm (leader-follower clustering) [12]. The pseudo code is provided in Fig. 3, where a new centroid is generated when a new speaker appears in a recording.

Each speaker period $\mathcal{P}_k$ is determined by

$$\tau \in \mathcal{P}_k \quad \text{if} \quad \mathbf{q}(\tau) \in C_k, \qquad (4)$$

where C_k is the k-th cluster. This $\mathcal{P}_k$ is the speaker indexing result for the real-time display (see Fig. 5). For the speaker indexing evaluation in Section 3, we removed the short fragments and short pauses by smoothing $\mathcal{P}_k$, and provide temporal information about the speech-onset and speech-offset for each speaker.

2.2. Speech enhancement

For speech enhancement, this paper employs a MaxSNR beamformer as in our previous paper [6]. The design criterion for the beamformer $\mathbf{w}_k(f)$ is to maximize the ratio $\lambda(f)$ of the output power between the target-speaker period $\mathcal{P}_k$ and the interference-and-noise-only period $\bar{\mathcal{P}}_k = \mathcal{P} - \mathcal{P}_k$. That is, the MaxSNR beamformer is one of the applications of the speaker indexing. The beamformer coefficients

Table 1. Conversation recordings. Each recording duration was five minutes.

Evaluation data ID	#Speaker	Overlap [%]	#Turn-taking	#Utterance	Noise sources
presentation rehearsal 1	4	1.4	40	119	PC1,2, projector, laughing voice of other speakers
presentation rehearsal 2 (Q&A)	3	6.0	75	145	
conversation	3	34.8	243	278	PC2
discussion	3	10.8	126	172	PC2, paper noise
crossword puzzle 1	4	18.6	149	185	PC2
crossword puzzle 2	4	13.0	183	218	PC2

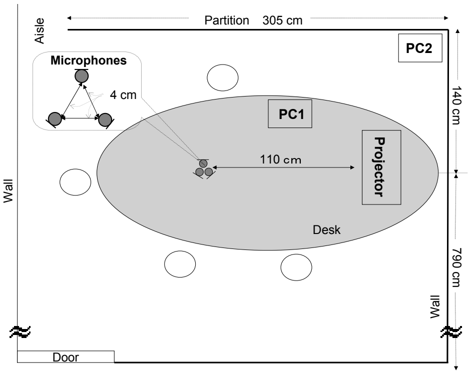

Fig. 4. Room setup. Small ellipses illustrate example speaker places. The reverberation time was around 350 ms.

$\mathbf{w}_k(f)$ is obtained by the eigenvector $\mathbf{e}_1(f)$, which corresponds to the largest eigenvalue of the generalized eigenvalue problem [6]

$$\mathbf{R}_{\mathbf{T}}^k(f)\mathbf{w}_k(f) = \lambda(f)\mathbf{R}_{\mathbf{I}}^k(f)\mathbf{w}_k(f), \tag{5}$$

where $\mathbf{R}_{\mathbf{T}}^k(f) = \frac{1}{|\mathcal{P}_k|}\sum_{\tau\in\mathcal{P}_k}\mathbf{x}(f,\tau)\mathbf{x}^H(f,\tau)$, $\mathbf{R}_{\mathbf{I}}^k(f) = \frac{1}{|\bar{\mathcal{P}}_k|}\sum_{\tau\in\bar{\mathcal{P}}_k}\mathbf{x}(f,\tau)\mathbf{x}^H(f,\tau), \mathbf{x}(f,\tau) = [x_1(f,\tau),\ldots,x_M(f,\tau)]^T$, and $|\mathcal{P}|$ denotes the number of elements of $|\mathcal{P}|$.

The enhanced speech for the k-th speaker is obtained by

$$y_k(f,\tau) = \mathbf{w}_k^H(f)\mathbf{x}(f,\tau)\,. \tag{6}$$

3. SYSTEM EVALUATIONS

3.1. Setup

Experiments were performed in the room shown in Fig. 4 whose reverberation time was around 350 ms. We recorded some conversations by three or four speakers in the room. The duration of each unit of recorded data was five minutes. The distance between the microphone array and the speakers was around 1 m. Personal computers (PC1, 2) and a projector in Fig. 4 could be the noise sources.

Because our recordings were conversations, they contain more speaker turn-takings and speaker overlaps than usual meeting recordings. Table 1 summarizes the conversation situations. It can be seen that our data contains many speaker turn-takings and overlaps, which make speaker indexing difficult and require speech enhancement. Reference VAD and speaker indexing labels were generated by employing a hand-labeled transcription, which includes temporal information about speech-onsets and speech-offsets.

The sampling rate was 16 kHz, the frame size for STFT was 64 ms, and the frame shift was 32 ms.

3.2. VAD results

In the evaluation, we compare the VAD performance of the proposed method with that of Sohn's method [13], which is a widely used statistical model-based VAD technique.

Table 2. Experimental results of VAD [%]

Evaluation data ID	Sohn				Proposed			
	FAR	FRR	Ave.	DER	FAR	FRR	Ave.	DER
presentation rehearsal 1	41.8	14.3	28.1	**20.0**	12.8	24.3	**18.6**	22.2
presentation rehearsal 2	24.3	28.5	26.4	34.2	22.5	19.7	**21.1**	**22.0**
conversation	33.2	44.1	38.7	56.6	47.9	21.7	**34.8**	**32.5**
discussion	56.4	17.0	36.7	45.7	14.8	22.4	**18.6**	**23.0**
crossword puzzle 1	12.6	37.1	24.9	45.3	16.4	13.7	**15.1**	**19.4**
crossword puzzle 2	32.0	22.8	27.4	37.3	26.8	13.6	**20.2**	**17.8**

The feature parameters for the PARADE-based VAD and SKF-based VAD in the proposed method were the 1st order periodic to aperiodic component ratio and the 24th order log-Mel spectra, respectively. These parameters were extracted by using a Hamming window with a 64 ms frame length and a 32 ms frame shift length. We trained the silence and clean speech GMMs for PARADE-based VAD and SKF-based by using phonetically balanced Japanese sentences. The training data consisted of 5,050 utterances spoken by 101 speakers. Each GMM had 32 Gaussian distributions.

The evaluation criteria are the false acceptance rate (FAR) and the false rejection rate (FRR):

$$\mathrm{FAR} = N_{FA}/N_{ns} \times 100\ [\%],\quad \mathrm{FRR} = N_{FR}/N_s \times 100\ [\%]\,,$$

where N_{ns}, N_s, N_{FA}, and N_{FR} are the total number of non-speech frames, the total number of speech frames, the number of non-speech frames detected as speech frames, and the number of speech frames detected as non-speech frames, respectively. We also evaluated the diarization error rate (DER) for VAD [3] (in [3], it is defined as DER for speech activity detection (SAD))

$$\mathrm{DER} = \frac{\text{Wrongly estimated speech period length}}{\text{Entire speech period length}} \times 100[\%].$$

Table 2 shows the VAD results. As seen in the table, the proposed method significantly improves the average FAR and FRR rates and the DER compared with Sohn's method. With the proposed method, the factors that contributed to the improvement were the updating of Kalman filter-based noise model and the adaptive stream combination. In particular, the expansion of the applicable noise environment based on the multi stream combination is the most crucial factor for VAD in real environments.

3.3. Speaker indexing results

We evaluated the diarization error rate (DER),

$$\mathrm{DER} = \frac{\text{Wrongly estimated speaker time length}}{\text{Entire speaker time length}} \times 100[\%],$$

which is established by NIST [3]. The diarization error includes the missed speaker time (MST), the false alarm speaker time (FAT), and the speaker error time (SET). We evaluated the DER after smoothing the speaker indexing result $\mathcal{P}_k$ as mentioned in Section 2.1.2. If the estimated number of speakers outnumbers the true number of speakers, such *ghost* speaker periods were regarded as the SET.

Table 3. Experimental results of speaker indexing [%]

Evaluation data ID	With Sohn's VAD				With proposed VAD			
	DER	MST	FAT	SET	DER	MST	FAT	SET
presentation rehearsal 1	27.2	21.6	4.0	1.7	**23.9**	20.7	3.1	0.1
presentation rehearsal 2	35.0	27.6	4.5	2.9	**31.2**	23.1	5.6	2.5
conversation	61.3	30.6	13.7	17.1	**38.7**	19.0	14.5	5.3
discussion	45.0	24.5	17.7	2.8	**34.8**	25.9	6.9	2.0
crossword puzzle 1	45.0	30.7	7.7	6.6	**36.9**	18.1	13.6	5.2
crossword puzzle 2	47.6	34.3	9.2	4.1	**32.7**	21.3	6.8	4.6

Table 3 summarizes the results. Please note that our recordings contained many speaker turn-takings and overlaps. With our noise robust VAD, we obtained better performance in the DER than with Sohn's VAD. In our implementation, we had small the SET rate. This means that our online clustering stage worked successfully. The DER results were affected by the high MST rate. The high MST rate was because we employed frame-wise DOA as the feature, which disregards the speaker overlap in one frame. The MST may be improved by using time-frequency-wise DOA information (e.g.,[11]).

3.4. Speech enhancement results

We also evaluated the speech enhancement performance provided by the MaxSNR beamformer, where the target-speaker period $\mathcal{P}_k$ and the interference-and-noise-only period $\overline{\mathcal{P}_k}$ were estimated by the speaker indexing. Here, $\mathcal{P}_k$ was not smoothed. For a quantitative evaluation, we utilized simulated conversation data that was made by following eq. (1) with impulse responses $\mathrm{h}_{jk}(l)$ and PC and projector noise $\mathrm{n}_j(t)$ obtained in the room (Fig. 4), and English speech sources $\mathrm{s}_k(t)$ sampled at 16 kHz. Each data segment lasted 90 s and the conversation involved three speakers. We tried six speaker combinations. The performance measures were the signal-to-interference plus noise ratio (SINR) and the signal to distortion ratio (SDR) [6].

Table 4 shows the speech enhancement results. We obtained high performance, although we did not smooth the speaker indexing result $\mathcal{P}_k$. Moreover, the performance does not depend on the VAD methods. This suggests that the MaxSNR beamformer does not necessarily require the low MST and FAT rates, and that it just needs the accurate speaker frame estimation (i.e., small SET rates).

We also confirmed informally that we can obtain good enhancement performance with our implemented system for the real recorded conversation. Even when we have more speakers than microphones, we can still expect that there are fewer speakers in a short data block (e.g, 5 seconds) in a real conversation. Therefore, it was still possible to employ the MaxSNR beamformer.

3.5. System

In our implementation, the VAD component was written in C and the speaker indexing parts and display drawing (see Fig. 5) were realized by MATLAB6. 5. In the display (Fig. 5), microphone-1 observation (Fig. 5(A)) and the speaker indexing results (Fig. 5(B), (C)) are drawn. Our speaker indexing system can work in real-time on a personal computer (AMD Athlon64, 2.4GHz). The real-time factor of the system, which is defined as (the total processing time)/(the total recording signal duration), was around 0.6.

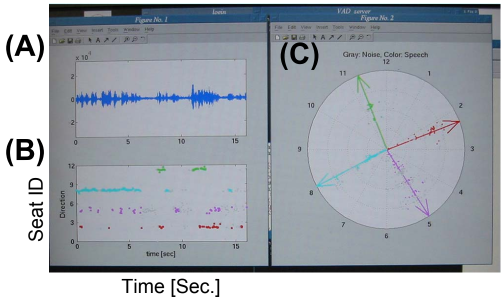

Fig. 5. Result image. (A) recording at microphone 1, (B) speaker indexing result, (C) speaker positions with respect to the microphone array (the center of the circle indicates the array position).

Table 4. Speaker enhancement results in SINR and SDR [dB]. The averaged input SINR was -3.1 [dB].

With Sohn's VAD		With proposed VAD	
SINR	SDR	SINR	SDR
11.8	16.6	11.9	16.2

4. CONCLUSION

We reported an evaluation of our proposed speaker indexing / enhancement system. Our proposed speaker indexing system, which consists of noise robust VAD, a GCC-PHAT based DOA estimator and a DOA classifier, works well even for conversation recordings. We also reported that speech enhancement with the MaxSNR beamformer can be realized by using roughly estimated speaker indexing results. Our future work will include improving the speaker indexing performance by employing time-frequency-wise DOA information.

5. REFERENCES

[1] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multi-microphone meetings using only between-channel differences," in *Proc. of MLMI'06 (LNCS 4299)*. Sept. 2006, pp. 257–264, Springer.

[2] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, pp. 2011–2022, Sept. 2007.

[3] `http://www.nist.gov/speech/test_beds/mr_proj/`

[4] D. Ellis and J. Liu, "Speaker turn segmentation based on between-channel differences," in *Proc. of NIST Meeting Recognition Workshop*, 2004, pp. 112–117.

[5] C. Busso, P. Panayiotis, G. Georgiou, and S. Narayanan, "Real-time monitoring of participants' interaction in a meeting using audio-visual sensors," in *Proc. of ICASSP'07*, Apr. 2007, vol. II, pp. 685–688.

[6] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," in *Proc. of ICASSP'07*, Apr. 2007, vol. I, pp. 41–45.

[7] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, "Noise robust front-end processing with voice activity detection based on periodic to aperiodic component ratio," in *Proc. of Interspeech '07*, 2007, pp. 230–233.

[8] M. Fujimoto and K. Ishizuka, "Noise robust voice activity detection based on switching Kalman filter," in *Proc. of Interspeech '07*, Aug. 2007, pp. 2933–2936.

[9] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A voice activity detection based on adaptive integration of multiple speech feature and signal decision scheme," in *Proc. of ICASSP '08*, Mar. 2008, (in submitting).

[10] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[11] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proc. of ICASSP'06*, May 2006, vol. 5, pp. 33–36.

[12] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley Interscience, 2nd edition, 2000.

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