A DOA BASED SPEAKER DIARIZATION SYSTEM FOR REAL MEETINGS

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

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
Email: shoko@cslab.kecl.ntt.co.jp

ABSTRACT

This paper presents a speaker diarization system that estimates who spoke when in a meeting. Our proposed system is realized by using a noise robust voice activity detector (VAD), a direction of arrival (DOA) estimator, and a DOA classifier. Our previous system utilized the generalized cross correlation method with the phase transform (GCC-PHAT) approach for the DOA estimation. Because the GCC-PHAT can estimate just one DOA per frame, it was difficult to handle speaker overlaps. This paper tries to deal with this issue by employing a DOA at each time-frequency slot (TFDOA), and reports how it improves diarization performance for real meetings / conversations recorded in a room with a reverberation time of 350 ms.

Index Terms— diarization, voice activity detector, direction of arrival

1. INTRODUCTION

Meeting recognition has been studied [1, 2, 3, 4, 5] and it has been pointed out that speaker diarization, i.e., estimating who spoke when, is an important topic. Speaker diarization 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 $\mathrm{s}_1, \ldots, \mathrm{s}_N$ are convolutively mixed and observed at M microphones,

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

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. Speech $\mathrm{s}_k(t)$ consists of intermittent signals. In this paper we assume that the speakers do not change their seats during one meeting / conversation. Our goal is to estimate "who spoke when" at each time point t, without knowing the number of speakers N, the speech sources s_k or the mixing process h_{jk}. We work in the time-frequency domain. That is, we utilize the time-frequency representation $x_j(f, \tau)$ of the observations $\mathrm{x}_j(t)$ (1), which we can obtain by a short-time Fourier transform (STFT). Here f is a frequency and τ is a time-frame index.

Recently, we proposed a diarization system based on a noise robust voice activity detector (VAD), a generalized cross correlation method with a phase transform (GCC-PHAT) based direction of arrival (DOA) estimator, and a DOA classifier [6]. That is, our diarization system relies on the speaker seat locations. The system worked very well for real meetings / conversations, even when the system did not know the number of speakers. The system is simple and portable: it can run in real-time and it uses a small number of microphones. However, the system has a problem: it estimates just one DOA per frame, because the previous approach employed the GCC-PHAT. This means that only one speaker is detected even if multiple speakers spoke in a frame. This causes a lot of missed speaker time (false rejection in VAD terminology), and degrades the performance. Some other approaches by the ICSI project [1] and the CHIL project [2] also employ the GCC-PHAT technique, and therefore they also experience the same problem as our approach.

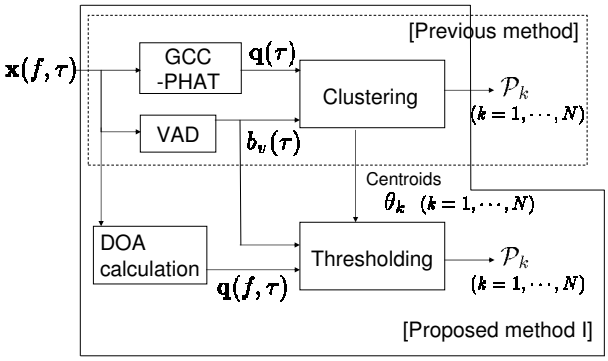

Fig. 1. Block diagram of previous method (framed by dashed line) and proposed method (Method I).

In this paper, we try to improve the performance of our diarization system by using the DOA at each time-frequency slot (TFDOA). The authors of [7] have successfully employed the TFDOA in a speech enhancement scenario. In addition to the TFDOA, we try to utilize the amplitude information of observations at each time-frequency slot, and a probabilistic representation of the VAD results. The experimental results obtained for real recordings of meetings / conversations show that such refinements improve the diarization performance.

2. PREVIOUS METHOD

This section describes our previous approach [6], and points out its problems.

2.1. Method

Figure 1 shows the system flow of our previous method [6]. With the method, first the speech periods $\mathcal{P}_S = \{\tau | b_v(\tau) = 1\}$ are estimated from a continuously observed signal by using a VAD, then the speech periods $\mathcal{P}_S$ are determined for each speaker period $\mathcal{P}_k$

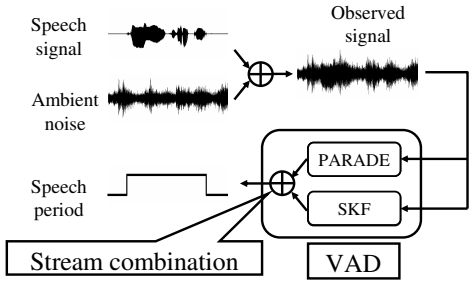

Fig. 2. Block diagram of VAD. PARADE: a Periodic to Aperiodic component RAtio-based DEtection, SKF: a switching Kalman filter.

($k = 1, \cdots, N$) by classifying the direction of arrival (DOA) information. That is, our previous method for estimating "who spoke when" is based on the speaker positions.

As the VAD, we utilized the MUlti Stream Combination of Likelihood Evolution of VAD (MUSCLE-VAD) [8], which integrates multiple speech features and a signal decision scheme. A block diagram of the VAD is shown in Fig. 2. The VAD is constructed by using two stream speech / non-speech discriminators, i.e., periodic to aperiodic component ratio-based detection (PARADE) [9] and a switching Kalman filter (SKF)-based approach [10]. 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 [8]. 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.

The VAD results $b_v(\tau)$ are given by binary labeling, i.e., the speech and non-speech frames are labeled 1 and 0, respectively. As we employ 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 DOA estimation was performed using the GCC-PHAT [11]

$$q'_{jj'}(\tau) = \text{argmax}_{q'} \sum_f \frac{x_j(f,\tau)x^*_{j'}(f,\tau)}{|x_j(f,\tau)x^*_{j'}(f,\tau)|} e^{j2\pi f q'}, \quad (2)$$

where $q'_{jj'}(\tau)$ is the time differences of arrival (TDOA) in between a microphone pair $j - j'$. The DOA vector $\mathbf{q}(\tau)$ is calculated by the TDOA information $\mathbf{q}'(\tau)$, which consists of the $q'_{jj'}(\tau)$ of all the microphone pairs, and the given microphone coordinate information $\mathbf{D}$ [12]:

$$\mathbf{q}(\tau) = c\mathbf{D}^+\mathbf{q}'(\tau) \quad (3)$$

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

$$\mathbf{q}(\tau) = [\cos\theta(\tau)\cos\phi(\tau), \sin\theta(\tau)\cos\phi(\tau), \sin\phi(\tau)]^T.$$

Here, we employ only the azimuth $\theta(\tau)$ for simplicity.

The individual speaker periods $\mathcal{P}_k$ are determined by clustering the estimated DOA $\theta(\tau)$ at all speech frames $\tau \in \mathcal{P}_S$:

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

where C_k is the k-th cluster. This $\mathcal{P}_k$ is the speaker diarization result.

Equation (4) can be rewritten as

$$\tau \in \mathcal{P}_k \quad \text{if} \quad b_v(\tau)\phi_k(\theta(\tau)) = 1 \quad (5)$$

where

$$\phi_k(\theta) = \begin{cases} 1 & \text{if } |\theta - \theta_k| \le th \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

θ_k is the centroid of the k-th cluster, and th is a threshold.

2.2. Problems with previous approach

2.2.1. Problem 1: Frame-wise DOA

Because the previous approach employed the GCC-PHAT, it estimates one DOA per frame. The system cannot estimate multiple DOAs even if there are some speakers in a frame. This causes a lot of missed speaker time (false rejection in VAD terminology), and degrades the performance.

2.2.2. Problem 2: Directional noise

Thanks to the VAD, noise periods (non-speech periods) are estimated successfully in most of the time frames. However, directional noise is sometimes detected as a speaker if the directional noise is detected as speech in terms of VAD, or if it becomes dominant in terms of the GCC-PHAT calculation when the other speaker is speaking with small power, e.g. speech onset/offset, and unvoiced speech part. Such false detections increase the false alarm speaker time (FAT, see Section 4.2) and degrade overall performance.

2.2.3. Problem 3: Binary VAD

The VAD performance changes when we change the threshold for the speech / non-speech discrimination. And, what is worse, overall diarization performance depends on the VAD performance. If the false acceptance rate (FAR) of the VAD is high, there is little missed speaker time (MST, see Section 4.2) but increased FAT. On the other hand, when the false rejection rate (FRR) of the VAD is high, there is little FAT but MST increases. Therefore, it is difficult to set a preferable VAD threshold for good diarization performance.

3. PROPOSED METHODS

This section proposes three refined methods to handle the problems discussed in the previous section.

3.1. Method I: Employ TFDOA

To solve problem 1, we employ the TFDOA. That is, instead of (2), the TDOA is estimated at each time-frequency slot by

$$q'_{jj'}(f,\tau) = \frac{1}{2\pi f} \arg\left[x_j(f,\tau)x^*_{j'}(f,\tau)\right] \quad (7)$$

and the TFDOA $\mathbf{q}(f,\tau)$ is estimated in the same way as (3). Here, we employ only the azimuth $\theta(f,\tau)$ for simplicity.

If τ is a speech frame $\tau \in \mathcal{P}_S$ and a sufficient number of frequency bins f have a DOA adjacent to the k-th centroid θ_k in frame τ, then speaker k is regarded as speaking in the frame τ. That is, the individual speaker periods $\mathcal{P}_k$ are determined by

$$\tau \in \mathcal{P}_k \quad \text{if} \quad b_v(\tau)\lambda_{1k}(\tau) > th1, \quad (8)$$

where $th1$ is a threshold,

$$\lambda_{1k}(\tau) = \sum_f \phi_k(\theta(f,\tau)) \quad (9)$$

is a DOA voting, and $\phi_k(\theta)$ has the same definition as (6). Figure 1 also shows a block diagram of this proposed method I.

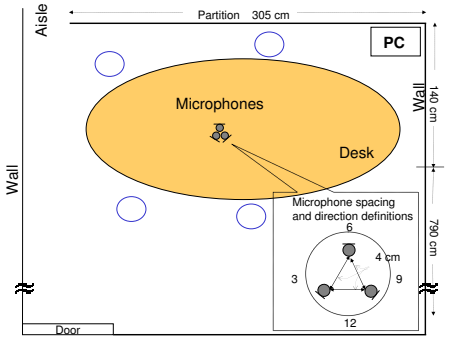

Fig. 3. Room setup. Small ellipses illustrate example speaker locations. The reverberation time was around 350 ms.

Table 1. Conversation recordings. Each recording lasted five minutes.

Evaluation data ID	#Speaker	Overlap [%]	#Turn-taking	#Utterance
crossword puzzle 1	4	18.6	149	185
crossword puzzle 2	4	13.0	183	218
discussion	3	10.8	126	172
conversation	3	34.8	243	278

3.2. Method II: Employ amplitude weight

To suppress the noise influence, here we assume that the signal to noise ratio is sufficiently high, i.e., the amplitude of the speech spectrum is greater than that of the noise spectrum. By exploiting this assumption, method II employs the amplitude weight for the DOA voting:

$$\lambda_{2k}(\tau) = \sum_f \frac{|x_1(f,\tau)|}{\sum_f |x_1(f,\tau)|} \phi_k(\theta(f,\tau)) \tag{10}$$

where $\phi_k(\theta)$ has the same definition as (6). From the amplitude weight, it can be expected that the speech time-frequency slot garners a large number of votes and noise time-frequency slot receives few votes.

The individual speaker periods $\mathcal{P}_k$ are determined in the same way as in Method I:

$$\tau \in \mathcal{P}_k \quad \text{if} \quad b_v(\tau)\lambda_{2k}(\tau) > th2 \tag{11}$$

where $th2$ is a threshold.

3.3. Method III: Employ probabilistic VAD

To handle the third problem, here we utilize the probability $p_v(\tau)$ of speech activity at each frame τ. Such a probability can be estimated with a VAD. Here, we also employ the MUSCLE-VAD introduced in Section 2.1, and utilize the forward probability (see (11) in [8]). As we employ an array of multiple microphones, first we calculate the probability for each channel independently and then take the average.

The individual speaker periods $\mathcal{P}_k$ are determined by

$$\tau \in \mathcal{P}_k \quad \text{if} \quad p_v(\tau)p_k(\tau) > th3 \tag{12}$$

where $th3$ is a threshold and $p_k(\tau)$ is the probability of speaker k speaking. Here as the probability $p_k(\tau)$, we employ the normalized DOA voting result

$$p_k(\tau) \equiv \lambda_{3k}(\tau) = \frac{\lambda_{2k}(\tau)}{\sum_k \lambda_{2k}(\tau)}. \tag{13}$$

4. SYSTEM EVALUATION

4.1. Setup

Experiments were performed in the room shown in Fig. 3 whose reverberation time was around 350 ms. We recorded some conversations between 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. The personal computer (PC) in Fig. 3 could be the noise source. Figure 3 also illustrates our direction definition.

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 diarization difficult. Reference diarization labels were generated by employing a hand-labeled transcription, which includes temporal information about the speech onsets and speech offsets of each speaker.

The sampling rate was 16 kHz for VAD and 8 kHz for DOA estimation. The frame size for STFT was 64 ms, and the frame shift was 32 ms. The thresholds for the proposed methods were $th1 = 40$, $th2 = 0.2$, and $th3 = 0.4$. They were decided in our preliminary experiments.

4.2. Evaluation measure

We evaluated performance with the diarization error rate (DER),

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

which was 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) [3]. If the estimated number of speakers exceeds the true number of speakers, such *ghost* speaker periods were regarded as the SET. We also evaluated the DER after smoothing (hangover) the speaker diarization result $\mathcal{P}_k$, where short fragments and short pauses were removed.

4.3. Results and discussion

Table 2 summarizes the results. With the previous method, which cannot estimate multiple DOAs in a frame, we had a large MST rate. When using Method I, where multiple DOAs can be estimated, the MST rate decreases. Figure 4 shows an example illustration. With the previous method (Fig. 4 (b)) we can estimate only one DOA for each frame. On the other hand, several DOAs can be estimated in a frame with Method I (Fig. 4 (c)). However, the FAT rate becomes a little worse with Method I. One reason for this is that the directional noise (PC) can be falsely detected easily by Method I. Figure 5 (c) shows an example of such false detection of the directional noise (from direction 8).

With Method II, where we utilize the amplitude weight, we can improve the FAT rate. Figure 5 shows this result. With the previous method (Fig. 5 (b)) and the proposed Method I (Fig. 5 (c)), some directional noise from direction 8 was detected for 6 to 8 seconds. By contrast, with Method II (Fig. 5 (d)), the directional noise is not detected thanks to the amplitude weight.

The performance further improved when we used Method III, which employs the probabilistic VAD result. When we test "crossword puzzle 2" data with Method II (Table 2 (b)), the FAR of the VAD was 26.8% and the FRR of the VAD was 13.6%. If we use the VAD of FAR=45.7% and FRR=2.8%, MST, FAT, SET and DER

Table 2. Diarization results [%] for each evaluation data. (Values in bracket indicate the DER after hangover.)

(a) Crossword puzzle 1

Method	MST	FAT	SET	DER
Previous	35.3	8.1	9.7	53.2 (41.1)
Method I	31.4	10.5	6.4	48.3 (37.0)
Method II	31.2	8.2	3.9	43.2 (28.5)
Method III	25.1	6.5	3.6	35.2 (21.9)

(b) Crossword puzzle 2

Method	MST	FAT	SET	DER
Previous	28.4	12.7	8.9	50.1 (40.0)
Method I	24.2	17.8	5.2	47.2 (36.1)
Method II	23.9	15.1	3.1	42.2 (29.2)
Method III	19.6	13.4	2.6	35.5 (25.0)

(c) Discussion

Method	MST	FAT	SET	DER
Previous	43.5	5.1	2.1	50.6 (35.9)
Method I	38.5	9.0	1.5	49.0 (31.9)
Method II	37.8	7.8	1.5	47.0 (32.6)
Method III	38.5	5.5	1.5	45.5 (29.9)

(d) Conversation

Method	MST	FAT	SET	DER
Previous	38.2	7.5	8.0	53.7 (40.7)
Method I	30.0	13.1	7.6	50.7 (37.7)
Method II	30.4	12.3	6.2	48.9 (32.3)
Method III	33.2	9.5	6.1	48.8 (34.3)

were 5.6, 30.7, 5.3, and 41.6, respectively. That is, the diarization performance depends on the VAD performance as pointed out in Section 2.2.3. With Method III, without tuning the VAD, we can obtain good diarization performance.

5. CONCLUSION

We proposed methods for improving the performance of our meeting diarization system. Instead of the GCC-PHAT approach, we utilized the DOA at each time-frequency slot for the diarization. This refinement reduces the missed speaker time, and improves the performance. In addition, we also showed that the use of the observation amplitude at each time-frequency slot effectively disregards directional noise. We also showed that the use of a probabilistic representation of the VAD results produces high levels of performance.

6. ACKNOWLEDGMENT

The authors thank Mr. Taku Hasegawa of Toyohashi University of Technology for his help in conducting the experiments.

7. REFERENCES

[1] 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.

[2] D. Macho, J. Padrell, A. Abad, C. Nadeu, J. Hernando, J. McDonough, M. Wölfel, U. Klee, M. Omologo, A. Brutti, P. Svaizer, G. Potamianos, and S. Chu, "Automatic speech activity detection, source localization, and speech recognition on the CHIL seminar corpus," in *Proc. of ICME'05*, July 2005, pp. 876–879.

[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.

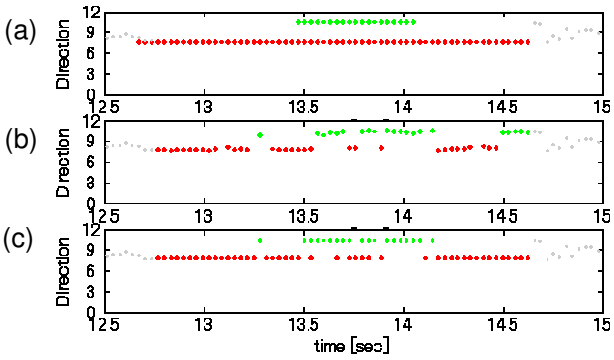

Fig. 4. Diarization result examples. (a) ground truth label, (b) with previous method and (c) with Method I. Gray dots indicate noise frames.

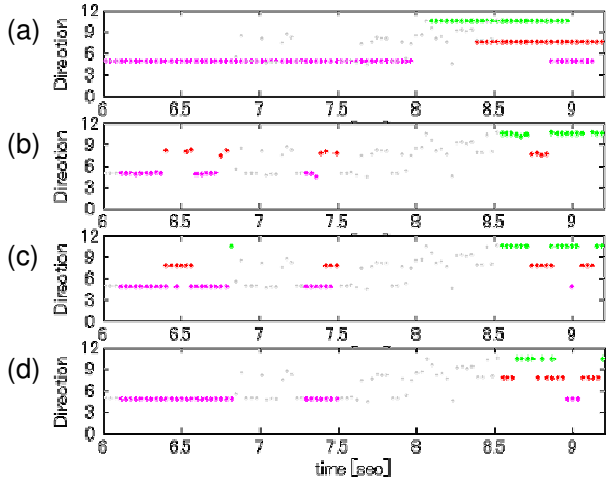

Fig. 5. Diarization result examples. (a) ground truth label, (b) with previous method, (c) with Method I and (d) with Method II. Gray dots show noise frames.

[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, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "Speaker indexing and speech enhancement in real meetings / conversations,," in *Proc. of ICASSP'08*, Mar. 2008, (to appear).

[7] M. L. Seltzer, I. Tashev, and A. Acero, "Microphone array post-filter using incremental Bayes learning to track the spatial distributions of speech and noise," in *Proc. of ICASSP'07*, Apr. 2007, vol. I, pp. 29–32.

[8] 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, (to appear).

[9] 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.

[10] 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.

[11] 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.

[12] 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.