HANDLING SPEAKER POSITION CHANGES IN A MEETING DIARIZATION SYSTEM BY COMBINING DOA CLUSTERING AND SPEAKER IDENTIFICATION

Tobias Hager, Shoko Araki, Kentaro Ishizuka, Masakiyo Fujimoto, Tomohiro Nakatani, Shoji Makino*

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

ABSTRACT

This paper presents a meeting diarization system that estimates who spoke when in a meeting, especially if there is a speaker position change. Our previous system utilized solely the direction of arrival (DOA) information for meeting diarization. Therefore, when someone moves from one place to another, our previous system mistakes the utterances from second place for another person's utterances. In order to handle such a speaker position change in a meeting, this paper tries to combine DOA information and Gaussian mixture model (GMM)-based speaker identification (SI). First, relying on the DOA information, we cluster the recorded meeting into segments. Several segments coming from the same DOA are utilized to construct GMMs for the Mel frequency cepstral coefficients (MFCC) for each speaker. The diarization result is obtained by evaluating every segment clustered by DOA against all speaker models. We obtained encouraging results for simulated meetings with a measured room impulse response and a recorded meeting, where the reverberation time of the room was about 350 ms.

***Index Terms*—** meeting diarization, speaker identification, voice activity detector, direction of arrival

1. INTRODUCTION

In recent years audio diarization has become an important topic and was studied extensively [1]. There are several applications like speaker diarization, also called "who spoke when" estimation, which can be split up into three categories: broadcast news diarization, telephone conversation and meeting diarization. Our paper concentrates on the last one.

In particular, this paper handles the problem of a speaker position change in meeting scenarios. This position change could occur often in real meetings, for example if the presenter is changing or a whiteboard is shared by speakers.

One of the most important technique for meeting diarization is speaker clustering. By speaker clustering, certain sound objects included in the recorded signals are separated into several clusters, each of which is assumed to correspond to one speaker in the meeting. Our previous system was build up solely on clustering DOA information. The effectiveness of this approach has been well confirmed by our experiments [2, 3]. However, only with DOA information, if a speaker changes his position, he was either recognized as another speaker or as a completely new speaker. To cope with such an issue we propose to combine DOA clustering and SI. In this paper, we assume that the speakers could change their seats, and that there is just one speaker at each direction at the same time.

*The author is on leave from the Chair of Multimedia Communication and Signal Processing, University Erlangen-Nuremberg.

Our proposed method in this paper is as follows: First, using the DOA information, we cluster the speech segments. We adopt a DOA estimation technique that can detect more than one source at the same time for this purpose [3]. Then, each speech segment is enhanced based on a microphone array speech enhancement technique, referred to as a maximum signal-to-noise ratio (MaxSNR) beamformer [11]. Finally the speech segments are clustered into individual speaker clusters by using SI techniques. Here, for the sake of simplicity, we assume all speakers stay and give a few utterances at their initial positions during first several seconds of the meeting. Based on this assumption, speaker segments over a specific period in the beginning are utilized to train the speaker models for each person. Then, each segment following the training period is assigned to one trained speaker by determining the GMM with the highest likelihood for its feature vector sequence [4].

Previously, the authors of [5] have proposed to combine DOA and speaker spectral features for meeting diarization. The paper just combines these features to cluster the speakers for a Bayesian information criterion (BIC) calculation, which has been widely employed in the meeting diarization area [1]. In [5], they extracted DOAs as time differences that give the cross correlation peaks between microphones. However, this is not a robust DOA feature in a real meeting situation. In addition, this feature does not allow us to handle time segments, in which more than one speaker talk at the same time. In contrast, our method can handle speaker overlaps by using time-frequency domain DOA and speech enhancement techniques. Because the performance with solely the DOA clustering has been confirmed [2, 3], it is possible to focus on the SI of each segment decided by the DOA clustering. The main contribution of this paper is to propose a new way for combining DOA and speaker spectral information, and to confirm whether a SI technique achieves better meeting diarization performance or not. The experimental results with a measured room impulse response and a recorded meeting, where the room reverberation time was about 350 ms, show that our new method can handle the speaker position change in a meeting.

2. PROPOSED SYSTEM

Let us denote the recorded data $\mathrm{x}_j(t)$, with the microphone index j ranging from one to three in our case. In the following we utilize the time-frequency representation $x_j(f,\tau)$ of our observations $\mathrm{x}_j(t)$, which can be obtained by the short-time Fourier transform (STFT). Here f is a frequency and τ is a frame index. The overall system flow is visualized by Fig. 1 and works in the following way:

First we apply a voice activity detector (VAD) on the recorded data, to detect whether a speaker is present or not. The next step contains the estimation of the DOA for every frame with the generalized cross correlation method with a phase transform (GCC-PHAT),

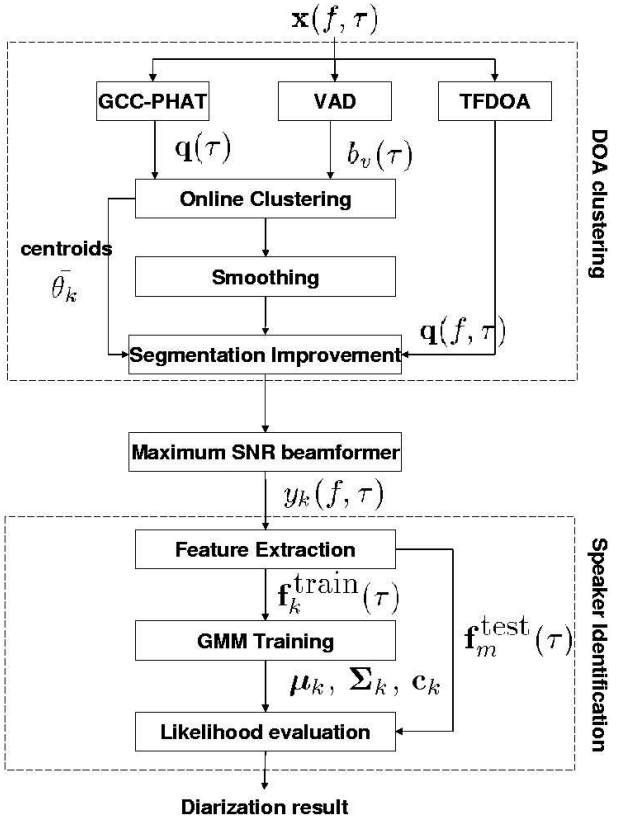

Fig. 1. Block diagram of overall system flow.

to determine the direction of the speakers. Based on the framewise DOA information we perform segmentation of recorded data with an online clustering [6] and use the cluster centroids $\bar{\theta}_k$ as the directions from where speaker k is speaking.

We additionally extract time-frequency direction of arrival (TF-DOA) information [3], which provides the DOA for every frame and every frequency bin separately. This result will be used later to improve the segmentation result as well as the detection of overlapping speaker segments.

To cope with the problem of clusters containing just a few frames as well as noisy results we apply a smoothing operation. Executing this operation results in smoother segments by closing small gaps and the elimination of isolated frames caused by noise. Using the results of the DOA clustering and the VAD, we obtain MaxSNR beamformer coefficients which are applied to filter the recorded meeting to suppress noise and other active speakers.

Based on our segmentation result the system extracts the feature vectors consisting of 12 MFCCs plus fundamental frequency for training and identification segments (see Section 2.5 and 2.6). The GMM model parameters are estimated then using the expectation maximization (EM) algorithm [7] on the obtained training feature sequences. For the GMM training, we assume that no speaker change occurs up to a certain time point.

SI includes the likelihood calculation for each identfication segment against all speaker models. Finally we assign the speaker with the highest likelihood to that interval. In the following sections each part of the system is described in detail.

2.1. VAD

Aim of this step is to find the speech periods in the recording. Regions of non-speech, which we have to differentiate can be manifold, such as silence or background noise. In order to construct such a VAD which is robust to various kinds of noise, it is based on a two stream approach using speech and non speech discriminators. These are periodic to aperiodic component ratio-based detection (PARADE) and a switching Kalman filter (SKF)-based approach [8].

In this paper the VAD results $b_v(\tau)$ are given by binary labeling (1 for speech frame, 0 for non-speech frame). For our microphone array VAD is done for every channel separatley first and joined afterwards by a single binary OR operation. The speech period is then defined as $\mathcal{P}_S = \{\tau \mid frames\ labeled\ as\ 1\}$.

2.2. DOA estimation

In this paper, we use VAD and DOA information for segmentation. The latter we get by first estimating the time difference of arrival (TDOA) $q'_{jj'}(\tau)$ for all microphone pairs j and j' using the GCC-PHAT [9]. Using the TDOA and the given microphone coordinate information we get the DOA estimates $q(\tau)$ [10]. Additionally we estimate the TFDOA at each time-frequency slot with

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

which we will use later to refine our segmentation results obtained by clustering the framewise DOA information $q(\tau)$. In our setup we only make use of the azimuth $\theta(\tau)$ for simplicity in both cases. Further details of this method can be obtained in [3].

2.3. Segmentation and smoothing

Relying on DOA information, the speech period $\mathcal{P}_S$ is then classified into each speaker period $\mathcal{P}_k$ $(k = 1, \cdots, \#speakers)$, which gives us the segmentation result. This is done by an online clustering (leader-follower clustering) [6] algorithm. The pseudocode of this algorithm can be found in [2].

Depending on the threshold to add new clusters or assign the point to an existing cluster, the final clustering result has clusters, that do not represent a speaker in reality. Those clusters have a very sparse frame density distributed over the time axis. A simple way to eliminate those frames consists of using a smoothing filter. In our case we apply a sliding window of odd length w and set the frame belonging to the midpoint of that window to one if more than 45 % of the frames of the total window size w are classified as speech. The initial and final $\frac{w}{2}$ frames are set to zero, assuming nobody is talking in the first and last 0.8 seconds of the recording.

All clusters, that have frames left after smoothing are accepted as real speaker segmentation information and the centroid $\bar{\theta}_k$ is used as the recognized speaker direction. In a further step this segmentation information is improved by the TFDOA data as explained in [3].

2.4. Maximum SNR beamformer

In a real meeting situation, like a discussion, speaker overlaps occur frequently. This overlaps would disturb the GMM training and decrease the SI performance significantly. Another common problem can result from projectors or personal computers in a meeting room causing directional noise, which can also be classified as an additional speaker in the worst case.

To cope with these problems and to reduce the influence onto SI we conduct a blind speech separation with a MaxSNR beamformer [11].

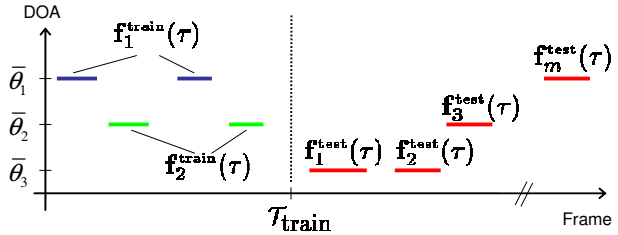

Fig. 2. Visualization of feature vector extraction for training and testing data. Each bar represents a segment by DOA clustering.

By applying the beamformer coefficients $\mathbf{w_k}$ to our recorded signal we get the enhanced speech for the speaker at the k-th position

$$y_k(f,\tau) = \mathbf{w}_k^H(f)\mathbf{x}(f,\tau). \quad (2)$$

2.5. Feature extraction and GMM training

Then, using the enhanced signals with the MaxSNR beamformer, we conduct the SI, which includes feature extraction, GMM training and likelihood evaluation stages (see Fig. 1). The training data for the estimation of our GMM parameters is defined by the first part of the meeting. Here, the training period $\tau = 1, \cdots, \mathcal{T}_{train}$ is set in advance. The enhanced signal y_k in the training period for each direction $\bar{\theta}_k$ is utilized as the training data for each speaker GMM.

In the feature extraction stage, from the training data, we calculate the feature vector $\mathbf{f}_k^{train}(\tau)$ for each frame τ belonging to speaker k. We extract 12 MFCCs and the fundamental frequency for every frame where speaker k is active before $\mathcal{T}_{train}$ and construct the 13-dimensional feature vector sequence by merging together all feature vector sequences for each segment of speaker k. This procedure is visualized in Fig. 2. The fundamental frequency is estimated by a maximum search in the autocorrelation function [12].

Then, using these feature vectors, the GMM for each direction $\bar{\theta}_k$ is trained. We train the GMM parameters using an iterative EM algorithm [7]. The GMM parameters include the mean μ_k^l, diagonal covariance matrix $\mathbf{\Sigma}_k^l$, and the weight c_k^l, where $l = 1, \cdots, 10$ denotes the mixture component index.

2.6. Likelihood evaluation

Finally we identify the speaker for each segment after $\tau = \mathcal{T}_{train}$. One segment consists of a consecutive sequence of frames stemming from the same direction. For each segment m we take the extracted feature vectors $\mathbf{f}_m^{test}(\tau)(m = 1, \cdots, \#segments)$ (see Fig. 2) and calculate the likelihood for all speaker models [4]. An identified speaker for segment m is obtained through selection of the model with the highest likelihood. This gives us the final diarization result.

3. EXPERIMENTAL RESULTS

3.1. Setup

We conducted experiments using simulated data with measured impulse responses and also one measured meeting. The measured meeting was performed in the room shown in Fig. 3 with a reverberation time around 350 ms. In our scenario four speakers, two males and two females were present. Additionally a personal computer has been placed in the room which acted as a noise source. The distance between speakers and microphone array was approximately one meter. The exact arrangement can be seen in Fig. 3. The recording time was set to five minutes.

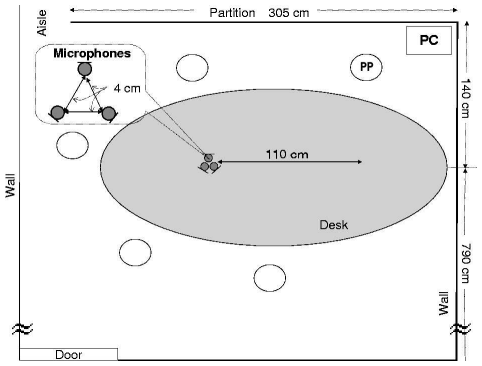

Fig. 3. Room setup. Small circles illustrate initial speaker locations each speaker moved to the presenter position (PP).

Simulations were build by convolving clean speech recordings with measured room impulse responses recorded in the same room as the measured meeting. We also recorded noise in this room including the personal computer. The noise was added with 10 dB signal to noise ratio (SNR). The meeting length was set to 90 seconds. We conducted four simulations: Simulations 1 and 2 had no overlapping segments and three speakers each. Simulation 3 contained also no overlap, but four speakers. In simulation 4, speaker overlaps (six seconds in total) occured and four speakers were present. We evaluated 10 different speaker combinations for each simulation and averaged over the results.

The sampling rate of our system was 16 kHz, STFT was done with a 64 ms window and 32 ms frame shift. The feature extraction stage, which operated on a 32 ms frame size and 8 ms frame shift, provided 12 MFCCs, excluding the 0'th coefficient. The GMM was build up with 10 mixture components. For the initialization we used random values for the means μ_k^l, equal probabilities for the mixture component weights c_k^l and estimates from the feature vector sequence for covariance matrices Σ_k^l. The EM iteration was executed 10 times or stopped if the total likelihood increase for the training data fell below a threshold. The smoothing filter length of w was set to 51 samples.

3.2. Evaluation

The diarization performance of our system was measured with the diarization error rate (DER),

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

established by NIST [13]. It includes missed speaker time (MST) (no speaker in estimation, but reference), false alarm time (FAT) (speaker in estimation, but not in reference) as well as the speaker error time (SET) (wrong speaker in estimation), which could be used to measure the speaker recognition performance.

If the number of detected speakers outnumbers the real number of participants this time was marked as SET. Ground truth for the measured meeting was generated by employing a hand-labeled transcription, including temporal information about the speech onsets and offsets of each speaker.

3.3. Results and discussion

In our measured meeting every speaker was introducing himself for 10 to 30 seconds. After this point each speaker changed his seat to

Table 1. Experimental results for measured meeting [%]

Evaluation data ID	DOA with SI				DOA without SI			
	DER	MST	FAT	SET	DER	MST	FAT	SET
meeting 1	23.1	10.1	4.1	**8.9**	55.2	10.1	4.1	41.0

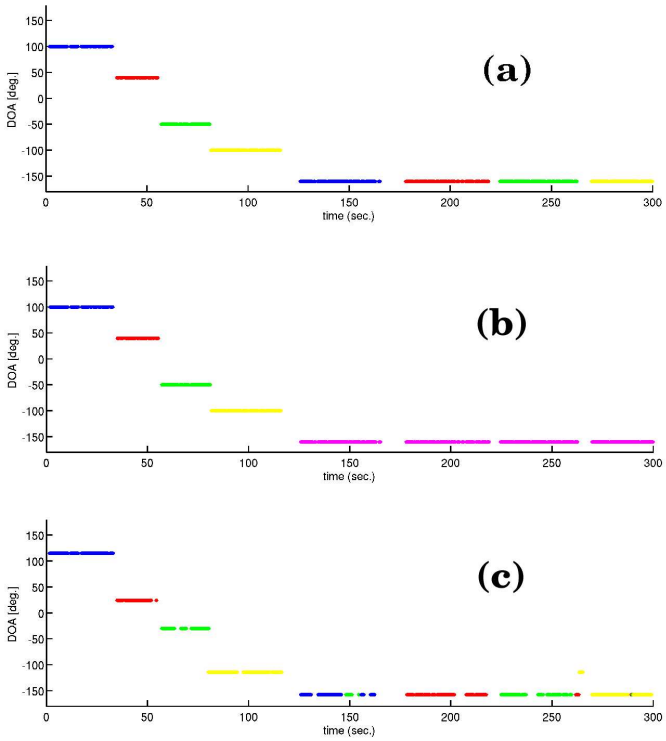

Fig. 4. Ground truth (a), diarization result of previous (b) and new approach (c) for the measured meeting. Each color represents one speaker.

the presenter position (marked with PP in Fig. 3) one after another and spoke there for 30 to 40 seconds. Ground truth (a), diarization result of our previous system (b), and diarization result of our new approach (c) are visualized in Fig. 4. We set $\mathcal{T}_{train}$ to 120 seconds and used the segmented data to build our speaker models.

An overview of the yielded diarization errors is depicted in Table 1. We compared the performance with and without SI. With our previous method, without SI and DOA only, we could not handle a speaker position change which led to a big SET. Such a fault is visualized in Fig. 4 (b). After 120 seconds each speaker spoke in rotation (Fig. 4 (a)). However, with our previous method, all the segments from a DOA of -160° are classified as one person. On the other hand our new method can identify each speaker from the same position (DOA of -160°, see Fig. 4 (c)) and outperforms the previous approach shown in Table 1. Our system achieves a SET of 8.9%.

Table 2 summarizes the simulation results. In the beginning of our simulations around 10 seconds of each speaker are used for the speaker model parameter training. The following speaker segments have a length of three to five seconds and are uttered from a different position. Due to the big window for smoothing and continous utterances over the whole interval MST and FAT were very low. We verified the capability of the speaker identifier for this short training time, achieving a 85% identification rate for training segments ten seconds and identification segments five seconds in length for 20 different speakers in our preliminary experiments. Simulation 4 contains speaker overlaps and therefore has a decreased performance. However we still obtain such an encouraging result, especially in terms of SET.

Table 2. Experimental results for simulated meetings [%]

Evaluation data ID	DOA with SI				DOA without SI			
	DER	MST	FAT	SET	DER	MST	FAT	SET
simulation 1	8.7	2.1	0.1	**6.4**	34.8	2.2	0.1	32.5
simulation 2	11.7	2.2	0.1	**9.4**	45.8	2.2	0.1	43.5
simulation 3	11.9	6.5	0.0	**5.4**	39.0	6.5	0.0	32.5
simulation 4	18.3	10.1	1.3	**6.9**	40.5	10.1	1.3	29.1

4. CONCLUSION

We proposed a method to handle a speaker position change in a meeting diarization system based on the combination of DOA clustering and SI. By utilizing the SI technique we are able to improve the diarization performance for speaker changes in a meeting successfully. Our future work includes the evaluation in meetings with more speaker overlap, more real recorded meetings, a comparision with a SI only method, and the detection of a person, who was not present in the training period.

5. REFERENCES

[1] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1557–1565, Sept. 2006.

[2] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "Speaker indexing and speech enhancement in real meetings / conversation," in *Proc. of ICASSP '08*, Mar. 2008, pp. 93–96.

[3] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "A DOA based speaker diarization system for real meetings," in *Proc. of HCSMA '08*, Apr. 2008.

[4] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[5] J. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple distant microphone meetings: mixing acoustic features and inter-channel time-differences," in *Proc. of ICSLP'06*, Sept. 2006, pp. 2194–2197.

[6] R. Duda, P. Hart, and D. Stork, Eds., *Pattern Classifycation*, Wiley Interscience, 2nd edition, 2000.

[7] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy Stat. Soc.*, vol. 39, pp. 1–38, 1977.

[8] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme," in *Proc. of ICASSP '08*, Mar. 2008, pp. 4441–4444.

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

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

[11] 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.

[12] W. Hess, Ed., *Pitch determination of speech signals*, Springer-Verlag, 1st edition, 1983.

[13] `http://www.nist.gov/speech/tests/rt/.`