Turn Taking-based Conversation Detection by Using DOA Estimation

Yohei Kawaguchi, Masahito Togami, and Yasunari Obuchi

Central Research Laboratory, Hitachi, Ltd.
1-280, Higashi-koigakubo Kokubunji-shi, Tokyo 185-8601, Japan
{yohei.kawaguchi.xk, masahito.togami.fe, yasunari.obuchi.jx}@hitachi.com

Abstract

We propose a new method that detects conversation groups when multi-conversation groups exist simultaneously. The proposed method uses hands-free microphone arrays without wearable microphones. It has two main features: (a) We integrate a conventional turn taking-based conversation detection method with Direction of Arrival (DOA) estimation-based Voice Activity Detection (VAD). (b) The proposed method estimates the number of speakers for DOA estimation-based VAD by using the turn-taking rules. Experimental results indicate that the performance of the proposed method with only microphone arrays setup in rooms is comparable to that of the conventional methods with wearable microphones.

Index Terms: microphone array, direction of arrival, conversation detection, turn taking, voice activity detection.

1. Introduction

Automatic conversation scene analysis is an important technology. This technology can allow to indexing the video/speech data of the conversations, for the rapid retrieval and summarization. It is also important for human symbiotic robots with a speech interface because the conversation scene analysis enables to estimate the information about the time when they are allowed to speak to users and the time when they should interpret utterances of users as commands.

The kinds of conversations can be categorized into two classes: “planned conversations” such as meetings planned in advance, etc. and “unplanned conversations” such as chance meetings that occur spontaneously, etc. Most of researches on the automatic conversation analysis handle with planned conversations. In automatic conversation analysis, speaker diarization [1] is a typical technology. The goal of this is to answer the question, “who spoken when?” Most of conventional speaker diarization systems [1] consist of voice activity detection (VAD) and speaker segmentation. These approaches are typically based on classification of Gaussian Mixture Model (GMM) with mel-frequency cepstral coefficient (MFCC). There are several approaches also been effects in scene understanding for more higher level of interaction than utterance in planned conversations such as meetings [2][3][4]. Some approaches recognize conversation patterns [2], estimate the interaction structure, such as “which behavior is triggered by which behavior” and track the locations of people [4]. They make use of multisensors such as not only microphones but also cameras because the visual information such as gaze and head orientations is important to infer conversation scenes.

On the other hand, there have some researches been efforts in unplanned conversations. Conversations of this type occur spontaneously and randomly. Thus, one of the important tasks of these researches is to detect conversations, in other words, it is to detect groups of people that belong to the same conversation when different multi-conversations exist. This task is called “conversation detection.” There are some researches on the conversation detection [5][6][7][8][9][10]. In unplanned conversations, it is difficult to configure cameras at such locations that all the faces are captured in real environments. Therefore, these existing researches use wearable microphones attached to each person, and deal with the conversation detection based on only the audio data from these microphones. Nakakura et al. [5] makes use of the assumption that the voice of the speakers of the same conversation are recorded louder at the microphones of these people than that of other conversations because people that belong to the same conversation exist at near locations relatively. This approach clusters people by correlation of input loudness between microphones. However this assumption does not hold when people are close together between different groups in such common office environments. Some researches detect conversations by focusing on the timing characteristic of utterances of conversations [6][7][8][9][10]. The typical timing characteristic of utterances is the turn-taking rules, the concept of which was pioneered by Sacks et al. [11]. Bridiczka et al. [6] proposed the approaches that use HMM the states of which correspond to possible partitioning of the speakers and whose observations are the voice activity of each speaker. This approach utilizes the turn-taking rules implicitly, and needs a learning phase of HMM. In contrast, other approaches [7][8][9][10] utilize the turn-taking rules explicitly. These researches employ the mutual information of voice activity between speakers. However, they use wearable microphones, and it is currently unrealistic that every person wear a microphone in unplanned conversations such as chance meetings.

In this paper, we propose the method that detects conversation groups with hands-free microphone arrays without wearable microphones for unplanned conversations. This paper has two main contributions: (a) We integrate turn taking-based conversation detection approach with DOA estimation-based VAD. (b) We propose the method that estimates the number of speakers for DOA estimation-based VAD by making use of the turn-taking rules. In Section 4, our experimental results show that the performance of the proposed method with only a microphone array setup in rooms is comparable to that of the conventional methods with wearable microphones.

2. Problem statements and notation

We assume that $K$ speakers exist. The voices of the speakers are recorded at a microphone array that consists of $M$ microphones. The recorded signals are AD-converted and analyzed by STFT. These multi-channel signals represented as $x(f, τ) = [x_1(f, τ) \cdots x_M(f, τ)]^T$, where $x_m(f, τ)$ is the in-
put signal of the m-th microphone, f is the index of a frequency bin, and τ is the index of a time-frame. \( x(f, \tau) \) is modeled as follows,

\[
x(f, \tau) = \sum_{i=1}^{K} a_i(f)s_i(f, \tau) + n(f, \tau),
\]

where \( a_i(f) \) is the steering vector of the i-th speaker, \( s_i(f, \tau) \) is the source signal of the i-th speaker, and \( n(f, \tau) \) is the background noise. \( a_i(f) \) is normalized such that \( |a_i(f)| = 1 \).

In this paper, we detect conversations from \( x(f, \tau) \). The conversation detection is to obtain the cluster set \( C \) for the speaker set \( S = \{1, \cdots, K\} \), where \( C = \{C_1, \cdots, C_L\} \), \( C_i \subset S \), all the speakers in the cluster \( C_i \) is the member of the same conversation represented by \( C_i \), and \( \forall C_1, C_2(\neq C_i) \in C, \forall i \in C_1, i \notin C_2 \).

3. Turn-taking-based conversation detection by using DOA estimation

3.1. Voice activity detection for each speaker

We assume the “sparseness” assumption that only one source is active in each time-frequency bin. Using this assumption, we estimate the Direction of Arrival (DOA) \( \hat{H}(f, \tau) \) in each \( (f, \tau) \) by Modified Delay-and-Sum Beamformer (MDSBF) [12],

\[
\hat{\Theta}(f, \tau) = \arg \max_{\theta} |a_\theta(f)x(f, \tau)|^2,
\]

where \( a_\theta(f) \) is the theoretical steering vector for the discrete direction \( \theta \), and this vector is calculated from the configuration of the microphones. We create DOA histogram \( H(\theta, \tau) \) by voting for the direction \( \Theta(f, \tau) \).

\[
H(\theta, \tau) = \sum_f |a_{\hat{\theta}(f, \tau)}(f)x(f, \tau)|^2
\]

We estimate the directions of each speaker \( \Theta_i \) by as the centroid of k-means clustering for \( \theta \) weighted by \( w(\theta) = \int H(\theta, \tau) d\tau \). However, k-means clustering has the difficulty that it needs the parameter of the number of the speakers \( K \) and \( K \) is unknown generally in real environments of unplanned conversations. In Section 3.3, we will introduce an estimation method of the number of the speakers.

Next, we estimate whether the speech of each speaker is present or not in each time-frame by voice activity detection (VAD) The voice activity \( v_i(\tau) \in \{0, 1\} \) is 1 when the i-th speaker is speaking, and it is 0 when i-th speaker is not speaking in the time-frame \( \tau \). Assuming that \( |n(f, \tau)| \) is sufficiently small, we can consider \( H(\hat{\Theta}_i, \tau) \) as the approximation of the power of the voice of i-th speaker \( s_i \), i.e., \( H(\hat{\Theta}_i, \tau) \approx \sum_f |s_i(f, \tau)|^2 \). We can obtain a posteriori SNR of \( \gamma_i(\tau) \) by Minima-Controlled Recursive Averaging (MCRA) [13] for \( \hat{\Theta}_i \).

We calculate \( v_i(\tau) \) by thresholding operation for \( \gamma_i(\tau) \).

\[
v_i(\tau) = \begin{cases} 
1 & \text{if } \gamma_i(\tau) > T_{\text{VAD}} \\
0 & \text{otherwise},
\end{cases}
\]

3.2. Turn-taking-based conversation detection

The turn-taking theory is characterized by Sacks et al. According to them, the turn-taking rules “minimize gap and overlap” between speakers in conversations. In other words, in the time-frames when one speaker is speaking, the probability that other speakers are speaking is low. We can detect conversations by using this inverse correlation of the voice activity. Here, we introduce the mutual information as the criteria of the inverse correlation. We call the criteria “turn-taking criteria.” The mutual information between speaker \( i \) and speaker \( j \) is defined in [8] as follows:

\[
\mu(i, j) = \sum_{b_i, b_j \in \{0, 1\}} P(v_i = b_i, v_j = b_j) \times \log \frac{P(v_i = b_i, v_j = b_j)}{P(v_i = b_i)P(v_j = b_j)},
\]

where

\[
P(v_i = b_i, v_j = b_j) = \begin{cases} 
\frac{1}{W} \sum_{t=0}^{W-1} v_i(t), & \text{if } (b_i, b_j) = (1, 1), \\
\frac{1}{W} \sum_{t=0}^{W-1} (1 - v_i(t)), & \text{if } (b_i, b_j) = (1, 0), \\
\frac{1}{W} \sum_{t=0}^{W-1} (1 - v_i(t))(1 - v_j(t)), & \text{if } (b_i, b_j) = (0, 1), \\
\frac{1}{W} \sum_{t=0}^{W-1} (1 - v_i(t)), & \text{if } (b_i, b_j) = (0, 0).
\end{cases}
\]

and \( W \) is the parameter of the window size of the calculation of the turn-taking criteria. The mutual information does not indicate the inverse correlation explicitly, and also in the case that the voice activities have the positive correlation between speakers, the mutual information criteria are high. However, assuming that there are no correlations between speakers in different conversations, the mutual information criteria of inter-conversation are small and neglectable. Therefore we can consider the mutual information as the inverse correlation criteria. We can cluster speakers by agglomerative clustering for the mutual information, and we detect conversation groups such that each group consists of the speakers in the same conversation.

3.3. Turn taking-based estimation of the number of speakers

As we mentioned in Section 3.1, the number of speakers \( K \) must be known in order to execute k-means clustering before the voice activity detection for each speaker. However \( K \) is the unknown parameter generally in real environments of unplanned conversations.

There are some researches on the estimation of the number of sources. To the best of our knowledge, all of these researches have proposed the approaches based on only the auditory transfer characteristics [14][15]. These approaches have the difficulty of the overlap between DOA distributions of sources that is caused by the reverberation in real environments and the variance of the positions of mouths. We propose another approach to estimate the number of speakers by using the turn-taking criteria.

Now, in the case that the estimate of the number of speakers \( \hat{K} \) is larger than the actual number of speaker \( K \), we detect the direction \( \hat{\Theta}_i \) and the direction \( \hat{\Theta}_j \) despite actually \( i \) and \( j \) are the same speaker. In this case, we can find strong positive correlations between the voice activity of direction \( \hat{\Theta}_i \) and that of \( \hat{\Theta}_j \) because \( \hat{\Theta}_i \) and \( \hat{\Theta}_j \) are either the direction of the direct sound or that of the reflection for the same speaker. The strong positive correlations make the mutual information \( \mu(i, j) \) a much larger value than that of pairs of turn-taking speakers. Therefore, we propose an approach for estimation of the number of speakers by using the turn-taking criteria that we introduced in Section 3.2. The proposed method consists of following steps:
1. Initialize $K$. $K \leftarrow K_{\text{MAX}}$, where $K_{\text{MAX}}$ is a large integer number.
2. Perform k-means clustering and voice activity detection for each speaker by (2)(3)(4).
3. Calculate the turn-taking criteria $\mu(i, j)$ by (5).
4. $K \leftarrow K - 1$ and go Step 2 if $\exists i, j, \mu(i, j) < T_{\text{MAX}}$, go Step 5 otherwise, where $T_{\text{MAX}}$ is a large real number.
5. Perform conversation detection by agglomerative clustering based on $\mu(i, j)$.

The proposed method estimates the number of speakers by the loop process of these steps.

4. Experimental results

4.1. Experimental setup

We evaluated the proposed conversation detection. Fig. 1 illustrates the configuration of the 4 speakers and the microphone array. The microphone array consists of 8 microphones configured in the semicircle the radius of which is 80mm and the shape of that is the semicircle. We recorded casual conversations at 8 kHz sampling rate and 16 bit-per-sample. As Fig. 2, the recorded data consists of 4 sessions: (a) Session 1 and 2 are for the combination illustrated by Fig. 2 (a). (b) Session 3 and 4 are for the combination illustrated by Fig. 2 (b). The lengths of all the sessions are 60 seconds.

4.2. Conversation detection

Fig. 3 shows the time-averaged DOA histogram $w(\theta) = \int H(\theta, \tau)d\tau$ of Session 1. This result indicates that we can find the peaks of the directions of all the speakers, on the other hand, the overlaps between DOA distributions of sources were caused by the reverberation. Fig. 4 shows an example of the results of VAD of Speaker A and Speaker C in Session 3. The results show that the voice activities of speakers in the same conversation are following the turn-taking rules.

In Fig. 5, we plotted the result of the accuracy of conversation detection as a function of the window size of the calculation of the turn-taking criteria $W$. The accuracy is calculated by $\text{Accuracy} = \frac{N_a}{N_A}$, where $N_A$ is the number of all the variations
of the windowed frames, and \( N_B \) is the number of the cases such that all the speakers are clustered into the actual conversation. Here, in order to separate the part of conversation detection from the part of the estimation of the number of speakers, the parameter of the number of speakers is given as \( K = 4 \). The result shows that, as \( W \) increased, the accuracy increased, and in the case that \( W \) is 60 seconds, the proposed method obtained the correct conversation groups for all the sessions. The accuracy was 83% in the case that \( W \) was 20 seconds. These results indicate that the performance of the proposed method with only microphone arrays setup in rooms is comparable to that of the conventional methods with wearable microphones [10]. On the other hand, as \( W \) decreased to 10 seconds, the accuracy decreased to 40% in some sessions. Thus, we can think that the proposed conversation detection works well for relatively long conversations of more than about 20 seconds, and the method does not work well for some shorter conversations. In relatively short-time conversations, the inverse correlation between speakers cannot be found, and the proposed approach does not obtain the actual conversation groups by using the inverse correlation.

From only these results, we cannot evaluate the proposed method completely. In this experiment, we evaluated it only for the case of two conversations of two speakers. In the future work, we should evaluate the method for more than 2 speakers in the same conversation. The proposed method might consider speakers as members of the same conversation in the case that a speaker changes the conversation that the speaker belongs to. The future work includes the evaluation of this point.

4.3. Estimation of the number of speakers

Next, we show the results of the estimation of the number of speakers. In Fig. 6 we plotted the accuracy of the estimation of the number of the speakers as a function of \( W \). The accuracy is calculated by Accuray = \( \frac{N_A}{N_C} \), where \( N_A \) is the number of the variations of the windowed frames as defined in Section 4.2, and \( N_C \) is the number of the cases such that the estimated number of speakers is correct. In the case that \( W \) is longer than 20 seconds, the proposed method estimates the actual number of speakers with an accuracy of more than 96%. These results show that the turn-taking criteria are effective for the estimation of the number of speakers in the case that the window size is large sufficiently. On the other hand, as \( W \) decreased to about 10 seconds, the accuracy decreased to 64% on the average similarly to the conversation detection. These results indicate that the inverse correlation cannot be found in short conversations. Therefore, for about 10 seconds or shorter conversations, we need to study the combination of the proposed method and conventional methods based on the auditory transfer characteristics.

5. Conclusion

We presented a conversation detection method using hands-free microphone arrays without wearable microphones for unplanned conversations. We integrated a turn taking-based conversation detection method with DOA estimation-based VAD. In addition, the proposed method estimates the number of speakers for DOA estimation-based VAD by using the turn-taking criteria. The experimental results have shown that the proposed method detects conversation groups with an accuracy of 83%, and it estimates the number of speakers with an accuracy of 96%. These results indicate that the performance of the proposed method with only a microphone array setup in rooms is comparable to that of the conventional methods with wearable microphones.

6. References