Detection and Separation of Speech Events in Meeting Recordings

Futoshi Asano, Jun Ogata

National Institute of Advanced Industrial Science and Technology,
Central 2, 1-1-1, Umezono Tsukuba 305-8568, Japan
{f.asano,jun.ogata}@aist.go.jp

10.21437/Interspeech.2006-648

Abstract

When applying automatic speech recognition (ASR) to meeting recordings including spontaneous speech, the performance of ASR is greatly reduced by the overlap of speech events. In this paper, a method of separating the overlapping speech events using an adaptive beamforming (ABF) framework is proposed. The main feature of this method is that all the necessary information for the adaptation of ABF, including microphone calibration, is obtained from meeting recordings based on the results of speech event detection. The performance of the separation is evaluated via ASR using real meeting recordings.

Index Terms: structuring, meeting recording, separation.

1. Introduction

The analysis, structuring, and automatic speech recognition (ASR) of meeting recordings has attracted considerable attention in recent years (e.g., [1, 2]). Especially for small informal meetings, a major difficulty of resaerch on such meetings is that the discussion consists of spontaneous speech, and various types of unexpected speech/non-speech events may occur. One such event is responses by listeners such as "Uh-huh" or "I see" being inserted in short pauses in the main speech. These responses are sometimes very close to or even overlap the speech of the main speaker, and is difficult to remove them by segmentation in the time domain. Due to the insertion of these small speech events, the performance of ASR is sometimes greatly reduced.

In the field of signal processing, various types of sound separation such as blind source separation (BSS) and adaptive beamforming (ABF) have been investigated. By using these methods, signals from different sound sources are separated in the spatial domain, and thus, can be effective for the separation of speech events that overlap in the time domain. Regarding BSS, however, it is difficult to employ this approach for speech event separation since the length of the overlapping section is often very short and data sufficient for separation cannot be obtained.

In this paper, a new approach for the separation of overlapping speech events based on the ABF framework is proposed. The disadvantage of ABF is that information on the location of the target and interference sources must be given as a form of the target steering vector and the noise spatial correlation as described later. Regarding the steering vector, in particular, precise calibration is required for an individual microphone array, and this hinders mass production. In this paper, a method of extracting this information from the non-overlapping section of the meeting recordings based on the results of speech event detection is proposed.

Figure 1 shows an outline of the proposed method. In the first half of the method, speech events are detected and the speaker for each event is identified. In the second half, the overlapping sections of the speech events are separated based on information on the detected speech events.

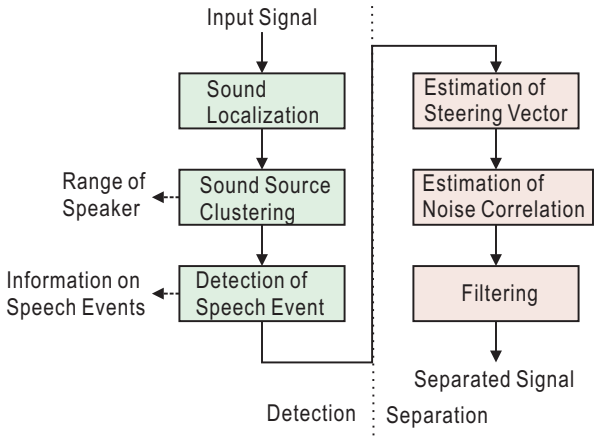


Figure 1: Outline of the proposed method.

2. Detection of Speech Events

2.1. Sound Localization

Meeting data recorded by using a microphone array are segmented into time blocks (the block length is 0.5 s in this paper). The spatial spectrum for each block is then estimated by the MUSIC method [3] extended to a broadband signal [4]. Peaks in the spatial spectrum indicate the position (direction) of the sound sources.

2.2. Clustering of Sound Sources

By clustering the positions of peaks in the spatial spectrum collected for the entire meeting, the range of each speaker is determined. For clustering, k-means was used in this paper. The number of participants is given to the system as the number of clusters. An example of the distribution of the peak positions and the clustering is depicted in Fig.2.

2.3. Detection of speech events

Figure 3(a) shows an example of the peak positions in each block. From this and the ranges of speakers determined in 2.2, the speaker is identified for each peak. These peaks with the speaker being identified is termed speech events. The adjacent speech events in which the same speaker is speaking is then merged into a single speech event. An example of the detected and merged speech events are shown in Fig.3(b).

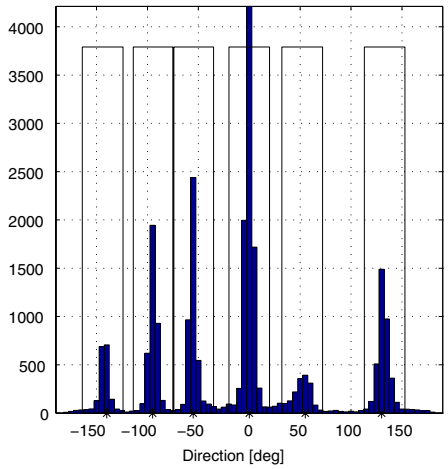


Figure 2: Distribution of the peak positions and results of clustering.

3. Separation of Speech Events

In this section, overlapping speech events are separated using adaptive/non-adaptive beamformer based on the information of the detected speech events.

Some types of beamformers are described in the frequency domain as follows (e.g., [5]):

$$y(\omega,t) = \mathbf{w}^H\mathbf{x}(\omega,t) \quad (1)$$

$$\mathbf{w} = \frac{\mathbf{R}^{-1}\mathbf{a}}{\mathbf{a}^H\mathbf{R}^{-1}\mathbf{a}}. \quad (2)$$

Here, the input vector $\mathbf{x}(\omega,t)$ consists of the short-term Fourier transform of the microphone inputs, while $y(\omega,t)$ represents the beamformer output. The vector $\mathbf{w}$ consists of the beamformer coefficients. Steering vector $\mathbf{a}$ consists of the transfer function of the direct path from the target speaker to the microphones. Matrix $\mathbf{R}$ is termed the noise spatial correlation matrix, and its general definition is

$$\mathbf{R} = E\left[\mathbf{x}_N(\omega,t)\mathbf{x}_N^H(\omega,t)\right], \quad (3)$$

where $\mathbf{x}_N(\omega,t)$ is the input vector generated by only the noise sources.

In the next sections, a method of obtaining the information required for constructing the beamformer coefficient vector, namely, $\mathbf{a}$ and $\mathbf{R}$, is proposed.

3.1. Estimation of the Steering Vector $\mathbf{a}$

It is difficult to estimate the steering vector directly from the data in a block including overlapping speech. Such a block to be separated is termed the current block for the sake of convenience. The steering vector is estimated from the blocks before or after the current block in which the target speaker is *solely* speaking, and the location of the target is the closest to that in the current block.

First, the blocks in which the target speaker is speaking are identified using the information of the speech events in the previous section as depicted in Fig.4. Next, the blocks in which the target speaker is solely speaking are identified. For identifying the events with only one speaker, the following index is used:

$$E = \frac{1}{N_\omega}\sum_{\omega_L}^{\omega_H}\frac{\lambda_1(\omega)}{\lambda_2(\omega)} \quad (4)$$

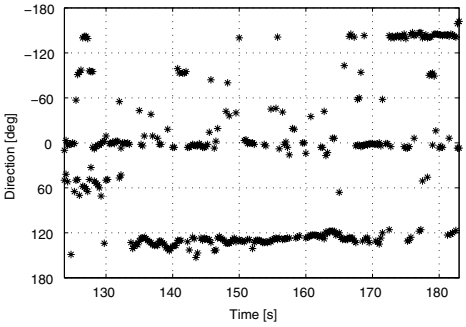
Here, $\lambda_1(\omega)$ and $\lambda_2(\omega)$ are the largest and the second-largest eigenvalues of the spatial correlation of the input $\mathbf{R}_I(\omega) = E\left[\mathbf{x}(\omega,t)\mathbf{x}^H(\omega,t)\right]$. The symbols $[\omega_L,\omega_H]$ and N_ω denote the range of frequencies of interest and the number of frequencies, respectively. It is known that the number of dominant eigenvalues of the spatial correlation matrix corresponds to the number of effective sound sources (e.g., [5]). When there is a single effective source, a single eigenvalue becomes dominant. When there are two or more sources, on the other hand, there are two or more dominant eigenvalues. Thus, in the blocks with only one speaker, it is expected that this index E becomes large. The blocks in which E is larger than a certain threshold are identified as the block with only one speaker. The identified set of blocks is denoted as Ψ.

Next, an optimal block for estimating the steering vector is determined from the set of blocks Ψ. As a criterion for selecting the optimal block, the following difference in the estimated source direction is used:

$$\tilde{n} = \arg\left[\min_{n\in\Psi}(|\theta_n - \hat{\theta}|)\right] \quad (5)$$

Here, $\tilde{n}$ is the selected block number. The symbol $\hat{\theta}$ is the estimated direction of the target source in the block to be separated (current block). The direction θ_n is the one estimated in the nth candidate block in Ψ. Since both of the estimated directions, $\hat{\theta}$ and θ_n, are digitized, the candidates selected using (5) may not be unique. When several candidates are selected, a single candidate

(a) Peaks in spatial spectrum in every block



(b) Detected speech events

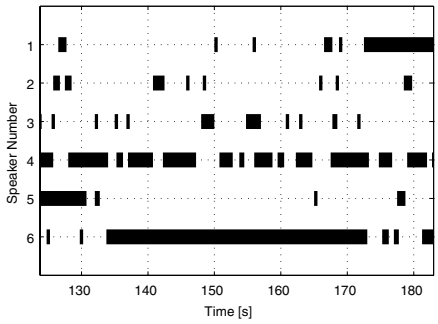


Figure 3: An example of detected speech events.

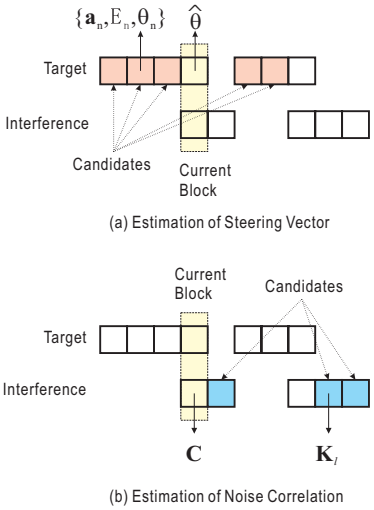


Figure 4: Estimation of the steering vector and the noise correlation.

is further selected by using the index E as follows:

$$\hat{n} = \arg\left[\max_{n\in\tilde{\Psi}} E_n\right] \tag{6}$$

Here, E_n denotes the index E at the nth candidate. The symbol $\tilde{\Psi}$ denotes the set of blocks selected by (5).

Once the optimal block is determined, the steering vector can be estimated by extracting the eigenvector of the correlation matrix corresponding to the largest eigenvalue. This is because the subspace spanned by the steering vectors and the eigenvectors corresponding to the dominant eigenvalues are identical (signal subspace) [5]. When there is only a single source, the dimension of this subspace becomes one, and it is obvious that the direction of the steering vector for the target speaker and that of the eigenvector corresponding to the single dominant eigenvalue are identical.

3.2. Estimation of the Noise Spatial Correlation R

Since $\mathbf{x}_N(\omega, t)$ cannot be observed separately in the current block, the ideal noise correlation $\mathbf{R}$ is also not available. In a manner similar to the estimation of the steering vector, the noise correlation is estimated from the blocks before or after the current block. First, the blocks in which the overlapping speaker (noise source) is speaking and the target speaker is *not* speaking are selected based on the information of the speech events as depicted in Fig.4. The set of the spatial correlation calculated in these blocks is denoted as $\Phi = [\mathbf{K}_1, \cdots, \mathbf{K}_L]$. When the noise correlation selected from these candidates has spatial characteristics close to that of the noise in the current block, the beamformer becomes an approximation of the maximum likelihood (ML) adaptive beamformer.

In addition to the set of the candidates Φ, two other candidates of the noise correlation are taken into account to enhance the performance of the separation and the speech enhancement. The first one is the identity matrix $\mathbf{I}$, which is the theoretical noise correlation when the noise is spatially white. A beamformer using $\mathbf{I}$ is termed a delay-and-sum (DS) beamformer. Even when the target speaker is solely speaking, there is room reverberation that reduces the performance of ASR. By applying this beamformer in the section with only one speaker, the effect of speech enhancement is expected.

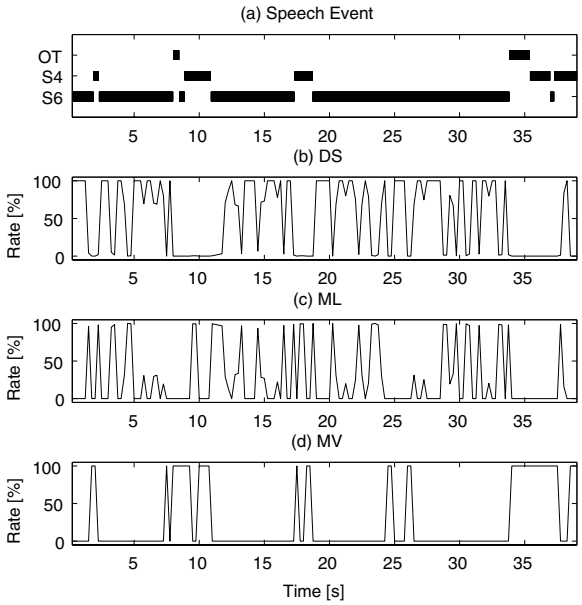


Figure 5: Selected beamforming algorithm.

Another candidate is the correlation calculated in the current block. This correlation is denoted as $\mathbf{C}$. The correlation $\mathbf{C}$ differs from the ideal noise correlation $\mathbf{R}$ since not only the noise but also the target signal is included in $\mathbf{C}$. When the noise is dominant in the current block, however, $\mathbf{R} \simeq \mathbf{C}$, and the noise is effectively reduced since the characteristics of noise used in the beamformer perfectly match those of the current block. When the level of the target is comparable to or larger than that of the noise, however, the noise reduction performance is lower than that of the ML beamformer. When $\mathbf{C}$ is employed, the beamformer is termed a minimum variance (MV) beamformer.

For selecting the noise correlation from the candidates described above, a criterion similar to that used in the MV beamformer, i.e., the output power of the beamformer in the current block, is used as follows:

$$\hat{\mathbf{R}} = \arg\min_{\mathbf{R}\in\Phi,\mathbf{I},\mathbf{C}} \mathbf{w}^H\mathbf{C}\mathbf{w} \tag{7}$$

$$\text{where } \mathbf{w} = \frac{\mathbf{R}^{-1}\hat{\mathbf{a}}}{\hat{\mathbf{a}}^H\mathbf{R}^{-1}\hat{\mathbf{a}}} \tag{8}$$

In (7), $\mathbf{w}^H\mathbf{C}\mathbf{w}$ represents the output power of the beamformer. As a steering vector in the beamformer coefficient vector $\mathbf{w}$, the one selected in the previous subsection, $\hat{\mathbf{a}}$, is used.

Figure 5 shows an example of the selection of the beamformer from the above three types of beamformers, namely, DS, ML, and MV. In this example, speaker #6 is the target, while speaker #4 is the overlapping speaker. The symbol "OT" denotes the other speakers. When the target speaker was solely speaking, the DS beamformer was mostly selected. When the overlapping speaker was speaking, the ML or MV beamformer was selected. When the noise is dominant such as the "OT" event at the time $t = 35$ s (sound of a cough by another speaker), the MV beamformer was selected. Therefore, it can be seen that three types of beamformers

Figure 6: Microphone array used as an input device.

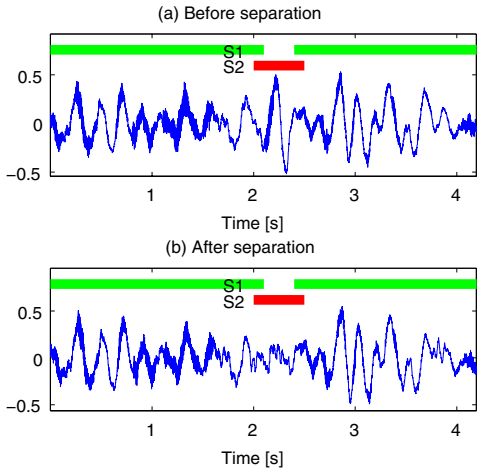


Figure 7: An example of speech event separation.

were selected by (7) according to the situation.

3.3. Filtering

Using the estimated steering vector $\hat{\mathbf{a}}$ and the noise correlation $\hat{\mathbf{R}}$, the beamformer coefficient vector $\mathbf{w}$ is updated in every block using (2). The microphone array inputs are then filtered by the updated coefficient vector using (1). In actual filtering, the beamformer coefficient vector $\mathbf{w}$ is inverse-Fourier-transformed into the time domain, and (1) is conducted in the time domain.

4. Experiment

Figure 6 shows a microphone array (with a camera array (Ladybug 2, PointGray Research)) used for the recording of the meeting in the experiment. The microphone array is circular with a diameter of 15 cm, and consists of eight microphones. The sampling frequency was 16 kHz. The recording was conducted in an ordinary meeting room with a reverberation time of around 0.5 s. The number of participants was six. The distance between the microphone array and the participants was 1.0-1.5 m.

Figure 7 shows an example of the separation of speech events. In this example, target speaker #1 was speaking while overlapping speaker #2 inserted a response "I see" in Japanese in the period of [2.0 2.5] s as shown by the bars in Fig.7. By comparing Fig.7(a) and (b), it can be seen that the response inserted by speaker #2 was reduced while the speech by speaker #1 remained intact.

Next, speech signals processed by the proposed method were evaluated using ASR. Speech events with durations of more than five seconds were selected from a single meeting recording and were fed into the ASR system. The number of the selected speech events were 367. Among the selected speech events, 100 events were used for adaptation of the acoustic model, while the remaining 267 events were used as test materials.

Subject	N	Input	Processed	Improvement
Overall	37970	44.55	53.18	8.63
S1	1793	39.88	39.43	-0.54
S2	6694	46.28	45.64	-0.64
S3	7389	53.12	53.90	0.78
S4	15818	39.63	54.74	15.11
S5	1572	38.42	52.48	14.06
S6	4704	49.00	62.99	13.99

Table 1: Phoneme accuracy [%]. "N" indicates the number of phonemes to be recognized. "Input" and "Processed" inditate the score for the microphone input and that for the signal processed by the proposed method, respectively.

Table 1 shows the results of evaluation using ASR. The phoneme accuracy, which mainly reflects the acoustic aspects of the test material, is shown. The overall improvement resulting from introduction of the proposed method was 8.63 %. From the individual score, it can be seen that the scores of S4, S5 and S6 were improved by more than 10 %, while little or no improvement was made for those of S1, S2 and S3. One of the reasons for this small improvement for some participants is considered to be the failure in finding candidates for the steering vector due to the small movement of the target speaker. This error can be improved by improving the tracking accuracy of the target speaker in a future study.

5. Conclusion

In this paper, overlapping speech events in a meeting recording were separated based on information on detected speech events. From the results of evaluation using ASR, it was shown that the phoneme accuracy was improved by approximately 8% by using the proposed method.

6. Acknowledgment

This work was partly supported by JSPS KAKENHI(A) 18200007.

7. References

[1] J. Ajmera, *et al.*, "Clustering and segmenting speakers and their locations in meeting," in *Proc. ICASSP 2004*, 2004, vol. I, pp. 605–608.

[2] T. Hain, *et al.*, "Transcription of conference room meetings: an investigation," in *Proc. Interspeech 2005*, 2005, pp. 1661–1664.

[3] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag*, vol. AP-34, no. 3, pp. 276–280, March 1986.

[4] F. Asano, *et al.*, "Detection and separation of speech event using audio and video information fusion and its application to robust speech interface," *EURASIP Jounal on Applied Signal Processing*, , no. 11, pp. 1727–1738, 2004.

[5] D. Johnson and D. Dudgeon, *Array signal processing*, Prentice Hall, Englewood Cliffs NJ, 1993.