Visualization of Conversation Flow in Meetings by Analysis of Direction of Arrivals and Continuousness of Utterance

Michiaki Katoh, Yuya Sugimoto, Shigeki Miyabe, Shoji Makino, Takeshi Yamada and Nobuhiko Kitawaki

University of Tsukuba, Tsukuba, Japan
E-mail: michy@mmlab.cs.tsukuba.ac.jp Tel: +81-29-853-6564

Abstract: In this paper, for providing a conversation flow as a visualization of entire a meeting, we propose a method for detecting intermissive and impulsive utterances from meeting recordings by discrete Fourier transform of block-series MUSIC scores. We explain why detecting intermissive and impulsive utterance in the meeting is important. We evaluate the method using meeting recordings in a real environment. Through the experimental result, we demonstrate that the method is efficient to display intermissive and impulsive utterances in the compressed representation of long meetings.
Key words: Meeting recording, Conversation flow, Compressed representation, Intermissive and impulsive utterance, MUSIC, Filter-bank analysis, Support vector machine

1. Introduction:

Recently, various technologies about meeting recordings are well progressed respectively, e.g., speaker indexing by estimating direction of arrival (DOA) [1,2,3], detection of interaction between participants from audios and movies [4], spatial communication framework for efficient transmission and remixing in teleconferences[5]. Araki *et al.* [1] presented a speaker indexing system with speech enhancement with a small number of microphones. Pardo *et al.* [2] realized a speaker dialization method by only between-microphone difference and Anguera *et al.* [3] also realized using multiple microphone in meetings by classic acoustic beamforming method with several algorithms. Otsuka *et al.* [4] presented automatic detection of cross-modal nonverbal interaction events from gazes, head gestures and utterances. MPEG is developing a standard for an efficient transmission method recorded in environments with multiple audio objects [5].

We considered to share meeting recordings attached speaker indexes and various information of events detected by such techniques with lots of people. For Example, we presented a recording system of meeting by a microphone array and a playback system with the speaker index [6]. The playback system has the speaker index with results of speech recognition, so we can use the index to playback efficiently e.g. playback focused only on some persons mainly uttered in meetings.

As an expansion of our system, we propose *conversation flow*, a new visualization framework of whole the meeting to aid the playback of recorded meeting. This system ease the user to overview the recorded meeting and find out the conversation of interest. The visualization is based on a compressive representation of whole the meeting conversation by classification of utterance, i.e., *continuous*, *intermissive* and *impulsive*, in addition to the spatio-temporal information of speakers by DOA estimation. By such visualization, the user can view *when, how long and by whom the utterance spoken* even if the meeting is as long as few hours. Intermissive and impulsive utterances shown in the conversation flow often include back-channel feedbacks e.g. “Ah, ah,” or “Wow.” Since seeing who gives the back-channel feedbacks leads to understanding of who takes main role in the conversation, we detect and emphasize them in the visualization. By showing the time flow of DOAs and classes of utterances, the conversation flow can simulataneously provide two perspectives of understanding, each speaker’s context of utterances and interactions between speakers.

In this paper, we propose a method to detect continuous, intermissive and impulsive utterances from meetings recorded by a microphone array. Our ideas of the method to be proposed are to take notice of time-series transition of DOA expressed by MUSIC score [7] to quantify and classify continuousness of utterances by Fourier analysis. In Fourier analysis, we obtain Fourier spectrum as *continuous feature spectrum*, then apply filter bank analysis of the feature spectrum to obtain *bank features*. Then we classify the bank features by *similarity vector* into three classes, continuous, intermissive and impulsive.

In the section II, we describe MUSIC method expanded into frequency domain [8] to calculate DOA as spatial spectrums and Fourier Analysis (FA) methods utilize to the spatial spectrums for detecting continuous, intermissive and impulsive signals. Results of compressed representation are to be shown in the section III as conversation flow by experiment using real recording.

2. Proposed Method:

2.1. Calculation of MUSIC Score

Let us suppose observed signals of the microphone array as $\boldsymbol{x}(t) = [x_1(t), x_2(t), \cdots, x_M(t)]^{\mathrm{T}}$ where M is number of microphones and t is time. Now we consider complex expression of the observed signals using short-term Fourier transform (STFT) as:

$$\boldsymbol{x}(\omega, r) = [x_1(\omega, r), x_2(\omega, r), \cdots, x_M(\omega, r)]^H, \quad (1)$$

where ω is angular frequency, r is STFT frame index and $x_m(\omega, r)$ is short-term complex spectrum of the signal observed at the *m*-th microphone.

To track the fluctuating DOAs, we devide the observed signal frames into blocks each of them consists of K frames. Suppose the *b*-th block starts from the *i*-th frame and we define the set A_b of the frames belonging to the *b*-th frames as $A_b = \{r_i, \ldots, r_{i+K-1}\}$. We calculate the spatial correlation matrix $\boldsymbol{R}_{xx}(\omega, b)$ of the *b*-th block as:

$$\boldsymbol{R}_{xx}(\omega, b) = \sum_{r \in A_b} \boldsymbol{x}(\omega, r) \cdot \boldsymbol{x}(\omega, r)^H, \quad (2)$$

where b is a index of block, and r_i is initial STFT frame in each block. The spatial correlation matrix $\boldsymbol{R}_{xx}$ is decomposed by complex eigenvalue expansion as follows:

$$\boldsymbol{V}(\omega, b)^H \boldsymbol{R}_{xx}(\omega, b) \boldsymbol{V}(\omega, b) = \mathrm{diag}[\lambda_1(\omega, b), \ldots, \lambda_N(\omega, b), 0, \ldots, 0] + \sigma^2 I, \quad (3)$$

where $\mathrm{diag}[\lambda_1(\omega, b), \ldots, \lambda_N(\omega, b), 0, \ldots, 0]$ is a diagonal matrix consists of eigenvalues λ_i ordered in descending, $\boldsymbol{V}(\omega, b) = [\boldsymbol{v}_1(\omega, b), \ldots, \boldsymbol{v}_N(\omega, b), \ldots, \boldsymbol{v}_M(\omega, b)]^H$ is a eigenvector matrix and σ^2 is noise power.

Assuming there are *N* speech sources, observation of the speech sources exist in the span of the first *N* eigenvectors. The span of the rest is regarded as noise subspace analyzing angle of noise space, the MUSIC score $P(\omega, b, \theta)$ to indicate the existence of speech in the direction θ is given by:

$$P(\omega, b, \theta) = \frac{1}{\sum_{i=N+1}^{M} |\boldsymbol{v}_i(\omega, b)^{\mathrm{H}} \cdot \boldsymbol{a}(\omega, \theta)|^2}, \quad (4)$$

where $\boldsymbol{a}(\omega, \theta)$ represents a steering vector of the direction θ. We calculate the wide-band MUSIC score $\bar{P}(b, \theta)$ by summing the narrow band MUSIC score of the frequency range of interest from the frequency bins ω_j to ω_k as:

$$\bar{P}(b, \theta) = \sum_{\omega=\omega_j}^{\omega_k} P(\omega, b, \theta). \quad (5)$$

In this paper, we call a *block-series MUSIC score* as:

$$\boldsymbol{p}(\theta) = [P(b_1, \theta), P(b_2, \theta), \ldots, P(b_k, \theta), \cdots]. \quad (6)$$

Fig. 1 shows a typical example of block-series MUSIC score by spectrogram representation.

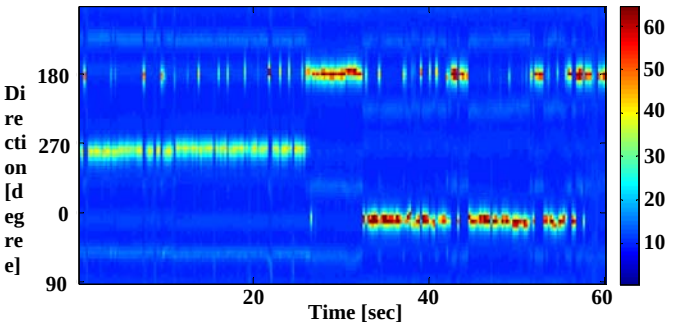

Fig.1 Typical result of block-series MUSIC score by spectrogram representation

2.2. Fourier analysis for block-series MUSIC score

To estimate the continuousness of signals, we apply Fourier analysis to each direction of the block-series MUSIC score. The problem to analyze continuousness of utterance in each direction reduces to the quantification of how often MUSIC score sequence rises and falls in a time period of certain length. Such quantification can be achieved with short-term Fourier analysis of MUSIC score sequence $\boldsymbol{P}(b, \theta)$ for each direction θ. The DFT spectrum $Q(l, s, \theta)$ is given by:

$$Q(l, s, \theta) = \left| \sum_{b \in B_s} w(b - b_k) P(b - b_k, \theta) e^{-j\frac{2\pi(b-b_k)l}{L}} \right| \quad l = 1, 2, \cdots, L \quad B_s = \{b_k, \cdots, b_{k+L-1}\} \quad (7)$$

$$w(b) = \begin{cases} \frac{1}{2}\sin\left(2\pi\frac{b - b_k}{L}\right) - \frac{1}{2} & (0 \le b \le L) \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where $Q(l, s, \theta)$ is a amplitude spectrum of DFT utilized to block-series MUSIC scores in each direction, l is a frequency index of DFT, L is a length of block that is wrapped and used by DFT, s is a wrapped block index, $w(b)$ is von Hann window, and b_i is initial block index in each DFT block. We call the amplitude spectrum the continuousness feature spectrum (CFS).

As shown in Fig. 2, each type of utterance tends to have the following spectral feature characteristics:

- Continuous utterance: having strong peak in direct current component, because of flat continuousness of MUSIC score.
- Intermissive utterance: having relatively large amplitude in direct-current and low-frequency components because of continuous large MUSIC score of medium length.
- Impulsive signal: having flat amplitude in entire frequency bands with gentle slope because of extremely short vise in MUSIC score.

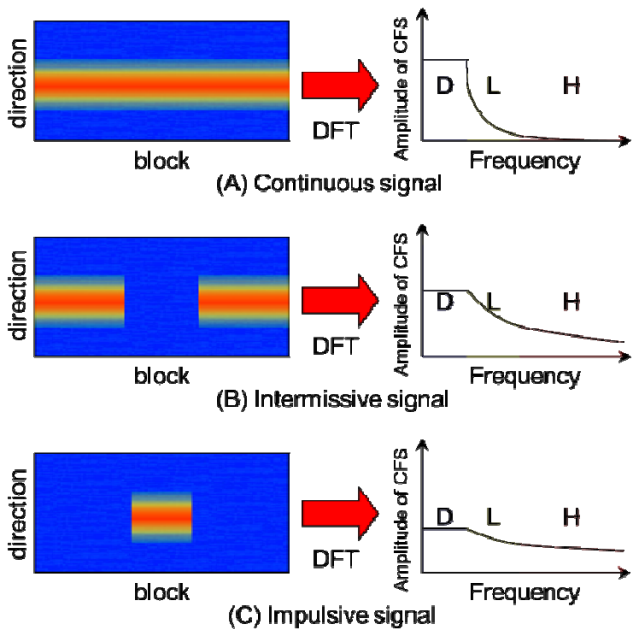

Fig. 2 Types of block-series MUSIC score (left) and expected results of CFS (right):
(A)continuous signal, (B)Intermissive signal and (C) Impulsive signal. Band D is direct current, band L is low frequency and band H is high frequency.

To capture these characteristics, we use three filter banks as feature extraction:

$$b_{\mathrm{DC}}(s,\theta) = |Q(l_{DC},s,\theta)|, \tag{9}$$

$$b_{\mathrm{low}}(s,\theta) = \sum_{l=l_{\mathrm{low}_1}}^{l_{\mathrm{low}_p}} |Q(l,s,\theta)|, \tag{10}$$

$$b_{\mathrm{high}}(s,\theta) = \sum_{l=l_{\mathrm{high}_1}}^{l_{\mathrm{high}_q}} |Q(l,s,\theta)|, \tag{11}$$

where $b_{\mathrm{DC}}(s,\theta)$ represents the values of CFS of direct component, $b_{\mathrm{low}}(s,\theta)$ represents the sum of CFS of low frequency bins and $l_{\mathrm{low}_1}$is lowest frequency bins of b_{low}, l_{low_p} is highest frequency bins of b_{low}. $b_{\mathrm{high}}(s,\theta)$ also represents the sum of CFS of high frequency bins. l_{high_1} is lowest frequency bin of b_{high} and $l_{\mathrm{high}_q}$is highest frequency bin of b_{high}. We call $b_{\mathrm{DC}}(s,\theta)$, $b_{\mathrm{low}}(s,\theta)$ and $b_{\mathrm{high}}(s,\theta)$ as filter bank feature of direct component, low frequency and high frequency respectively.

Note that distribution of the bank features depends on the length of block *L* used by DFT. So we calculate the filter bank features by various lengths of block preliminary and use a suitable set of filter bank features by our subjective. This reason is described in the following section.

2.3. Utterance classification by similarity of bank feature

With the bank features, we introduce *similarity* measure to classify the filter bank feature. A similarity vector of continuous utterance is defined as $\boldsymbol{d}_{\mathrm{cont}} = [d_{\mathrm{cont1}}\ d_{\mathrm{cont2}}\ d_{\mathrm{cont3}}]$, whose columns correspond to direct current, low frequency and high frequency components, respectively. A similarity of continuous utterance $E_{\mathrm{cont}}(s,\theta)$, intermissive utterance $E_{\mathrm{int}}(s,\theta)$ and impulsive utterance $E_{\mathrm{imp}}(s,\theta)$ are calculated as inner product of the similarity vector and the vector of the filter bank features:

$$\begin{aligned} E_{\mathrm{cont}}(s,\theta) &= \boldsymbol{d}_{\mathrm{cont}} \cdot \left[\, b_{\mathrm{DC}}(s,\theta)\ b_{\mathrm{low}}(s,\theta)\ b_{\mathrm{high}}(s,\theta)\right]^T \\ &= [d_{\mathrm{cont1}}\ d_{\mathrm{cont2}}\ d_{\mathrm{cont3}}]\left[\, b_{\mathrm{DC}}(s,\theta)\ b_{\mathrm{low}}(s,\theta)\ b_{\mathrm{high}}(s,\theta)\right]^T, \end{aligned} \tag{12}$$

$$\begin{aligned} E_{\mathrm{int}}(s,\theta) &= \boldsymbol{d}_{\mathrm{int}} \cdot \left[\, b_{\mathrm{DC}}(s,\theta)\ b_{\mathrm{low}}(s,\theta)\ b_{\mathrm{high}}(s,\theta)\right]^T \\ &= [d_{\mathrm{int1}}\ d_{\mathrm{int2}}\ d_{\mathrm{int3}}]\left[\, b_{\mathrm{DC}}(s,\theta)\ b_{\mathrm{low}}(s,\theta)\ b_{\mathrm{high}}(s,\theta)\right]^T, \end{aligned} \tag{13}$$

$$\begin{aligned} E_{\mathrm{imp}}(s,\theta) &= \boldsymbol{d}_{\mathrm{imp}} \cdot \left[\, b_{\mathrm{DC}}(s,\theta)\ b_{\mathrm{low}}(s,\theta)\ b_{\mathrm{high}}(s,\theta)\right]^T \\ &= [d_{\mathrm{imp1}}\ d_{\mathrm{imp2}}\ d_{\mathrm{imp3}}]\left[\, b_{\mathrm{DC}}(s,\theta)\ b_{\mathrm{low}}(s,\theta)\ b_{\mathrm{high}}(s,\theta)\right]^T. \end{aligned} \tag{14}$$

Using the similarities, we finally decide which a class the signal is assigned to as following condition:

$$\begin{aligned} &\textbf{if} \\ &\operatorname{argmax}\left(E_{\mathrm{cont}}(s,\theta), E_{\mathrm{int}}(s,\theta), E_{\mathrm{imp}}(s,\theta)\right) = E_{\mathrm{cont}}(s,\theta) \\ &\quad \textbf{then class} = \textbf{continuous,} \\ &\textbf{else if} \\ &\operatorname{argmax}\left(E_{\mathrm{cont}}(s,\theta), E_{\mathrm{int}}(s,\theta), E_{\mathrm{imp}}(s,\theta)\right) = E_{\mathrm{int}}(s,\theta) \\ &\quad \textbf{then class} = \textbf{intermissive,} \\ &\textbf{else if} \\ &\operatorname{argmax}\left(E_{\mathrm{cont}}(s,\theta), E_{\mathrm{int}}(s,\theta), E_{\mathrm{imp}}(s,\theta)\right) = E_{\mathrm{imp}}(s,\theta) \\ &\quad \textbf{then class} = \textbf{impulsive.} \end{aligned} \tag{15}$$

A merit of using the similarity is that the result of classification can be controlled by tuning the similarity vectors. In other words, it is possible to provide results that is suitable for user's preference with tuned similarity vectors. Thus in the experiment in the following section, we search the similarity vectors and the block length of DFT described in II.B that results of classification between intermissive utterance and impulsive utterance are suitable for our preference.

3. Applying the Methods to Real Recordings:

To verify effectiveness of the proposed method described in previous section, we perform an experiment using meeting recordings in real environment.

The experimental condition was following: we used a round microphone array put on table in conference room of 7.5 m squared. The array consisted of 8 microphones, measured 20 cm in diameter, and 10 centimeters as height. Each microphone in the microphone array was located equally on the round. A size of the table on the center of the room was 1.8 m × 1.2 m, and there were 3 participants around the table in the recording, one was an interviewer and the others were interviewees. Distance between the array and each participant was about 1.0 m. Sampling Rates of the recording microphone array was 16

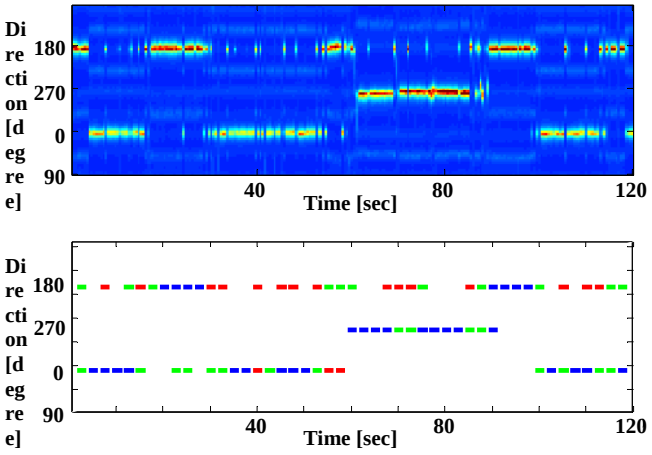

Fig. 3 Result of block-series MUSIC score within 2 minutes by spectrogram representation

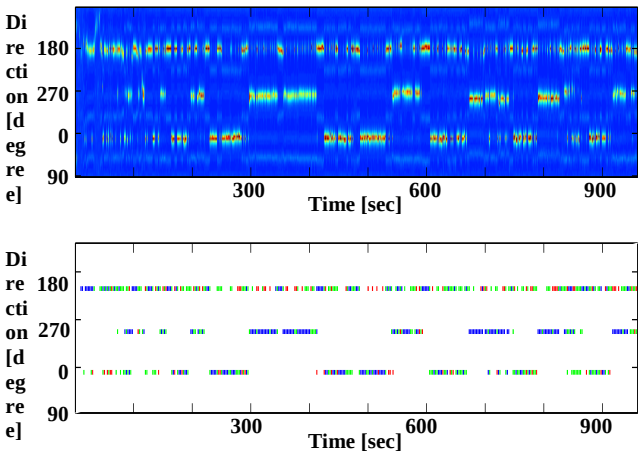

Fig. 4 Result of block-series MUSIC score within 16 minutes by spectrogram representation

We calculated block-series MUSIC scores described in II.A from the recording. Time length of a block of MUSIC was 0.5 sec. In each block STFT was executed with hamming window where number of point was 512 and frame shift was 128. Also in each block, MUSIC score was calculated according to equation (5), where frequency range ω_j to ω_k was 1 to 4 kHz.

Next, as described in II.B, we utilized DFT into L length of block in each direction of the scores shown as equation (7). A number of blocks was 10, so the number of DFT points of CFS was 6. We assigned the lowest frequency of CFS to direct current component, the next two CFSs to low frequency component and the last three CFSs to high frequency component. Then we calculated these components to bank features correspond to equation (9), (10) and (11), respectively.

We finally classified the bank features into 3 clusters; continuous, intermissive and impulsive, respectively in every wrapped block with similarity vectors. In our experiment, the similarity vector adjusted to our preference were as:

$$\begin{aligned} \boldsymbol{d}_{\mathrm{cont}} &= [1.10 \quad 0.42 \quad 0.05], \\ \boldsymbol{d}_{\mathrm{int}} &= [0.65 \quad 0.97 \quad 0.57], \\ \boldsymbol{d}_{\mathrm{imp}} &= [0.28 \quad 0.85 \quad 1.35]. \end{aligned} \tag{16}$$

In the similarity vector of continuous utterance $\boldsymbol{d}_{\mathrm{cont}}$, first elements was relatively large (1.10), second was decent (0.42) and third was small (0.05). The opposite result from $\boldsymbol{d}_{\mathrm{cont}}$ was the similarity vector of impulsive utterance $\boldsymbol{d}_{\mathrm{imp}}$, which first elements was relatively small, second was decent and third was large. In the similarity vector of intermissive utterance $\boldsymbol{d}_{\mathrm{int}}$, every elements were decent.

Fig. 3 shows input block-series MUSIC score within 2 minutes on the top and obtained a conversation flow as results of classification of feature banks at the bottom. In the bottom, we define a color bar corresponds to 3 classes we have described; blue represents continuous, green represents intermissive and red represents impulsive. Though there are some failures, e.g., the continuous utterance on the top is classified intermissive at the bottom, but from the result we can understand each speaker generally, e.g., a speaker at 270 degrees spoke within 60 to 90 seconds, talked to a speaker at 180 degrees, partially uttered intermissively.

Top of Fig. 4 shows input entire the block-series MUSIC scores within 16 minutes. On the top, we easily notice that if meeting was long, the representations of block-series MUSIC scores were clouded and intermissive or impulsive utterance were invisible partially. So providing the compressed representation we proposed as shown on the bottom of Fig. 4 are more effective if the meeting were for many hours. In the figure, some invisible parts on the top were represented exactly, e.g., a speaker around 0 degree and time were around 700 seconds.

4. Conclusion:

In this paper, we proposed the method as detecting intermissive and impulsive utterance for providing the conversation flow using Fourier analysis of MUSIC scores and calculating similarities of bank features. A merit of the method is that can provide efficient compressed representations of meetings if these are long. Our future work is to evaluate error rates of the results of the classification and efficiencies of compressed representation quantitatively.

References:

[1] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, S. Makino, “Speaker indexing and speech enhancement

in real meetings/ conversations," *Proc. ICASSP 2008*, pp.93−96, 2008.

[2] J. Pardo, X. Anguera, C. Wooters, "Speaker dialization for multi-microphone meetings using only between-channel differences," *Proc. MLMI'06*, pp.257−264, 2006, Springer.

[3] X. Anguera, C. Wooters, J. Hernando, "Acoustic bemforming for speaker dialization of meetings," *IEEE. Trans. Audio, Speech and Language Processing*, vol.15, pp.2011−2022, 2007.

[4] K. Otsuka, H. Sawada, J.Yamato, "Automatic inference of cross-modal nonverbal interactions in multiparty conversation," *Proc. ICMI2007*, pp. 255−262. 2007.

[5] J. Engdegard, *et al*, "Spatial audio object coding (SAOC) – the upcoming MPEG standard on parametric object based audio coding," *Proc. 124th Audio Engineering Society Convention*, preprint 7377, 2008.

[6] M. Katoh, *et al*, "State estimation of meetings by information fusion using bayesian network," *Proc. Interspeech 2005*, pp. 113−116. 2005.

[7] R.O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas and Propagation*, vol.34, no.3, pp. 276−280, 1986.

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