A METHOD FOR ESTIMATING THE GROUPING OF PARTICIPANTS IN CLASSROOM GROUP WORK USING ONLY AUDIO INFORMATION

Osamu Ichikawa[1]*, Yuuto Shima*[1]*, Takahiro Nakayama*[2]*, Hajime Shirouzu*[3]

[1]Shiga University, Japan
[2]The University of Tokyo, Japan
[3]National Institute for Educational Policy Research, Japan

ABSTRACT

This paper proposes a novel method for estimating which microphone belongs to the same group in a situation where there are multiple discussion groups in one room, using only audio information. The assumption is that each member wears one close-talk microphone, and that the audio is recorded on their own audio track. Each microphone records the main speech of the associated speaker, as well as the speech of neighboring others that have leaked into the microphone. If the neighboring speaker's leaked speech is coming in clearly, the neighboring speaker can be in proximity. An undirected network is constructed with speakers as nodes and the degrees of leaked speech as the edge weights. Given a target number of discussion groups in a room, network clustering can be applied to obtain subgroup information about which audio tracks belong to the same group. An evaluation experiment was conducted using audio data recorded in workgroup classes at the actual junior high school. In this experiment, the average Rand index of the grouping was 0.995, confirming that practical accuracy can be obtained.

Index Terms— Multi-channel signal processing, spectral clustering, correlation analysis, graph

1. INTRODUCTION

Active learning, in which students think and learn on their own, has been proposed as a new form of education. The "knowledge constructive jigsaw methodology," promoted by Consortium for Renovating Education of the Future (CoREF) promotes heuristic learning by dividing the students in a classroom into small groups of three to seven for mutual discussions. To visualize this learning process, each student is fitted with a close-talk microphone to record what they speak, and subsequently, ASR is used to convert their speech data into text [1][2]. By observing the interactions between participants in a group, we can learn how students gain understanding and discover further questions.

Because there were multiple discussion groups in one classroom, the use of far-field microphones arrays placed on the centers of desks is not recommended. Therefore, the use of close-talk microphones has become the de facto standard in this field. In that form, a single audio track is associated with a corresponding single participant. Since the analysis of the interactions between participants is done on a group basis, the recorded tracks need to be linked to the respective groups.

Managing grouping information has become a major workload for teachers. For example, the recording device is assumed to be an IC recorder or tablet PC. In that case, the students can use a recording device with the same number as in the class membership list. Since the students will belong to one of the groups, the teachers should prepare a table of which students (recorders) belong to which group before starting the class. However, even if the grouping is prepared in advance, if there is an imbalance in the number of groups due to a sudden absence of students, etc., it is quite possible to order a temporary move. To reduce the workload of managing the grouping information, it is required to provide a function to automatically estimate which students (recorders) were in the same group after recording.

2. CONVENTIONAL TECHNOLOGY

Instead of estimating groupings, methods for estimating the location of speakers and microphones have been actively researched. Using a microphone array placed in the center of the desk, the position and direction of the speaker can be estimated by the MUSIC [3][4] method or other methods. Research is also being conducted to estimate the location of sound sources from data collected by multiple microphones distributed on a desk [5][6]. Speech separation is also explored [7-9]. However, it cannot work in the case where a speaker with a directional microphone attached to his/her mouth turns his/her face in various directions, as in this study.

One possible solution would be to record a reference voice, such as a teacher's voice or a chime, at a high level on any of the recorder's microphones and estimate the distance from the source by measuring the time difference between the recorders. If there are multiple reference voices and their physical locations are known, the location of the microphones on the recorders can be estimated. There is a study combining this technology with watermarking [10]. If the locations of the speakers are determined, we can presume that the people nearby are in one group. However, this method assumes that the times of multiple recorders are perfectly synchronized, which is difficult to achieve with IC recorders that are manually switched on. It could be possible to use special software to synchronize the start of recording, but the problem is that the time synchronization gradually slips due to subtle differences in the clock frequency in the equipment. There is a technique to correct it sequentially [11], but it is difficult to compensate for the time synchronization of devices located far enough apart in a classroom that they do not mix with each other's audio.

Previous studies focusing on sound source localization often implicitly assume that there is only one group in a room and cannot cover the case of multiple discussion groups in a single classroom, as in this study.

In the following explanation, students are referred to as speakers, and the audio data from the recording device (microphone) associated with one speaker is referred to as the audio track. The challenge is to estimate the group to which a speaker belongs while the speaker and the audio track are tied together. This may be accomplished by linguistic or acoustic approaches.

In the linguistic approach, the speech tracks of all speakers are converted to text by automatic speech recognition, and groupings are estimated by accumulating pairs of utterances that match the context of the texts. In discourse analysis, the context is discussed as cohesion and coherence [12-14]. However, it is very difficult to judge the context because everyone from different groups may be discussing the same topic. Also, since most utterances are short, only a few words, it is difficult to track the complete context. Also, it should be noted that there may be taciturn people who never say a word, and it is almost impossible to estimate which group they belong to, in the linguistic approach.

Therefore, the acoustic approach is explored in this paper.

3. PROPOSED METHOD

The method proposed in this paper focuses on the "leaked speech" of others that are mixed into each person's microphone. It is expected that each audio track contains the target speaker's voice at a sufficient volume. On the other hand, it has been found that the voices of the surrounding speakers also come in at a very low volume. This is called leaked speech. Even with a noise-canceling close-talk microphone, leakage cannot be completely eliminated. Surrounding speakers vary in loudness, but in general we could say that nearby speakers produce loud leaked speech, while distant speakers produce small leaked speech. In other words, the volume of the leaked speech of speakers in the group is expected to be relatively high, while the volume of the leaked speech of speakers outside the group is expected to be low.

In the above, the degree of leaked speech was explained as volume, however it can be measured stably by correlation, especially in a noisy space such as a classroom. In this paper, we employed the cross-spectrum phase analysis (CSP) [15], which can sensitively detect correlations without depending on the volume or tone. In other words, the CSP coefficient between the leaked speech and its source speech are calculated and used as a metric of the distance between the speakers. Compute this for all speaker pairs and set it as an adjacency matrix as in Figure 1.

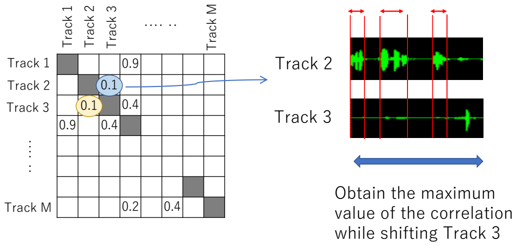

Fig. 1 Correlation analysis of audio tracks.

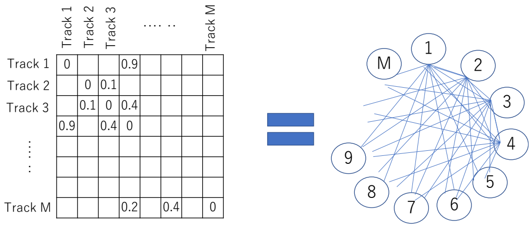

Fig. 2 Constructing undirected graph.

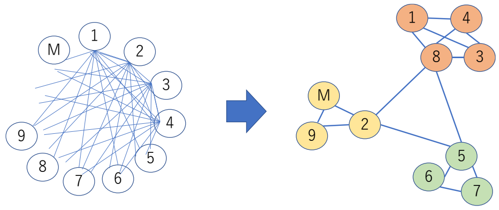

Fig. 3 Image of graph clustering.

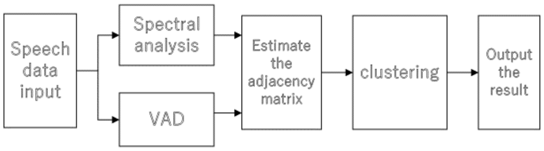

Fig. 4 Block diagram of the proposed method.

An undirected graph is constructed by using speakers as nodes and giving the weights (edges) between nodes with reference to the adjacency matrix as shown in Figure 2. By performing graph clustering and estimating the subgraphs within it, we can estimate multiple subgroups connected by large weights of edges as shown in Figure 3.

Note that the target number of clusters should be given in the above graph clustering. This is because it is natural for the teacher to know the total number of groups. Also, if the target number of clusters is not given, the following problems might be expected. Let's consider one speaker who speaks very loudly. His/her speech will be mixed into everyone else's microphones at a high level. Then, it may happen that the speaker acts as a hub connecting multiple clusters, and as a result, only one cluster can be estimated. Given a target number of clusters, such speaker can also be expected to belong to one cluster with the highest affinity.

The block diagram of the proposed method is shown in Figure 4. The details of each function are described below.

3.1. Spectral Analysis

The input data is the PCM data (audio waveform data) of M tracks, which is stored on the computer. In our experiment, the audio data was recorded at a sampling frequency of 16 kHz. Each audio track is windowed with a frame shift of 10 ms and the complex spectrum is obtained using a 512-point Discrete Fourier Transform (DFT). The complex spectrum of the t-th frame of the m-th audio track is denoted by $X(m,k,t)$. k is the index corresponding to the frequency.

3.2. VAD

For each audio track, a flag is determined for each frame as to whether the speaker associated with that track is speaking or not, and set as Voice Activity Detection (VAD) information $V(m,t)$. Here, a value of 1 indicates that the m-th speaker is speaking at frame t, and a value of 0 indicates that the speaker is not speaking. The determined speech segment will be used as a segment to detect leaks to other audio tracks.

There are two main types of VAD techniques: model-based methods [16] that use MFCCs and spectral features, and power-based methods [17][18] that use speech power as a feature. In the assumed scenario, the voice of the speakers next to the subject speaker will also be mixed into the microphone, so we used a power VAD with stricter judgment criteria. In other words, the threshold should be set high enough so that the neighboring speaker's speech is not judged as the target speaker's speech. Weak utterances of the target speaker may be dropped by the setting, this is not a problem.

3.3. Adjacency Matrix Estimation

In this part of the process, the CSP coefficients for each pair of audio tracks are obtained for all the M×(M-1)/2 possible pairs of audio tracks. As the CSP coefficients have dimensions corresponding to the time shift, the largest value in the time dimension is selected and set to the corresponding part of the adjacency matrix. In other words, when the adjacency matrix is α and the indices of the two specified audio tracks are p and q, assign the same value to $\alpha(p,q)$ and $\alpha(q,p)$.

To calculate the CSP coefficients, it is first necessary to decide which of the two specified audio tracks p and q should be the primary track and which should be the secondary track. The primary and secondary correspond to the direction of speech "leakage". We consider that the speech uttered by the speaker of the primary track is mixed (leaked) into the speech of the secondary track. Leakage essentially occurs in both directions, but it is necessary to consider that there are also silent students in groups. Therefore, the speech of the student with the longer speech segments is treated as the primary track, while the speech of the more reticent student is treated as the secondary track.

Therefore, if the Equation (1) holds, p should be the primary track and q should the secondary track.

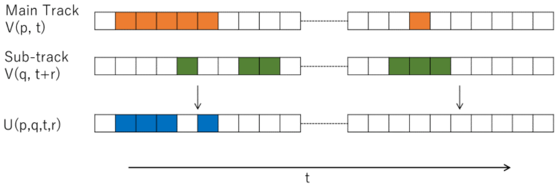

Fig. 5 Configuring segments for leak detection.

$$\sum_t V(p,t) \geq \sum_t V(q,t). \tag{1}$$

The CSP coefficients are obtained as a vector of the same dimension as the input width of the Discrete Fourier Transform. This dimension corresponds to the time in sampling units and represents the time difference in synchronization between the two inputs. However, it should be noted that the time difference of synchronization identified by CSP is only within the width of a single audio frame. Since it is necessary to consider the possibility of temporal misalignment beyond the frame between the two audio tracks, we shift the frame of one of the tracks in the range of -R to +R to find the maximum CSP coefficient. The value of R can be given arbitrarily depending on the recordings.

A new metric representing the entire tracks is obtained as an average of the maximum CSP coefficients over the segments U that is the leak detection target, where the speaker of the primary track is speaking, and the speaker of the secondary is not.

If p is the primary track and q is the secondary track, then U is given by Equation (2).

$$U(p,q,t,r) = V(p,t)(1 - V(q,t+r)) \tag{2}$$

where r is the frame shift varying from -R to +R (see Figure 5).

The CSP coefficient ϕ for frame t with frame shift r is obtained by Equation (3). IDFT is an inverse discrete Fourier transform, where * denotes complex conjugate.

$$\phi(p,q,t,r) = IDFT\left[\frac{X(p,k,t)X(q,k,t+r)^*}{|X(p,k,t)||X(q,k,t+r)|}\right]. \tag{3}$$

ϕ is the vector, whose dimension is mapped to the time axis by IDFT. Let C be the maximum value of elements of ϕ. C represents the degree of correlation between both tracks for the frame t.

$$C(p,q,t,r) = \max_d \phi_d(p,q,t,r) \quad . \tag{4}$$

The average value of C is obtained by collecting the C of the frames in the segments where the leak detection is made. This is named $\acute{C}$.

$$\acute{C}(p,q,r) = \frac{\sum_t C(p,q,t,r)U(p,q,t,r)}{\sum_t U(p,q,t,r)} \quad . \tag{5}$$

$\acute{C}$ is the CSP coefficient value when the frame is shifted by r. For this variable $\acute{C}$, find the maximum value C_{max} by varying r from -R to R.

$$C_{max}(p,q) = \max_r \acute{C}(p,q,r). \tag{6}$$

Thus, this is the maximum value of the CSP coefficient set to the adjacency matrix α as

$$\alpha(p,q) = \alpha(q,p) = \begin{cases} C_{max}(p,q) & ,p \neq q \\ 0 & ,p = q \end{cases} . \qquad (7)$$

3.4. Clustering

This process uses the adjacency matrix estimated in the previous step to cluster the audio tracks into multiple subgroups. There are various methods for network clustering, but here we will use spectral clustering [19], which allows us to specify the target number of clusters.

For the adjacency matrix α, $\alpha(p,q)$ is the edge weight that links node p to node q. They correspond to speaker p to speaker q. The matrix α is a symmetric matrix and the diagonal terms are zeros. Since the number of speakers is M, the matrix has $M \times M$ dimensions.

In spectral clustering, the graph Laplacian φ is first estimated from the adjacency matrix α. Here we use the unnormalized graph Laplacian matrix.

$$\varphi = \nu - \alpha \qquad (8)$$

where the order matrix ν is given by

$$\nu(p,q) = \begin{cases} \sum_{\acute{q}} \alpha(p,\acute{q}) & ,p = q \\ 0 & ,p \neq q \end{cases} . \qquad (9)$$

Next, find the eigenvalues of the graph Laplacian φ, select N eigenvalues in order of decreasing eigenvalue, and find N corresponding eigenvectors. Let them be $\gamma_1, \gamma_1, \cdots, \gamma_N$. These vectors are aligned in the column direction to form the matrix Γ. This will be a matrix with M-dimensional rows and N-dimensional columns.

$$\Gamma = (\gamma_1, \gamma_1, \cdots, \gamma_N) \ . \qquad (10)$$

Next, take the matrix Γ in the row direction and construct M N-dimensional vectors. These are denoted as $\rho_1, \rho_1, \cdots, \rho_M$. Each of them corresponds to M speakers.

$$\Gamma^T = (\rho_1, \rho_1, \cdots, \rho_M). \qquad (11)$$

This can be divided into N groups using a vector clustering method such as k-means.

4. EVALUATION

4.1. Data

For the evaluation, we used the real audio data recorded in group work classes in the actual junior high school working with CoREF. The classrooms were furnished with woods. It is the standard size of a Japanese junior high school. The desks were about 60 cm wide. The desks of students in the same group were closely arranged. The closest distance between students of different groups was about 1 meter, as interpreted from the pictures. There were two group work sessions in one class, the first half is called Expert session and the second half is called Jigsaw session. The length of the former is about 5 to 8 minutes, and the latter about 20 to 24 minutes. Three classes (mathematics, science, and Japanese) were evaluated. We measured the performance of automatic grouping for a total of six sessions.

4.2. Experimental Results

Examples of the results of automatic grouping are shown in Figure 6 and Figure 7. Each node in the figure has an ID such as "A-06" or "01-A" (corresponding to the audio track name of the speaker), so you can read from the ID whether it belongs to the correct group or not.

For each session, we calculated the Rand index, which is a metric of the accuracy of the grouping. Table 1 shows the results. The grouping error occurred only once in all six sessions, where it resulted in one node belonging to the wrong group.

5. CONCLUSION

Assuming that there are multiple discussion groups in a classroom, we have proposed a novel method to estimate the grouping using only audio information, recorded by the microphones attached to each student. An evaluation experiment was conducted using audio data recorded in actual junior high school classes, and an average Rand index of 0.995 was obtained. This means that out of the six sessions, only one person was misassigned. It was verified that the method has a practical level of accuracy.

6. ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI (Grant Numbers JP17H06107 and JP19K02999).

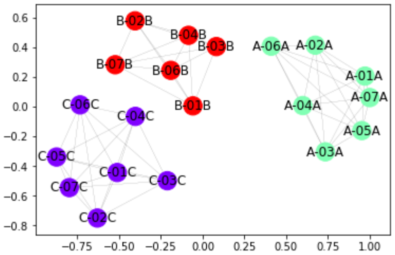

Fig. 6 Results of automatic grouping in Expert Session of the science class.

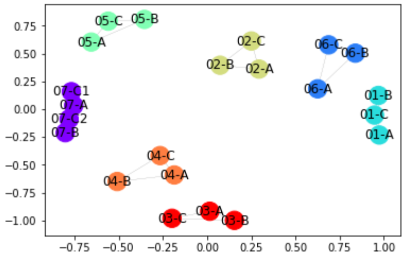

Fig. 7 Results of automatic grouping in Jigsaw Session of the mathematics class.

Table 1 Experimental Results

Class	Session	Num. groups	Rand index
Mathamatics	Expert	7	1.00
	Jigsaw	7	1.00
Science	Expert	3	1.00
	Jigsaw	6	1.00
Japanese	Expert	6	0.97
	Jigsaw	6	1.00

7. REFERENCES

[1] H. Shirouzu, M. Saito, S. Iikubo, T. Nakayama, and K. Hori, “Renovating Assessment for the Future: Design-Based Implementation Research for a Learning-in-Class Monitoring System Based on the Learning Sciences,” *Proceedings of International Conference of the Learning Sciences (ICLS) 2018*, pp. 1807-1814, Jul. 2018.

[2] "IBM Cloud API Docs / Speech to Text", *https://cloud.ibm.com/apidocs/speech-to-text*, Last updated: 2021-06-11.

[3] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276-280, 1986.

[4] F. Asano, H. Asoh and T Matui, "Sound Source Localization and Separation in Near Field," *IEICE Transaction*, vol. E83-A, No. 11, pp. 2286-2294, 2000.

[5] F. Jacob, J. Schmalenstroeer and R. Haeb-Umbach, "Microphone Array Position Self-Calibration from Reverberant Speech Input," *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*, pp. 1-4, 2012.

[6] K. Tanaka, Y. Wakabayashi, N. Ono and R. Miyazaki, "Multiple sound source localization with iterative updates of DOA permutations in distributed microphone arrays," *Proceedings in Spring Meeting of Acoustical Society Japan 2021* (in Japanese), 3-1-19, 2021.

[7] K. Ochi, N. Ono, S. Miyabe, and S. Makino, "Multi-talker Speech Recognition Based on Blind Source Separation with Ad Hoc Microphone Array Using Smartphones and Cloud Storage," *Proceedings of Interspeech 2016*, pp. 3369-3373, 2016.

[8] H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada, and S. Makino, "Amplitude-based Speech Enhancement with Nonnegative Matrix Factorization for Asynchronous Distributed Recording," *Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC) 2014*, pp. 203-207, Sep. 2014.

[9] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Meeting Recognition with Asynchronous Distributed Microphone Array Using Block-Wise Refinement of Mask-Based MVDR Beamformer," *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018*, pp. 5694-5698, 2018.

[10] Y. Nakashima, R. Tachibana and N. Babaguchi, "Watermarked Movie Soundtrack Finds the Position of the Camcorder in a Theater," *IEEE Transactions on Multimedia*, vol. 11, no. 3, pp. 443-454, 2009.

[11] S. Araki, N. Ono, K. Kinoshita and M. Delcroix, "Meeting Recognition with Asynchronous Distributed Microphone Array," *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) 2017*, pp. 32-39, 2017.

[12] T. Nomoto and Y. Matsumoto, "Discourse parsing: a decision tree approach," *Proceedings of the 6th Workshop on Very Large Corpora*, pp. 216-224, 1998.

[13] D. Cristea, N. Ide, D. Marcu and V. Tablan, "An empirical investigation of the relation between discourse structure and co-reference," *Proceedings of the 18th conference on Computational linguistics (COLING 2000)*, vol. 1, 2000.

[14] J. Li, E. Hovy, "A model of coherence based on distributed sentence representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2039-2048, 2014.

[15] M. Omologo and P. Svaizer, "Acoustic Event Localization Using a Crosspower-Spectrum Phase Based Technique," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '94.)*, pp. 273–276, 1994.

[16] J. Sohn, N.S. Kim and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1-3, Jan. 1999.

[17] M. H. Moattar and M. M. Homayounpour, "A simple but efficient real-time Voice Activity Detection algorithm," *Proceedings of European Signal Processing Conference (EUSIPCO) 2009*, pp. 2549-2553, 2009.

[18] O. Ichikawa1, K. Nakano, T. Nakayama and H. Shirouzu, "Multi-Channel VAD for Transcription of Group Discussion," *Proceedings of Interspeech 2021*, pp. 336-340, 2021.

[19] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, 2007.