INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Multi-Channel VAD for Transcription of Group Discussion

Osamu Ichikawa[1], *Kaito Nakano*[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
osamu-ichikawa@biwako.shiga-u.ac.jp, s5017216@st.shiga-u.ac.jp,
nakayama@coref.u-tokyo.ac.jp, shirouzu@nier.go.jp

Abstract

Attempts are being made to visualize the learning process by attaching microphones to students participating in group works conducted in classrooms, and subsequently, their speech using an automatic speech recognition (ASR) system. However, the voices of nearby students frequently become mixed with the output speech data, even when using close-talk microphones with noise robustness. To resolve this challenge, in this paper, we propose using multi-channel voice activity detection (VAD) to determine the speech segments of a target speaker while also referencing the output speech from the microphones attached to the other speakers in the group. The conducted evaluation experiments using the actual speech of middle school students during group work lessons showed that our proposed method significantly improves the frame error rate (38.7%) compared to that of the conventional technology, single-channel VAD (49.5%). In our view, conventional approaches, such as distributed microphone arrays and deep learning, are somewhat dependent on the temporal stationarity of the speakers' positions. However, the proposed method is essentially a VAD process and thus works robustly. It is the practical and proven solution in a real classroom environment.

Index Terms: Voice activity detection, Multiple microphones, Automatic speech recognition, Noise robustness

1. Introduction

Educators are seeking new teaching methodologies to implement active learning, in which students play an active role and deepen their learning by engaging in discussions. The "knowledge constructive jigsaw methodology," promoted by the University of Tokyo's Consortium for Renovating Education of the Future (CoREF) promotes heuristic learning by dividing the students in a classroom into small groups of three to four for mutual discussions, both inside and outside their groups. 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]. A detailed analysis of this text is conducted to allow educators to visualize aspects of the learning process, such as how students gain understanding, discover further questions, and deepen their understanding.

Far-field microphones located on desks show poor performance in current ASR systems. The use of microphone arrays placed in the center of the desks [2] was also considered but was not as effective in a situation where there were multiple discussion groups in one classroom. Therefore, the use of close-talk microphones has become the de facto standard in this field. However, the voices of nearby students frequently become mixed with the output speech data, even when using close-talk microphones with noise robustness [3]. This is problematic because if speeches become mixed at a high level of volume, the speech of the non-targeted speakers may be transcribed.

The need for accurate transcription of group discussions is not limited to schools. Even in a business meeting, it is possible to imagine a scenario in which each participant brings his or her own smartphone and uses a Bluetooth earpiece microphone to transcribe their discussion using ASR. We have not yet evaluated the use of earpiece microphones, but in principle they are expected to work in the same way as a close-talk microphone.

2. Conventional Technologies

The simplest method for canceling the speech from adjacent speakers is applying single-channel VAD [4][5] to the speech of the target speaker. This involves dividing the speech data into frames along a timeline, calculating the speech power, and allowing speech only when its power is higher than a predefined threshold. However, this method does not yield sufficiently accurate results, and it can be impossible to differentiate between speakers when a person is speaking softly or when the voices in the adjacent speakers are loud.

One alternative method is to distribute several far-field microphones on a desk to configure a distributed microphone array to extract the speech of the target speaker [6][7][8][9]. In this system, an NMF, a BSS, or an MVDR is used after synchronizing the audio tracks, and it should be trained on the actual observation to extract only the target speech. If the distance between the speaker and the microphone changes, retraining is required. Therefore, it is unsuitable for use in the scenario considered in this study, in which speakers frequently change their direction to face adjacent speakers with attached close-talk microphones.

Another relevant method involves the combined use of special microphones attached to the throats of the speakers [10]. These microphones can detect vibrations in the throat, and consequently, determine whether the target speaker or an adjacent speaker is speaking. However, the use of distinct microphones presents a barrier for practical use because of the additional cost of purchasing these microphones, and the psychological and physical burden on the speakers to whose throats these microphones are attached.

If it is permissible to wear two or more microphones in each ear of a student, then multi-channel speech separation technology including ICA and deep learning [11][12][13] can be applied. However, as in the case of throat microphones,

http://dx.doi.org/10.21437/Interspeech.2021-200

there are problems of the burden of wearing and the cost of the equipment.

Other conventional methods include using deep learning to differentiate between speakers speaking into a single microphone [14][15] and using acoustic modeling and Bayesian information criterion (BIC) to distinguish between speakers [16]. However, these methods require a separate process to identify the target speaker after differentiating between the speakers. Furthermore, they do not exclude the possibility that the speakers may be partially interchangeable.

The above methods are front-end approaches, but there is also an approach to adjust the word insertion penalty in the decoder of the ASR [17]. However, this depends largely on the acoustic environment and is difficult to determine in advance. Also, most commercial speech recognition systems do not allow the user to make this adjustment.

In this paper, to resolve the abovementioned challenge, we propose robust VAD approach. This is a practical and workable solution, although quite simple compared to the above techniques. It combines the output of multiple close-talk microphones in a group to perform multi-channel voice activity detection (MVAD) to identify the speech segments of a target speaker.

3. Multi-Channel VAD

The MVAD proposed in this paper determines the speech segment of the target speaker while also referring to the speech measured from the close-talk microphones attached to other speakers in the group. The main points of this system are summarized below.

- It conducts VAD. Unlike the microphone array method, it is stable regardless of the direction or position of the speaker.
- The system judges by the all-win condition. Taking a group of four as an example, the system compares the power per frame for each of the three pairs formed by the target speaker and the other three group members. Furthermore, the system verifies that a speech segment belongs to the target speaker only if the target speaker is dominant in all pairs.
- The system uses ambient noise as reference. The power of ambient noise is used as a reference value to normalize the gain for each recording track, in scenarios where the amplifier setting and the distance between the microphone and mouth vary from speaker to speaker. Since ambient noise should be common to all speakers, this reference is justified.

3.1. Normalization of speech power using ambient noise

Quasi-stationary ambient noise segments are those in which the target speaker and nearby speakers are not speaking; these exclude the non-stationary noise segments. A single-channel VAD system based on a Gaussian mixture model (GMM) trained on human speech, or a system that identifies frames with minimum speech power within a moving time window, could be used here. The latter was used for the experiments described in this paper.

Then, for the time series of speech power measured from all microphones attached to the speakers in the group, the average log power of the ambient noise was subtracted in the log domain (in dB). In this paper, we call this the local signal-to-noise ratio (SNR). The difference from the standard SNR is that it includes the index of the frame number. Therefore, it can be used instead of the naive speech power to compare the loudness of speech regardless of the gain setting of the recording device.

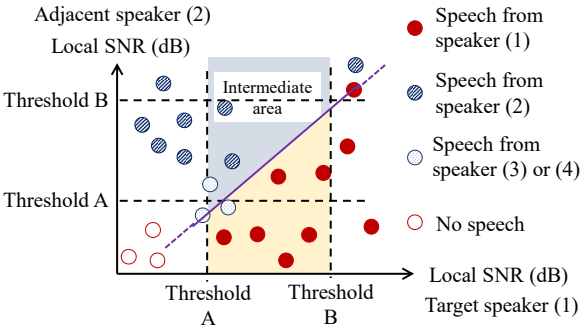

Figure 1: *Local SNR distribution of paired channels.*

3.2. Details of the all-win condition and two-stage threshold

In the experiments conducted in this study, a single microphone is associated with each individual speaker. Thus, when a particular speaker is indicated, it is the same as indicating the associated microphone.

Pairs are formed between the target speaker and each of the other speakers in the group, in an order. For example, if there are four speakers in a group, the target speaker has three pairs to consider. With the pairs, the all-win condition is tested in order as follows.

Figure 1 shows the distribution of the speech power for each pair schematically. It is plotted by frame, and the frame rate was 100 frames per second in the experiments conducted by us. The speech power converted into the local SNRs from the microphones of the target speaker and adjacent speakers are represented on the horizontal and vertical axes, respectively. A diagonal line is drawn in the intermediate area between two thresholds in the figure. When the speech is from the target speaker, the speech power in that frame should be below this diagonal line. If the speech power is above this diagonal line, it suggests a high probability that an adjacent speaker is speaking.

This concept is employed for each pair of microphones. Note that even if a single pair claims that the speech is from an adjacent speaker, it is automatically deduced that the speech is not from the target speaker. Thus, a speech segment will not be identified as belonging to the target speaker ("all-win condition").

Here, the problem is finding a method to draw the abovementioned diagonal line, which serves as the decision boundary. The simplest method is to draw a 45° diagonal line; this is referred to as the "diagonal approximation" in this paper. If the instructions to wear close-talk microphones are followed thoroughly and the students speak clearly and with sufficient volume, the local SNR of each speaker will be distributed equally, and the system should operate effectively even with a fixed decision boundary.

However, this may not be assumed in actual use, and instead, it is preferable to set the decision boundary based on online learning during use. This is achieved as follows. First,

the conventional technology of single-channel VAD is used to identify the speech segments of the target speaker and adjacent speakers. This is not accurate as many frames will be classified as speech segments of both speakers. However, it is used here only as a first approximation. The results of the speaker identification during this first approximation are subsequently used to learn the decision boundary and finally to obtain improved speech segments. These results are used to perform iterative processing, by which the decision boundary is updated, and subsequently, the improved speech segments are determined.

The decision boundary is obtained iteratively via machine learning on a model using speech segments obtained by the previous model. A conventional classification approach, such as a Gaussian model or a support vector machine (SVM), should be applied here. Figure 2 presents a simple implementation. The centroids of the elements in each group are determined, and a dotted line is drawn connecting the two centroids. Then, the perpendicular bisector of this line is the decision boundary. This method was used in the experiments described in this paper. These experiments were conducted both with and without iterations, and their performance was subsequently compared.

The process discussed thus far can be used to determine the dominance of the speech power of the two speakers in a single pair for each frame. To conclude that the speech segment belongs to the target speaker, the dominance must be proved in all the pairs.

The conditions for the two thresholds (A and B) shown in Figure 1 are also applied. As in single-channel VAD, the speech power must be higher than threshold A to be considered as the speech segment of the target speaker. When the speech power becomes higher than threshold B, the frame can be determined as a speech segment regardless of the all-win condition previously discussed in context of the decision boundary. Although students do not speak simultaneously very frequently during group work, when this occurs, the second threshold attempts to retrieve a speech segment as the target speaker, when the voice is sufficiently loud, even though the all-win condition is not met. Conversely, back-channel feedback like “ya” is a small voice. It will not be retrieved when its power is not higher than threshold B.

3.3. Post-processing for speech segment determination

The process discussed thus far is used to identify speech segments at an individual speech frame level. Because decisions are made simply at the frame level, decisions are either often flipped in short intervals or do not form continuous segments with realistic length. Therefore, as in single-channel VAD, speech segments are adjusted by post-processing, such as removing independent short speech segments and filling short gaps in between continuous speech segments. The head and tail of the speech segments are extended. In the conducted experiments using ASR, the amount of extension was set to 20 frames (0.2 s).

4. Evaluation Experiment

4.1. Evaluation data

We conducted two group work classes in a junior high school and recorded the audio data, and subsequently, the time information for the speech segments was manually added to

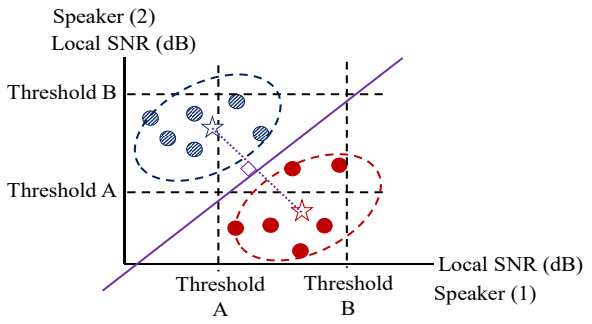

Figure 2: *Decision boundary estimation using centroids.*

the text transcribed from the speech of each student. Each class has two sessions. The first half is called "expert" session and the second half is called "jigsaw" session. Therefore, we have four sessions for evaluation: three sessions have six to seven groups with three to four members per group. one session has three groups with six to seven members per group. Note that these groups share a regular-sized classroom for discussion. The "expert" session lasted about 6 minutes and the "jigsaw" session about 15 minutes.

4.2. VAD

The following four types of VAD were applied to the output obtained from the microphones attached to each individual student to determine the speech segments of each speaker.

1. Single-channel VAD (SVAD)
2. MVAD diagonal approximation (MVAD Diag)
3. MVAD learning without iteration (MVAD No-iter)
4. MVAD learning with iteration (MVAD Iter)

Several VAD threshold candidates were provided for each method. These were selected so to find a balance point of the insertion and the deletion errors. They were converted into thresholds A and B as follows:

Threshold A = Threshold parameter / 2, and
Threshold B = Threshold parameter + 10.

SVAD has threshold A only. The number of iterations for “MVAD Iter” was set as 2.

4.3. Evaluation of segmentation performance

The speech segment information obtained from the four types of VAD are compared to the correct speech segment information (VAD Oracle) obtained manually, following which the performance is evaluated for only VAD. The error rate is defined as (1).

$$Error\ rate = \frac{No.of\ inserted\ frames\ +\ No.of\ deleted\ frames}{Total\ no.of\ speech\ frames\ in\ oracle\ segment} \quad (1)$$

Table 1 lists the results of the evaluation. In all three cases of MVAD, the performance is superior to that of SVAD. This is largely due to the reduction in the insertion errors, indicating that the MVAD is working as expected.

Comparing the three cases of MVAD, the method of determining the decision boundary without iteration showed the highest performance. It is reasonable that it showed better performance than the diagonal approximation without learning. Contrary to expectations, the method learning with iterations

did not show the highest performance. However, the average of the performances measured at various thresholds turned out to be the best.

Note that this evaluation was conducted without extending the head and tail of the speech segments. This was done so that the beginnings and endings of the speech segments given by VAD Oracle seemed trimmed closely.

4.4. Evaluation of transcription performance

Each speech was segmented based on the speech segment information obtained from the four types of VAD and subsequently processed by ASR. IBM Watson STT [18] was used as the speech recognition system for the evaluation. The evaluation was done in July 2020, while the speech recognition performance of the cloud system may be updated. For comparison, two additional cases were performed: the "None" case, in which the speech was input as is without segmentation into the ASR system, and the "Oracle" case, in which the segmentation was performed using correct speech segments. Following this, the character error rate (CER) for ASR was measured by comparing the recognized text against the correct text. The method with a lower CER was regarded as the better method.

Table 2 summarizes the results of the evaluation. It was observed that the CER for SVAD was lower than that of the "None" case, showing that even SVAD is effective to some degree. However, further improvement is desirable because of the wide gap between this and the Oracle case, which is considered to yield ideal results. Performance of each of the three MVAD cases exceeded that of SVAD, and even came close to that of the Oracle case.

Comparing the three cases of MVAD, the benefit of iterative processing was small when evaluating the accuracy of ASR, as it was when evaluating the performance of VAD. Therefore, in actual use, MVAD without iterative processing must be sufficient.

Note that the optimal VAD threshold is smaller for the transcription purpose than for the segmentation purpose. The evaluation of SVAD using ASR should have been performed also with a threshold higher than 40. However, the threshold of 40 is still meaningful because it is the balance point between insertion errors and deletion errors. (As an excuse, the cloud system had been updated and we lost the opportunity to do the additional testing.)

The reason for the Oracle case having an even higher deletion error rate than the "None" case can be interpreted as follows. As shown by the high number of insertion errors, the "None" case generates a large amount of useless text. Therefore, the low rating of the deletion errors is a result of this text occasionally matching with the parts that will become deletion errors.

Table 1: *Frame-based error rate of VAD (%).*

	Threshold	Ins Error	Del Error	Total Error
SVAD	35	33.15	23.88	57.03
	40	19.14	30.86	49.99
	45	10.10	39.39	**49.49**
	50	4.89	49.10	53.99
MVAD Diag	25	24.27	21.01	45.28
	30	14.80	25.23	40.03
	35	8.87	29.98	**38.85**
	40	5.27	35.87	41.14
MVAD No-iter	25	23.75	20.87	44.62
	30	14.38	25.25	39.62
	35	8.64	30.08	**38.72**
	40	5.12	35.99	41.11
MVAD Iter	25	21.27	21.70	42.98
	30	13.33	25.85	39.18
	35	8.23	30.62	**38.85**
	40	4.97	36.44	41.41

Table 2: *Character error rate of ASR using VAD (%).*

	Threshold	Sub Error	Del Error	Ins Error	CER
None		30.35	14.92	21.56	66.82
Oracle		25.55	22.91	2.66	51.12
SVAD	20	27.60	18.91	12.99	59.51
	30	27.34	20.94	10.94	59.21
	40	24.05	24.86	9.49	**58.39**
MVAD Diag	15	26.70	21.35	4.48	52.53
	20	25.74	22.87	3.21	**51.82**
	25	25.28	24.05	2.73	52.07
MVAD No-iter	15	26.77	21.20	4.51	52.48
	20	25.80	22.65	3.25	**51.71**
	25	25.32	23.84	2.82	51.98
MVAD Iter	15	26.70	21.35	4.48	52.53
	20	25.74	22.87	3.21	**51.82**
	25	25.28	24.05	2.73	52.07

5. Conclusion

In this study, we evaluated MVAD by which the speech segments of a target speaker can be accurately estimated by referencing the recording tracks of all the members in a group.

Since this method is essentially a VAD technique, it is robust to temporal non-stationarity in the speakers' positions. Also, while beamforming can only attenuate, this method can completely mute when the target speaker is not speaking, which is an advantage since it has been found that ASR systems tend to output misrecognized words even for attenuated speech. Therefore, we believe that this is a practical and optimal solution for the purpose of this project, which is to transcribe group discussions in the classroom.

The evaluation experiment was conducted using recordings of group discussions held at an actual junior high school. Looking at the details of the results, we can see that even the non-iterative method gives good results in the ASR application. This implies that it can be used as part of a real-time system.

6. Acknowledgements

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

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] S. Araki, M. Okada, T. Higuchi, A. Ogawa and T. Nakatani, "Spatial correlation model based observation vector clustering and MVDR beamforming for meeting recognition," *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2016, pp. 385-389, 2016.

[3] O. Ichikawa, T. Fukuda and R. Tachibana, "Effective speech suppression using a two-channel microphone array for privacy protection in face-to-face sales monitoring," *Acoustical Science and Technology*, vol. 36, no. 6, pp. 507-515, 2015.

[4] 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.

[5] Jongseo Sohn, Nam Soo Kim and Wonyong Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1-3, Jan. 1999.

[6] 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.

[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] 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.

[9] T. Yoshioka, Z. Chen, D. Dimitriadis, W. Hinthorn, X. Huang, A. Stolcke and M. Zeng, "Meeting Transcription Using Virtual Microphone Arrays", *Microsoft Technical Report*, MSR-TR-2019-11, 2019.

[10] Y. Otaka, T. Tsunakawa, M. Nishida, and M. Nishimura, "Voice Activity Detection Using Throat and Lavalier Microphones for Multi-Party Conversations," *Proceedings of International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP) 2017*, 2AM2-3-2, pp. 369-372, 2017.

[11] P. Comon, "Independent component analysis, a new concept?," *Signal Processing, Elsevier*, vol. 36, pp. 287–314, 1994.

[12] R. Gu, J. Wu, SX. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou and D. Yu, "End-to-End Multi-Channel Speech Separation," *arXiv*:1905.06286, 2019.

[13] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari and N. Ono, "Independent Deeply Learned Matrix Analysis for Determined Audio Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1601-1615, 2019.

[14] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, J.R. Hershey, "Single-Channel Multi-Speaker Separation using Deep Clustering," *arXiv:*1607.02173, 2016.

[15] Y. Luo and N. Mesgarani, "TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018*, pp. 696-700, 2018.

[16] M. Nishida and T. Kawahara, "Speaker Model Selection Based on the Bayesian Information Criterion Applied to Unsupervised Speaker Indexing," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 583-592, 2005.

[17] C. Leggetter, V. Valtchev, S. Young, P. Woodland and J. Odell, "The 1994 HTK large vocabulary speech recognition system," *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1995*, pp. 73-76, 1995.

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