IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 January 2023, accepted 6 March 2023, date of publication 9 March 2023, date of current version 22 March 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3254529*

RESEARCH ARTICLE

# Audio-Visual Overlapped Speech Detection for Spontaneous Distant Speech

**MINYOUNG KYOUNG, HYUNGBAE JEON, AND KIYOUNG PARK**
Electronics and Telecommunications Research Institute (ETRI), Daejeon 34129, Republic of Korea

Corresponding author: Minyoung Kyoung (mykyoung@etri.re.kr)

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korean Government (MSIT; 2019-0-00004, "Development of Semi-Supervised Learning Language Intelligence Technology and Korean Tutoring Service for Foreigners," and 2022-0-00989, "Development of Artificial Intelligence Technology for Multi-speaker Dialog Modeling").

**ABSTRACT** Although advances in deep learning have brought remarkable improvements to Overlapped Speech Detection (OSD), the performance in far-field environments is still limited owing to the lack of real-world overlapped speech and a low signal-to-noise ratio. In this paper, we present an end-to-end audiovisual OSD system based on decision fusion between audio and video modalities. Firstly, we propose a simple yet powerful audio data augmentation method for spontaneous distant speech data. Secondly, to maximize the effectiveness of the video modality, we design a video OSD system based on a cross-speaker attention module that explores the visual correlation between multiple speakers. Lastly, we present cross-modality attention module to make the final decision more accurate. Our experimental results demonstrate that our approach outperforms current state-of-the-art methods on a real-world distant speech dataset. Moreover, our approach can robustly detect overlapped speech when compared with its counterpart, which uses audio modality alone.

**INDEX TERMS** Overlapped speech detection, far-field audio data augmentation, audiovisual speech processing, multimodal deep learning.

## I. INTRODUCTION

Overlapped speech detection (OSD) is an essential component of speech-based systems, particularly in spontaneous multiparty conversations. Advances in deep learning have achieved remarkable success, and a variety of neural network-based audio OSD systems [1], [2], [3], [4] have been proposed to improve the performance of close-talk speech data. However, the performance in far-field conditions is still limited owing to the degraded signal quality and the low signal-to-noise ratio (SNR).

Recently, there has been increasing interest in incorporating visual information into speech-based systems such as automatic speech recognition (ASR) [5], [6], [7], [8], [9] and voice activity detection (VAD) [10], [11], [12], [13], [14], [15]. Note that visual information contains additional cues, such as lip movements and number of speakers. In addition, video modality is entirely invariant to acoustic signal corruption, such as high levels of acoustic noise or transient interferences. For these reasons, visual information enhances the performance of speech processing compared with its counterpart, which uses audio modality alone.

Our goal is to develop an end-to-end audiovisual OSD system that improves the performance, especially in multi-speaker distant-talking speech such as meeting scenarios, by incorporating audio and visual information. In recent works, S. Cornell et al. [16], [17] showed that the Temporal Convolutional Network (TCN) and Transformer-based approaches for joint VAD+OSD systems can outperform recent state-of-the-art methods on spontaneous distant speech data, such as the AMI meeting corpus [18]. Although these studies have brought remarkable improvements to VAD systems, there are still challenges in OSD systems, such as class imbalance. This imbalance, which arises because speech overlap is very rare in real-world human conversations, results in relatively low performance. Building on these

The associate editor coordinating the review of this manuscript and approving it for publication was Easter Selvan Suviseshamuthu.

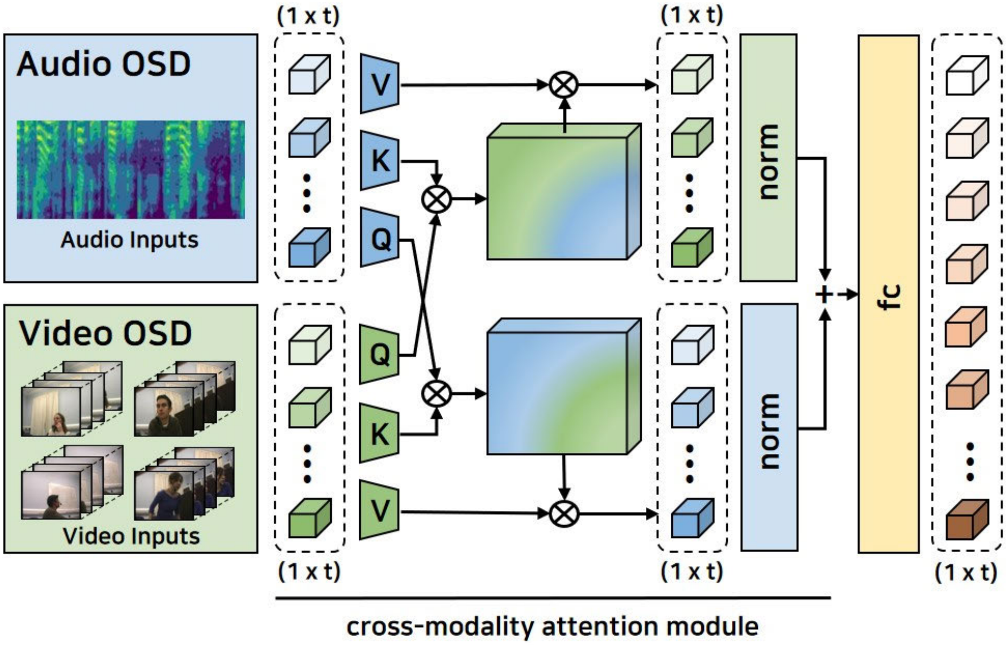

**FIGURE 1.** Proposed end-to-end audiovisual OSD system architecture.

previous studies, we study how to incorporate visual information into the audio modality to improve the performance of the OSD system.

To successfully develop our approach, we emphasize the following three points: 1) Most existing audio OSD systems have relied on close-talk or synthetic mixed speech for training. However, there are differences between close-talk and far-field speech in terms of acoustic characteristics, such as SNR and distortion from reverberation. To alleviate this problem, we propose an audio data augmentation algorithm for far-field OSD. 2) To maximize the effectiveness of video modality, we designed a video OSD system based on a cross-speaker attention module that explores the visual correlation between multiple speakers. This approach enables the detection of overlapping speech without audio information. 3) To better leverage both audio and video modalities for OSD systems, we present an end-to-end audiovisual OSD system that combines the decision vectors of audio and video OSD systems based on the cross-modality attention module to encourage features across modalities compared to a simple concatenation of the decision vectors. Considering these points, audio OSD, video OSD, and audiovisual OSD systems were implemented and evaluated. A more detailed discussion of our approach is provided in Section III.

Our primary contributions of this paper are summarized as follows:

- We developed an end-to-end audiovisual OSD system that works robustly for multi-speaker distant-talking speech when compared with stand-alone audio modality. To the best of our knowledge, this is the first study to use a multimodal approach for far-field OSD.
- The proposed audio data augmentation algorithm makes it possible to improve the performance of current state-of-the-art audio OSD systems in a real-world meeting dataset.
- The proposed multimodal approach effectively captures the correlation between audio and video modalities, even in the presence of background acoustic noise.

Through the experiments described in Section IV, we demonstrate that these three advantages help improve the far-field OSD performance.

## II. DATASET

In this work, we used the AMI meeting corpus [18], which consists of over 100 hours of meeting recordings. The corpus was recorded using a wide range of devices, including microphone arrays, per-speaker headsets, lapel microphones, and individual cameras, and had almost 4 participants. And it is manually annotated for many different phenomena including speaker activity and orthographic transcription.

In [16] and [17], which is our baseline system, the author claims that using training targets obtained via forced-alignment (FA) brings substantial improvement, even when official manual annotation is used as the ground truth in the evaluation stage. For this reason, we used the same FA-based labels for training, and the test set was evaluated using official annotation.

## III. PROPOSED OVERLAPPED SPEECH DETECTION

In this study, we aim to develop an end-to-end audiovisual OSD system for multi-speaker distant-talking speech. The complete system is illustrated in Fig. 1. First, we present audio data augmentation algorithm to improve the performance of existing audio OSD systems in Section III-A. Next, we propose a video OSD system to detect overlapped speech using only visual information by analyzing the visual correlation between speakers, as described in Section III-B. Finally, in Section III-C, we explain an end-to-end audiovisual OSD system based on a cross-modality attention module for combining the decision vectors of audio and video OSD.

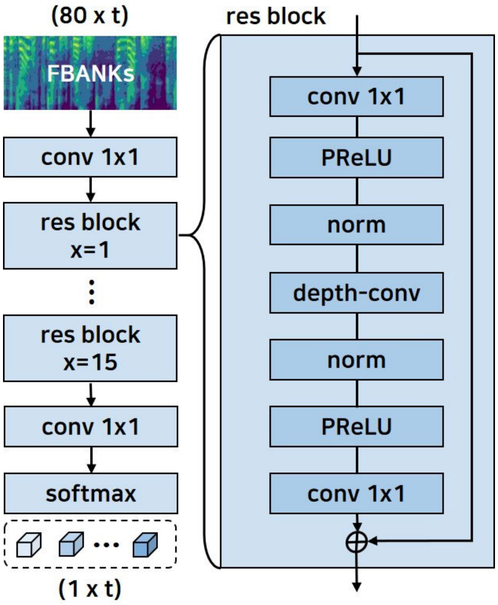

**FIGURE 2.** Proposed audio OSD system architecture.

**TABLE 1.** Frame-level speaker counting frequency (%) for the AMI development and test sets.

| AMI | | No Overlap | | Overlap | | |
|---|---|---|---|---|---|---|
| | | 0 spk | 1 spk | 2 spk | 3 spk | 4 spk |
| Official Annotation | dev | 17.90 | 65.55 | 13.61 | 2.53 | 0.41 |
| | test | 16.65 | 67.16 | 12.40 | 3.04 | 0.75 |
| FA-based Labels | dev | 37.22 | 57.08 | 5.36 | 0.32 | 0.01 |
| | test | 37.17 | 57.86 | 4.65 | 0.31 | 0.01 |

### A. AUDIO OSD

For audio OSD system, we use the TCN-based model as shown in Fig. 2. As a sequence labeling model, we input $x = [x_1, x_2, \ldots, x_t]$, where t is the length of the input sequence, and the corresponding class label $y = [y_1, y_2, \ldots, y_t]$ is used to guide the training, where 0 refers to no overlap and 1 refers to overlap. In this paper, we use 80 log-mel filterbank features computed using a Hann window length of 400 with a 10 ms shift as the input sequence.

As mentioned earlier, the OSD task on real-world data is affected by class imbalance. Table 1 shows the speaker-counting statistics for the AMI. The number of 2-speakers, 3-speakers, and 4-speakers frames corresponding to overlapped speech is a very small fraction of the total number of frames, especially in the case of FA-based labels. To avoid suffering from this imbalance, most of the existing audio OSD systems use artificially mixed audio chunks randomly selected from close-talk speech data (headset and lapel mixes) and far-field speech data (microphone arrays). However, these synthetic data may degrade the system performance, partly because of the mismatch between the training data (close-talk mixed speech) and testing data (far-field speech).

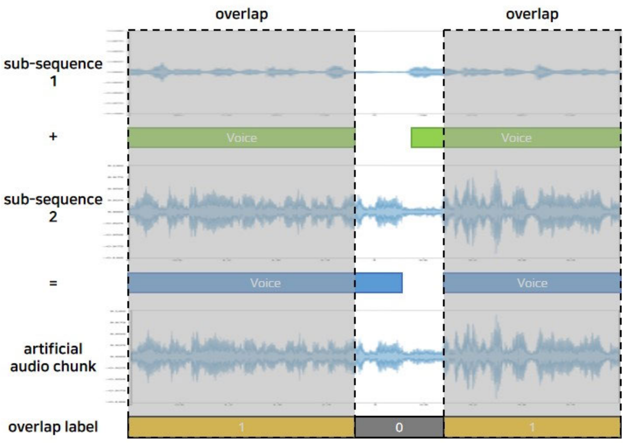

**FIGURE 3.** Example of proposed audio data augmentation.

**TABLE 2.** Frame-level class frequency (%) for the overlapped speech detection task on the AMI training set and after augmentation method.

| AMI | no overlap | overlap |
|---|---|---|
| before augmentation | 94.73 | 5.27 |
| after augmentation | 56.60 | 43.40 |

To overcome this problem, we propose a simple yet effective audio data augmentation algorithm for real-world far-field OSD. First, we segment all sequences of the 1 channel audio data of the first microphone array into a fixed-length sub-sequence with a 1 s stride, where each sub-sequence corresponds to 6 s audio chunks. Next, to increase the number of positive training samples, the artificial audio chunks are created by summing up two randomly chosen sub-sequences only if the percentage of utterance frames of each sub-sequence is more than 80%. For class balance, almost half of the training dataset consists of artificially mixed audio chunks as shown in Table 2. Note that the maximum possible number of overlapping speakers in an artificial audio chunk is 8, because each sub-sequence can consist of up to 4 speakers. The example of proposed audio data augmentation method is illustrated in Fig. 3.

### B. VIDEO OSD

The video OSD system aims to detect overlapped speech using only visual information. Through this approach, we expect to enhance the robustness against high levels of acoustic noise or degraded signal quality. Existing works in this area, such as lip-reading [19], [20], [21], have focused on recognizing lip movements for a single speaker. However, these approaches are difficult to use in real-world multi-speaker scenarios because the speaker's mouth may sometimes be undetectable (e.g., the speaker may turn his/her head to talk to others).

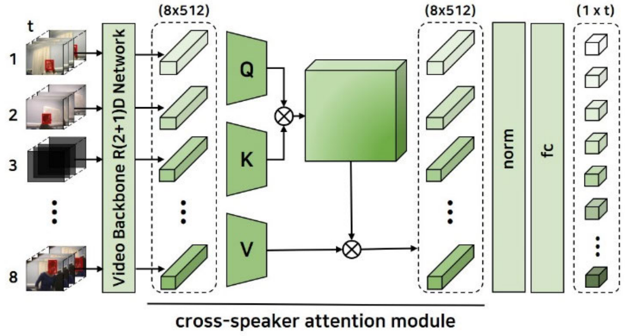
t
1
2
3
8
Video Backbone R(2+1)D Network
(8x512)
Q
K
V
(8x512)
norm
fc
(1 x t)
cross-speaker attention module

**FIGURE 4.** **Proposed video OSD system architecture.**

In this work, we focus on analyzing the interaction between multiple speakers to avoid the partial occlusion of individual speakers. The architecture of the proposed video OSD system, which adopt the sequence labeling model like the audio OSD system, is illustrated in Fig. 4. Note that to pair with artificial audio chunks of audio OSD system for multimodal training, the duration and number of speaker video sequences should be set to 6 s and 8, respectively. First, to extract speaker feature vector from the each speaker video sequence, we use video backbone network inspired from ResNets with (2+1)d convolutions [22], as shown in Fig. 5. If the number of speaker video sequences is fewer than 8, or the input data are absent for some reason such as some meetings have fewer than 4 participants, the input image is replaced with a black image, as shown in Fig. 4. Therefore, we obtained an 8 × 512 dimensional feature vector (8 in the speaker domain, and 512 in the context domain) for the multi-speaker video sequences. Here, the parameters of video backbone network are shared across all speakers. Next, to capture the visual relationship between multiple speakers for OSD task, we adopted the cross-speaker attention module consisting of a scaled dot-product multi-head self-attention layer [23] with four attention heads, followed by a position-wise feed-forward layer. The module aims to find interactive information between the feature vectors from multiple speakers. The output of the fully-connected layer as decision vectors is constrained to the range of 0–1 and can be considered as the probability of the absence or presence of overlapped speech.

### C. AUDIOVISUAL OSD

Through the above tasks for audio and video OSD, we obtained decision vectors for both audio and visual modalities. To make the final decision more accurate, we proposed a cross-modality attention module, which contains two transformer layers with four attention heads, as shown in Fig. 1. The module aims to explicitly model the correlation between the decision vectors in different modalities. The two inputs of the module are denoted as the audio-modality decision vector $\mathrm{f}^{\mathrm{a}}$ and video-modality decision vector $\mathrm{f}^{\mathrm{v}}$. Then, we used the module where the decision vectors from the audio OSD system attend to the decision vectors from the video OSD system and vice versa:

$$\phi_{cross1}\left(f^{a}, f^{v}, f^{v}\right) = \mathrm{Softmax}\left(\frac{f^{a} f^{vT}}{\sqrt{d}}\right) f^{v} \tag{1}$$

$$\phi_{cross2}\left(f^{v}, f^{a}, f^{a}\right) = \mathrm{Softmax}\left(\frac{f^{v} f^{aT}}{\sqrt{d}}\right) f^{a} \tag{2}$$

where $\phi_{cross1}(\cdot)$ is responsible for measuring the correlation from the audio to video, $\phi_{cross2}(\cdot)$ is responsible for finding multi-speaker interactive information from video to audio, and d is the decision vector dimension. These are fed to normalization layers followed by fully-connected layer performing the final decision (the output dimension of the fc layer matches the number of audio frames). This design enables the module to properly fuse the two decision vectors made from both modalities so that the complete system can robustly detect overlapped speech when compared to a simple approach, such as the concatenation of decision vectors. In the following experiments, we demonstrate that our approach improves OSD performance in multi-speaker distant-talking speech.

## IV. EXPERIMENTAL RESULTS

In this section, we compare our approach with a recent state-of-the-art system proposed for OSD on AMI far-field data. Our approach consists of audio, video, and audiovisual OSD. To evaluate our proposed audio data augmentation for audio OSD, we used Long-Short Term Memory (LSTM) [24], Convolutional-Recurrent Neural Network (CRNN) [25], and TCN implemented in [17] for fair comparison. As audio input features, we used 80 log-mel filterbank features extracted for each 6 s audio chunks.

For video OSD, it is necessary to reduce the size of the input image to achieve computational efficiency. In this study, we extracted 10 consecutive facial images from each speaker video sequence at 0.6 s intervals and resized to a fixed resolution of 112 × 112. To achieve this, we adopted MediaPipe [26] to estimate the locations of the face. Note that if there is no input data available, such as when a face cannot be detected, the input image is replaced with a black image, as mentioned earlier. To extract speaker feature vector from each speaker

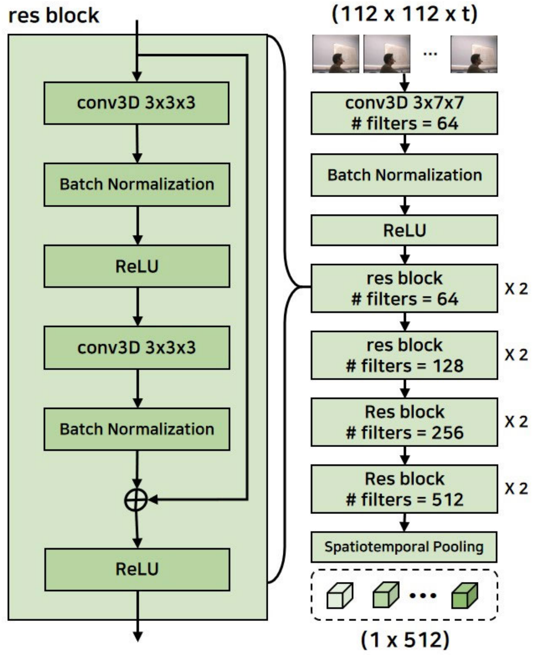

**FIGURE 5.** Proposed video backbone network architecture.

**TABLE 3.** Performance (AP, %) comparison between our proposed data augmentation algorithm and baseline approach [16] for existing audio OSD systems.

| Audio OSD | Baseline [16] | Ours |
|---|---|---|
| LSTM [24] | **34.3** | 32.2 |
| CRNN [25] | 38.9 | **53.7** |
| TCN [16] | 54.2 | **62.7** |
| Transformer [17] | 57.8 | - |

**TABLE 4.** Performance (AP, %) comparison by modality. Audiovisual OSD system uses TCN as audio OSD.

| Modality | Method | Performance |
|---|---|---|
| Audio | TCN | 62.7 |
| Video | Cross-speaker attention module | 20.0 |
| Audiovisual | Concatenation | 65.1 |
| | Cross-modality attention module | **67.2** |

**TABLE 5.** Performance (AP, %) comparison for different SNR levels.

| SNR (dB) | Audio OSD | | | Audiovisual OSD |
|---|---|---|---|---|
| | LSTM | CRNN | TCN | TCN + Video OSD |
| -10 | 23.2 | 38.5 | 50.7 | **57.9** |
| -5 | 27.0 | 46.9 | 56.5 | **60.2** |
| 0 | 30.3 | 51.3 | 59.1 | **61.9** |
| 5 | 30.6 | 52.9 | 61.2 | **64.7** |
| 10 | 30.9 | 53.0 | 62.1 | **66.1** |

**TABLE 6.** Performance (Recall) comparison for different number of overlapping speakers. Both methods use TCN as audio OSD.

| Method | Overlap | | |
|---|---|---|---|
| | 2 spk | 3 spk | 4 spk |
| Audio OSD | 0.633 | 0.890 | 0.929 |
| Audiovisual OSD | **0.783** | **0.967** | **0.986** |

video sequence, we utilize the R(2+1)D network having 18 layers. To implement our approach, we used the PyTorch framework. And we used the Rectified Adam (RAdam) optimizer [27] to minimize the cross-entropy loss between the estimated frame-level posterior probabilities and the true class distribution with a learning rate of 0.001.

For testing, we used the 1 channel audio data of the first microphone array of the test set in the AMI corpus. The performance of the OSD is evaluated based on the Average Precision (AP) widely used in tasks that exhibit strong class imbalance, such as object detection. Table 3 shows the performance comparison between our proposed audio data augmentation algorithm and baseline approach under the same conditions (e.g., network architecture and training techniques). The proposed algorithm significantly improves the detection performance without additional training techniques, especially for CRNN (+14.8% AP). From the results, we infer the following three points. 1) When using far-field speech data for testing, the existing data augmentation method of baseline system [16], which uses the training data consisting of close-talk and synthetic mixtures, had a relatively low performance, partly because of the mismatch between training and testing data in acoustic characteristic, such as the level of background noise and reverberation. 2) The baseline approach created the same number of training samples for 2, 3, and 4 concurrent speakers in order to address the class imbalance problem. However, since it is very rare for more than three people to talk at the same time in a real-world meeting scenario dataset (e.g., approximately 3.8% of the test set in the AMI corpus), our approach focuses on generating new simultaneous speaker samples by overlapping two random audio chunks consisting of predominantly single-speaker from the original array data. 3) The data augmentation technique by summing multiple chunks from real dataset is similar to baseline approach, but our approach has significantly higher performance than baseline approach for various audio OSD systems. For these reasons, our approach can be a more suitable and effective solution to the class imbalance problem in the far-field OSD task.

To demonstrate the advantage of the decision fusion of the audio and video OSDs, we evaluated both single-modality (audio and video OSDs) and multi-modality (audiovisual OSD) systems as shown in Table 4. The performance of the video OSD system is not sufficient compared with the performance of the audio OSD system as a result of the relatively rough video sampling. However, it can be seen that our multimodal approach significantly outperforms the stand-alone audio modality system, despite the relatively low performance of the video modality system. The results confirm that using video information helps to enhance the performance of speech processing compared with its counterpart, which uses audio modality alone. It also proves that the proposed cross-modality attention module can make final decisions more accurately than simple concatenation of the decision vectors (+2.1% AP).

To evaluate the acoustic noise robustness of our approach, the test data were generated with five levels of SNR settings by synthesizing the test set in the AMI corpus and the office noise from the DEMAND noise dataset [28]. As shown in Table 5, our multimodal approach improves robustness at lower SNR when compared to the degradation of performance in audio-only OSD systems. The results demonstrate that our approach effectively captures the correlation between audio and video modalities, even in the presence of background noise such as babbling.

As shown in Table 6, we report performance comparison for different number of overlapping speakers. Recall is the ability of the method to find the overlapped speeches, measured as the ratio of correctly detected overlapped speeches to the number of actual overlapped speeches for each speaker counting. From the results, we observed the following two points. 1) Although our audio data augmentation algorithm is designed to focus on scenarios with two concurrent speakers, the audio OSD system detects more accurately as more people speak at the same time. 2) The multi-modality system consistently outperforms the audio single-modality system regardless of the number of overlapping speakers, with the most significant gain being in the sub-set with 2 speakers (+0.15 Recall).

## V. CONCLUSION

In this paper, we presented an end-to-end audiovisual OSD system for multi-speaker distant-talking speech. Through experimental results, we demonstrated that our proposed audio data augmentation algorithm can improve the performance of recent state-of-the-art audio OSD systems in far-field conditions. The proposed cross-speaker attention module-based video OSD system can capture spatiotemporal features that are invariant to acoustic features. Finally, the proposed cross-modality attention module can make more accurate final decisions than that of a simple concatenation of decision vectors. However, owing to the relatively rough video sampling, the performance of the video OSD system is not sufficient compared with the performance of the audio OSD system. One possible solution is to increase the number of consecutive images using denser video sampling. However, this is more costly; therefore, our future work will be to further improve the performance with effective video modality representation and explore the audio-visual relationship and synchronization information. Moreover, we will evaluate our proposed system in the presence of visual artifacts (e.g., partial occlusions) to analyze the effect of the interactive correlation between multiple speakers.

## REFERENCES

[1] J. T. Geiger, F. Eyben, B. Schuller, and G. Rigoll, "Detecting overlapping speech with long short-term memory recurrent neural networks," in *Proc. Interspeech*, Aug. 2013, pp. 1668–1672.

[2] V. Andrei, H. Cucu, and C. Burileanu, "Detecting overlapped speech on short timeframes using deep learning," in *Proc. Interspeech*, Aug. 2017, pp. 1198–1202.

[3] M. Kunesova, M. Hruz, Z. Zajic, and V. Radova, "Detection of overlapping speech for the purposes of speaker diarization," in *Proc. Int. Conf. Speech Comput.*, 2019, pp. 247–257.

[4] L. Bullock, H. Bredin, and L. P. Garcia-Perera, "Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7114–7118.

[5] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8717–8727, Dec. 2018.

[6] T. Makino, H. Liao, Y. Assael, B. Shillingford, B. Garcia, O. Braga, and O. Siohan, "Recurrent neural network transducer for audio-visual speech recognition," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 905–912.

[7] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7613–7617.

[8] D. Serdyuk, O. Braga, and O. Siohan, "Audio-visual speech recognition is worth 32 × 32 × 8 voxels," 2021, *arXiv:2109.09536*.

[9] O. Braga, T. Makino, O. Siohan, and H. Liao, "End-to-end multi-person audio/visual automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6994–6998.

[10] R. Ahmad, S. P. Raza, and H. Malik, "Unsupervised multimodal VAD using sequential hierarchy," in *Proc. IEEE Symp. Comput. Intell. Data Mining (CIDM)*, Apr. 2013, pp. 174–177.

[11] F. Tao and C. Busso, "End-to-end audiovisual speech activity detection with bimodal recurrent neural models," *Speech Commun.*, vol. 113, pp. 25–35, Oct. 2019.

[12] I. Ariav and I. Cohen, "An end-to-end multimodal voice activity detection using WaveNet encoder and residual networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 265–274, May 2019.

[13] O. Kopuklu, M. Taseska, and G. Rigoll, "How to design a three-stage architecture for audio-visual active speaker detection in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1193–1203.

[14] J. Xiong, Y. Zhou, P. Zhang, L. Xie, W. Huang, and Y. Zha, "Look&Listen: Multi-modal correlation learning for active speaker detection and speech enhancement," 2022, *arXiv:2203.02216*.

[15] S. Thermos and G. Potamianos, "Audio-visual speech activity detection in a two-speaker scenario incorporating depth information from a profile or frontal view," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2016, pp. 579–584.

[16] S. Cornell, M. Omologo, S. Squartini, and E. Vincent, "Detecting and counting overlapping speakers in distant speech scenarios," in *Proc. Interspeech*, Oct. 2020, pp. 3107–3111.

[17] S. Cornell, M. Omologo, S. Squartini, and E. Vincent, "Overlapped speech detection and speaker counting using distant microphone arrays," *Comput. Speech Lang.*, vol. 72, Mar. 2022, Art. no. 101306.

[18] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus," in *Proc. 5th Int. Conf. Methods Techn. Behav. Res.*, 2005, pp. 137–140.

[19] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. CVPR*, 2017, pp. 3444–3450.

[20] S. Fenghour, D. Chen, K. Guo, and P. Xiao, "Lip reading sentences using deep learning with only visual cues," *IEEE Access*, vol. 8, pp. 215516–215530, 2020.

[21] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic, "Lip-reading with densely connected temporal convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2857–2866.

[22] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 1–11.

[24] N. Sajjan, S. Ganesh, N. Sharma, S. Ganapathy, and N. Ryant, "Leveraging LSTM models for overlap detection in multi-party meetings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5249–5253.

[25] F.-R. Stoter, S. Chakrabarty, B. Edler, and E. A. P. Habets, "CountNet: Estimating the number of concurrent speakers using supervised learning," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 2, pp. 268–282, Feb. 2019.

[26] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Guang Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for building perception pipelines," 2019, *arXiv:1906.08172*.

[27] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proc. ICLR*, 2020, pp. 1–14.

[28] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multichannel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Meetings Acoust. ICA*, 2013, Art. no. 035081.

**MINYOUNG KYOUNG** received the B.S., M.S., and Ph.D. degrees in computer engineering from Hanbat National University, Daejeon, Republic of Korea, in 2014, 2016, and 2021, respectively. Since 2021, he has been with the Artificial Intelligence Research Laboratory, Electronics and Telecommunications Research Institute (ETRI), Daejeon. His research interests include computer vision, speech recognition, and multi-modal representation.

**HYUNGBAE JEON** received the B.S. degree in electronics engineering from Yonsei University, Seoul, Republic of Korea, in 1999, and the M.S. degree in electrical engineering and the Ph.D. degree in bio and brain engineering from Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea, in 2001 and 2016, respectively. Since 2001, he has been with the Spoken Language Processing Research Section, Electronics and Telecommunications Research Institute (ETRI), where he is currently a Principal Researcher. His main research interests include speech recognition, language modeling, and machine learning.

**KIYOUNG PARK** received the M.S. and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, in 1999 and 2003, respectively. From 2003 to 2005, he was with the Samsung Advanced Institute of Technology (SAIT), Yongin, Republic of Korea, where he contributed to the research and development of human–machine interaction systems. Since 2005, he has been with the Electronics and Telecommunication Research Institute (ETRI), Daejeon, where he is currently a Principal Researcher. His main research interests include speech recognition, signal processing, and machine learning.