Audio-Visual Speech Enhancement With Selective Off-Screen Speech Extraction

Tomoya Yoshinaga[1*], Keitaro Tanaka[1*], Shigeo Morishima[2]
[1] School of Advanced Science and Engineering, Waseda University, Tokyo, Japan
[2] Waseda Research Institute for Science and Engineering, Tokyo, Japan

***Abstract*—This paper describes an audio-visual speech enhancement (AV-SE) method that estimates from noisy input audio a mixture of the speech of the speaker appearing in an input video (on-screen target speech) and of a selected speaker not appearing in the video (off-screen target speech). Although conventional AV-SE methods have suppressed all off-screen sounds, it is necessary to listen to a specific pre-known speaker's speech (e.g., family member's voice and announcements in stations) in future applications of AV-SE (e.g., hearing aids), even when users' sight does not capture the speaker. To overcome this limitation, we extract a visual clue for the on-screen target speech from the input video and a voiceprint clue for the off-screen one from a pre-recorded speech of the speaker. Two clues from different domains are integrated as an audio-visual clue, and the proposed model directly estimates the target mixture. To improve the estimation accuracy, we introduce a temporal attention mechanism for the voiceprint clue and propose a training strategy called the muting strategy. Experimental results show that our method outperforms a baseline method that uses the state-of-the-art AV-SE and speaker extraction methods individually in terms of estimation accuracy and computational efficiency.**

***Index Terms*—Audio-visual speech enhancement, speaker extraction, multimodal, deep learning**

I. Introduction

Audio-visual speech enhancement (AV-SE) aims to extract a target speaker's speech from a noisy input signal (a.k.a. speech enhancement) by using an additional visual clue of the speaker, typically lip movement [1], [2]. As lip movement helps us to track the synchronizing target speech contaminated by non-speech noise or interfering speech, AV-SE works robustly to various kinds of noise. AV-SE has the potential for practical applications, such as hearing aids [2], [3], telecommunication [4], and automatic speech recognition front end [5].

The standard approach of AV-SE is to extract only the speech of the speaker appearing in an input video (on-screen target speech) and suppress the other sounds (off-screen sounds). Based on the fact that humans improve their speech perception by watching a speaker's face [6], a pioneer study on AV-SE [7] applied a statistical speech enhancement model that used lip shape features. Since deep neural networks (DNNs) appeared, DNN-based methods [1], [5] have been the major approach of AV-SE because of DNNs' high capability of extracting speech and fusing multimodal information. Recently, several studies have also tackled practical problem settings such as temporal occlusion of the speaker's mouth [8], [9].

*These two authors contributed equally to this work. This work is supported in part by JSPS KAKENHI Nos. 19H04137, 21H05054, 22J22424, 22KJ2959. We thank Mr. Masaki Kuribayashi for his valuable feedback.

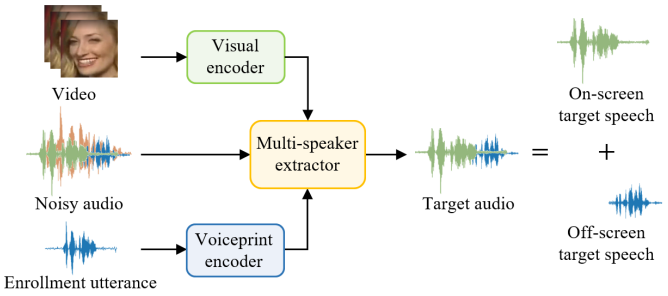


Fig. 1. The proposed method directly extracts the mixture of on-screen and off-screen target speech from noisy audio using the corresponding video and enrollment utterance.

In some practical situations, however, users need to listen to a pre-known speaker's speech in the off-screen sounds as well as the on-screen target speech. For example, when AV-SE is applied to hearing aids that extract an interlocutor's speech, young users should always pay attention to what their parents or teachers say, even when users' sight does not capture the speaker's face. Announcements in stations also should not be suppressed for user safety. These situations call for a method that can simultaneously extract the on-screen target speech and selected off-screen speech (off-screen target speech) from a noisy input signal, where the voice characteristics of the off-screen target speaker are pre-known.

A straightforward approach to this situation is to extract each target speech individually and mix them. The on-screen target speech can be extracted by AV-SE with the help of the corresponding visual clue. The off-screen one can be extracted by speaker extraction [10], [11] with the help of a voiceprint clue from a pre-recorded speech of the speaker (enrollment utterance). However, the mixing operation deteriorates the output signals because of the accumulation of estimation errors, such as artifacts and remaining non-speech noises or interfering speech in each output. This approach is also computationally inefficient because two independent models are required to obtain each output from the same input audio, which would be undesirable in low-resource devices [12].

In this paper, we propose a unified AV-SE model with selective off-screen speech extraction that directly estimates a mixture of the on-screen and off-screen target speech (Fig. 1). At the heart of our method is the multimodal fusion of two clues obtained from different domains. Specifically, we extract the visual clue for the on-screen target speech from the input video and the voiceprint clue for the off-screen target speech

from the enrollment utterance. These clues are integrated as an audio-visual clue, with which a multi-speaker extractor extracts the target mixture from the input audio. End-to-end training enables our model to perform more accurately than the mixing-based approach.

The main contribution of this study is to propose a computationally efficient yet high-performance method for a novel practical problem setting in AV-SE. Our model further improves its performance by attention mechanism and muting strategy. When the off-screen target speech temporally does not exist in the noisy input audio, the model does not need to refer to the voiceprint clue. We thus estimate the voice activity of the off-screen target speech and calculate temporal attention, which controls how much the voiceprint clue contributes to the audio-visual clue. We also utilize a muting strategy, where either on-screen or off-screen target speech in the noisy input audio is muted during training. This encourages the model to strongly bind each clue and the characteristics of the corresponding signal. Experimental results show that our method outperforms a straightforward baseline method (even the combination of the state-of-the-art AV-SE and speaker extraction methods) in terms of the quality of output signals and the number of model parameters.

II. Related work

This section reviews target speaker extraction methods in terms of the modalities of their clues. We also briefly consider denoising methods as alternative approaches to our problem.

A. Target speaker extraction

Target speaker extraction aims to extract the target speech from a noisy input signal using additional information about the speaker. Existing methods can mainly be categorized into two approaches: visual-clue-based and audio-clue-based. The visual-clue-based approach, namely AV-SE, utilizes lip movements [1], [2] or crops of face images [5] synchronized with the target speech. On the other hand, the audio-clue-based approach, namely speaker extraction, utilizes a speaker-dependent voiceprint [10], [11] obtained from an enrollment utterance. Although these approaches have flourished independently, some studies are recently trying to integrate them as an audio-visual-clue-based approach. In this approach, the visual and voiceprint clues can work complementarily, and thus models become robust against temporal occlusion of lip movements [8], [9] or contamination in enrollment utterances [9]. Note that both clues are designed to extract the same on-screen target speech. Our method is also one of the audio-visual-clue-based approaches, but we use the two clues to extract different target speech signals.

B. Denoising

While the proposed method estimates the target mixture by additively extracting two speech signals from the input audio, our goal might be achieved by subtractively suppressing sounds other than the two speech signals contained in the input audio. Such approaches are called selective noise suppression [13] or audio-only speech enhancement [14], [15]. Selective noise suppression removes only unnecessary noises using their enrollment recordings without removing *necessary noises* (e.g., alarms), and audio-only speech enhancement removes all non-speech sounds uniformly. However, these approaches are actually not appropriate for our goal because they require all enrollment recordings of unnecessary noises prepared in advance or fail to remove unnecessary speech. In contrast, our additive approach requires only one enrollment utterance in advance to work in any unknown acoustic conditions.

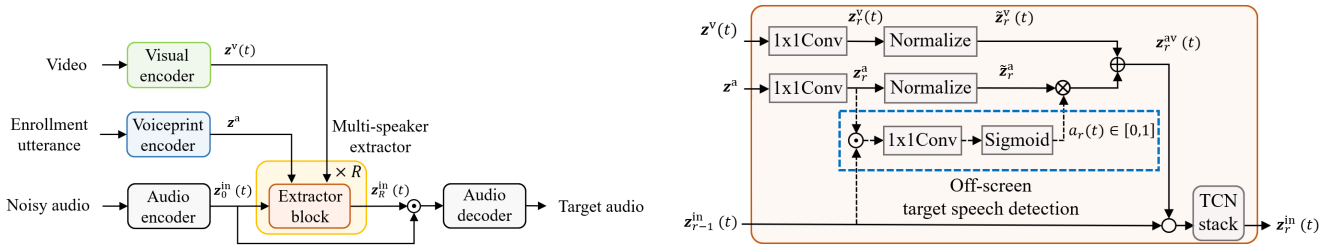
III. Proposed method

This section explains the proposed framework that directly extracts the mixture of the on-screen and off-screen target speech from the noisy input audio. We also present the attention mechanism and muting strategy that are uniquely motivated by the framework to improve the model's performance in multi-target and multi-modal speaker extraction.

A. Framework

Our model is a time-domain encoder-decoder model overall (Fig. 2(a)). It consists of five parts: audio encoder, visual encoder, voiceprint encoder, multi-speaker extractor, and audio decoder. The audio encoder transforms the noisy input audio waveform into initial time-variant latent representations $\mathbf{z}_0^{\mathrm{in}}(t) \in \mathbb{R}^{D^{\mathrm{in}}} (t = 1, ..., T)$, where D^{in} and T are the dimension of each latent representation and the number of time frames, respectively. The visual encoder extracts time-variant embeddings $\mathbf{z}^{\mathrm{v}}(t) \in \mathbb{R}^{D^{\mathrm{v}}}$ from a sequence of cropped lip images in the input video synchronized with the on-screen target speech, where D^{v} is the dimension of each embedding. Note that the last layer of the encoder conducts upsampling to match the temporal resolution of $\mathbf{z}_0^{\mathrm{in}}(t)$. The voiceprint encoder extracts a time-invariant speaker embedding $\mathbf{z}^{\mathrm{a}} \in \mathbb{R}^{D^{\mathrm{a}}}$ from the enrollment utterance to represent the voice characteristics of the off-screen target speaker (e.g., pitch and timbre), where D^{a} is the dimension of the embedding. Given $\mathbf{z}^{\mathrm{v}}(t)$, $\mathbf{z}^{\mathrm{a}}$, and $\mathbf{z}_0^{\mathrm{in}}(t)$, the multi-speaker extractor calculates time-variant masks $\mathbf{z}_R^{\mathrm{in}}(t) \in \mathbb{R}^{D^{\mathrm{in}}}$ which are applied to $\mathbf{z}_0^{\mathrm{in}}(t)$, where R is the number of iterations (explained below). Then, the audio decoder estimates the target audio from the masked initial latent representations of the input mixture.

For the visual and voiceprint encoder, we use the corresponding encoding modules of existing AV-SE and speaker extraction models, respectively. The other three parts are based on a time-domain speech separation model Conv-TasNet [16]. Below, we focus on the multi-speaker extractor, which is the heart of the proposed framework. The multi-speaker extractor consists of R consecutive extractor blocks (Fig. 2(b)) and they form $\mathbf{z}_R^{\mathrm{in}}(t)$ by iteratively processing $\mathbf{z}_0^{\mathrm{in}}(t)$ conditioned by an audio-visual clue instead of a single-modal clue as in AV-SE or speaker extraction. Let r be an iteration index $(r = 1, ..., R)$. $\mathbf{z}^{\mathrm{v}}(t)$ and $\mathbf{z}^{\mathrm{a}}$ are individually processed by pointwise convolution layers and transformed to $\mathbf{z}_r^{\mathrm{v}}(t) \in \mathbb{R}^{D^{\mathrm{av}}}$ and $\mathbf{z}_r^{\mathrm{a}} \in \mathbb{R}^{D^{\mathrm{av}}}$, where D^{av} is the dimension of each embedding. We sum



(a) Overall architecture of our model.

(b) The r-th extractor block.

Fig. 2. Illustration of our model. $\odot$, $\otimes$, $\oplus$, and $\bigcirc$ represent element-wise multiplication, multiplication between a vector and scalar, summation, and concatenation, respectively. (a) Overall architecture based on Conv-TasNet [16]. (b) Details of the r-th extractor block. The dotted arrows are only performed when we use the attention mechanism.

their normalized embeddings, $\tilde{\mathbf{z}}_r^{\mathrm{v}}(t)$ and $\tilde{\mathbf{z}}_r^{\mathrm{a}}$, and obtain an audio-visual embedding $\mathbf{z}_r^{\mathrm{av}}(t) \in \mathbb{R}^{D^{\mathrm{av}}}$. $\mathbf{z}_r^{\mathrm{av}}(t)$ is concatenated with $\mathbf{z}_{r-1}^{\mathrm{in}}(t)$ and processed by temporal convolutional network (TCN) stack [17]. The output of the TCN stack, $\mathbf{z}_r^{\mathrm{in}}(t)$, is used for the next iteration. After R iterations, the initial latent representation $\mathbf{z}_0^{\mathrm{in}}(t)$ is multiplied by the estimated mask $\mathbf{z}_R^{\mathrm{in}}(t)$ element-wisely and put into the audio decoder. The entire model is trained end-to-end so that a scale-dependent signal-to-noise ratio (SNR) [18] is maximized:

$$\mathrm{SNR}_{\mathrm{dB}} = -\mathcal{L}_{\mathrm{on+off}} = 10\log_{10}\frac{\|s\|^2}{\|\hat{s}-s\|^2}, \tag{1}$$

where s and $\hat{s}$ are the clean and estimated signals, respectively.

When we consider extracting each target speech individually and mixing them, one of the main drawbacks is the accumulation of estimation errors, such as artifacts and remaining non-speech noises or interfering speech. In contrast, since the proposed model extracts both target speech signals in a lump, we can avoid this problem and further conduct end-to-end training. The proposed model is also computationally efficient because, unlike the mixing-based approach, we need only one model to obtain the output mixture.

B. Attention mechanism

While the visual clue is time-variant and depends on (i.e., synchronized with) the on-screen target speech [1], the voiceprint clue is time-invariant and independent of (i.e., not synchronized with) the off-screen target speech [10]. To fill this gap, we prompt the multi-speaker extractor to refer to the voiceprint clue only when the off-screen target speech temporally exists in the noisy input audio by the attention mechanism [19]. Specifically, we use the speaker-dependent voice activity detection (SDVAD) network [20], which takes noisy speech as an input and estimates whether the target speaker is active at each frame using the speaker embedding.

The off-screen target speech detection network for the SDVAD method takes as an input the element-wise multiplication between $\mathbf{z}_r^{\mathrm{a}}$ and $\mathbf{z}_{r-1}^{\mathrm{in}}(t)$ and estimates $a_r(t) \in [0,1]$ at each time frame as shown in Fig. 2(b). $a_r(t)$ represents the confidence that t-th frame of the noisy input contains the active off-screen target speech. We utilize $a_r(t)$ as an attention that controls the contribution of $\tilde{\mathbf{z}}_r^{\mathrm{a}}$ to $\mathbf{z}_r^{\mathrm{av}}(t)$ as follows:

$$\mathbf{z}_r^{\mathrm{av}}(t) = \tilde{\mathbf{z}}_r^{\mathrm{v}}(t) + a_r(t)\tilde{\mathbf{z}}_r^{\mathrm{a}}. \tag{2}$$

Following [20], the off-screen target speech detection network is trained such that the cross-entropy $\mathcal{L}_{\mathrm{CE}}$ between $a_r(t)$ and the oracle voice activity (1 if the speech exists and 0 otherwise) is minimized. The entire model is trained in a multi-task learning manner under the total loss function $\mathcal{L}_{\mathrm{total}}$:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{on+off}} + \lambda\mathcal{L}_{\mathrm{CE}}, \tag{3}$$

where λ is a hyperparameter to control the weight of the cross-entropy loss function.

C. Muting strategy

To further improve the performance of our model, we introduce a novel training strategy, which we call the muting strategy. While the proposed model refers to two clues from different domains, it outputs a mixture of two speech signals without distinction. Thus the model does not explicitly interpret the correspondence between each clue and the characteristics of the target speech of the clue. Each clue can further contribute to extracting the precise target speech by encouraging the model to learn the correspondence more precisely. We, therefore, mute one of the on-screen or off-screen target speech signals at certain probabilities during training (at p_{on} for on-screen and p_{off} for off-screen). Since this forces the model to output only either of the two target speech signals, the model strongly binds each clue and the characteristics of the corresponding signal.

IV. Evaluation

This section describes experiments to evaluate the performance of our method for multi-speaker extraction. We consider four conditions where the target mixture is contaminated by environmental sound noise and/or interfering speech.

A. Data

We used the VoxCeleb2 dataset [21], WSJ0 corpus [22], and AudioSet [23] for the on-screen target speech, off-screen target speech, and environmental sound noise, respectively. The VoxCeleb2 consists of speech signals and their synchronizing videos of the speaker's face region, while WSJ0 consists of only speech signals labeled with speaker identity. AudioSet contains audio clips labeled with multiple classes (527 classes in total), such as human voices, music, and sounds of things. For the training and validation, we used 25,000 (800 speakers), 12,776 (101 speakers), and 18,870 clips in the VoxCeleb2 training set, WSJ0 "si_tr_s" set, and the balanced training

TABLE I
COMPARISON OF THE BASELINE METHOD AND THE PROPOSED METHOD ON THE CONDITIONS WITH ENVIRONMENTAL SOUND NOISE ("NOISE") AND WITH AN INTERFERING SPEAKER ("SPK"). AM AND MS REFER TO THE ATTENTION MECHANISM AND MUTING STRATEGY, RESPECTIVELY.

Methods	AM	MS	"noise" condition		"spk" condition		#Params ↓
			SI-SDRi (dB) ↑	SDRi (dB) ↑	SI-SDRi (dB) ↑	SDRi (dB) ↑	
Baseline_A [20], [24]	-	-	7.34	7.19	7.58	7.96	29.8M
Proposed_A	-	-	7.56	7.37	8.25	8.56	25.1M
	-	✓	7.67	7.47	8.44	8.78	25.1M
	✓	-	7.77	7.57	8.29	8.56	25.1M
	✓	✓	8.06	7.88	8.73	9.11	25.1M
Baseline_r [11], [17]	-	-	7.78	7.65	8.22	8.60	53.0M
Proposed_r	-	-	8.40	8.19	9.49	9.77	**19.2M**
	-	✓	8.55	8.37	9.80	10.10	**19.2M**
	✓	-	8.42	8.21	9.71	9.98	**19.2M**
	✓	✓	**8.58**	**8.41**	**9.91**	**10.21**	**19.2M**

↑ means higher is better, and ↓ means lower is better.

subset of the AudioSet, respectively. Each set was split into a training set (80%) and a validation set (20%). For the test, we used 3,000 (118 speakers), 1,857 (18 speakers), and 3,000 clips in the VoxCeleb2 test set, WSJ0 "si_dt_05" and "si_et_05" sets, and the evaluation subset of the AudioSet, respectively. Note that, in the VoxCeleb2 and WSJ0, speakers in the test set were unseen in the training and validation set.

With these datasets, we generated 20,000, 5,000, and 3,000 pairs of noisy input audio and the synchronizing video for the training, validation, and test sets. All sounds and videos were resampled to 16 kHz and 25 fps. We generated the input audio by mixing an off-screen target speech and a four-second environmental sound noise into a four-second on-screen target speech with random SNR between -2.5 and 2.5 dB. The off-screen target speech was cropped to have a random duration between two and four seconds for the training and validation and between zero and four seconds for the test. We randomly selected the enrollment utterance from the other utterances of the off-screen target speaker. We also conducted an experiment where the target mixture is contaminated by another speaker ("spk") instead of AudioSet noise ("noise"). Moreover, experiments with both AudioSet noise and an interfering speaker ("noise+spk") and with two interfering speakers ("2spk") were conducted. In these situations, interfering speech was randomly selected from the VoxCeleb2 or WSJ0.

B. Model configuration

For the visual encoder, we used the visual encoder of AV-ConvTasnet [24] (Proposed_A) or the attractor encoder of the reentry [17] (Proposed_r), with D^{v} of 512 or 256, respectively. Note that the attractor encoder takes the additional noisy audio as an input as well as the video. In our implementation, AV-ConvTasnet was a slightly modified version [24] of the original [25]. For the voiceprint encoder, we used WASE [20], where D^{a} was 256. Finally, we used AV-ConvTasnet [24] for the audio encoder, TCN stack, and audio decoder. The parameters of the extractor block D^{av}, D^{in}, and R were set to 256, 256, and 4, respectively. λ, p_{on}, and p_{off} were set to 1.0, 20%, and 20%, respectively.

We trained all models with an Adam optimizer for 200 epochs. The learning rate was initialized to 0.001 and halved if the validation loss did not improve for three epochs. Early stopping was applied if the learning rate dropped four times. Note that the optimizations for Proposed_A and Proposed_r follow the methods used in AV-ConvTasnet and the reentry, respectively. Specifically, for Proposed_A, we optimized our entire model except for the pre-trained visual encoder's front-end [26], which consists of a 3D convolutional layer and ResNet18 block. For Proposed_r, we optimized our entire model except for the SLSyn network [17] of the visual encoder, which was pre-trained to extract speech-lip synchronization embeddings in reentry, and then fine-tuned the entire model with the re-initialized optimizer.

C. Evaluation metrics

We evaluated our method in terms of estimation accuracy and computational efficiency. We did the estimation accuracy for the multi-speaker extraction using scale-invariant signal-to-distortion ratio [27] improvement (SI-SDRi) and signal-to-distortion ratio [18] improvement (SDRi), measuring the amount of distortion in estimated signals. We did the computational efficiency by the number of model parameters.

D. Baseline methods

We set two baseline methods, which individually extract the on-screen and off-screen target speech using AV-SE and speaker extraction, respectively, and then mix them. Specifically, we used the combination of AV-ConvTasnet and WASE (Baseline_A) and that of the reentry and SpEx++ [11] (Baseline_r). We selected AV-ConvTasnet and WASE because their network and Propoed_A consist of Conv-TasNet conditioned with a clue (note that only WASE has the skip-connection paths [16]). This enables a fair comparison between the direct and separative approaches at a framework-level since the audio-visual clue is obtained with the same encoders as those of the direct one. The reentry and SpEx++ are state-of-the-art in each task. These AV-SE and speaker extraction methods were originally evaluated on the VoxCeleb2 and WSJ0, respectively. Thus we used the datasets not to deteriorate the two baseline methods rather than the proposed method. In all loss functions, we used SNR instead of the original SI-SDR to retain the scale of signals for the mixing operation.

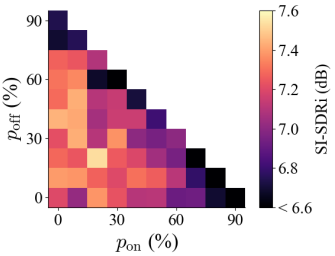


Fig. 3. Grid search of p_{on} and p_{off} for Proposed_A on the conditions with environmental sound noise ("noise").

TABLE II
EVALUATION UNDER COMPLEX INTERFERING CONDITIONS.

Interference	Methods	SI-SDRi (dB) ↑	SDRi (dB) ↑
noise+spk	Baseline_A	6.17	6.70
	Proposed_A	6.35	6.96
	Baseline_r	7.34	7.91
	Proposed_r	**7.53**	**8.14**
2spk	Baseline_A	6.28	6.73
	Proposed_A	5.66	6.19
	Baseline_r	7.46	7.94
	Proposed_r	**7.59**	**8.08**

E. Experimental results

Table I shows the experimental results on the conditions with environmental sound noise and with an interfering speaker. Proposed_A outperformed Baseline_A in SI-SDRi and SDRi, which indicates that our direct estimation method works well for accurate estimation. Our method can improve performance only by changing the visual encoder to a state-of-the-art encoder. Then, Proposed_r outperformed even the combination of the state-of-the-art methods, Baseline_r. Further, the ablation studies on the attention mechanism and muting strategy show that each method can improve the performance of our framework although the improvement is small in Proposed_r. Figure 3 shows that the muting strategy tends to be effective only when $p_{\mathrm{on}} + p_{\mathrm{off}}$ is not so large.

The computational efficiency of the proposed method is shown in the rightmost column of Table I. As we adopted the single-path framework (single sequence of the audio encoder, TCN stacks, and audio decoder) unlike the dual-path baseline framework, the number of parameters of Proposed_A is fewer than that of Baseline_A by 16%. Proposed_r is significantly lightweight compared to Baseline_r because the reentry and SpEx++ have modules that do not appear in our simple network for their state-of-the-art performance. The rightmost column also shows that we can use the attention mechanism with a very slight parameter increase and apply the muting strategy without increasing a single model parameter. Table II compares the baseline and the proposed method on the more complex interfering speech conditions. Here again, our method outperformed the combination of the state-of-the-art methods.

V. CONCLUSION

This paper presented an AV-SE method that estimates the mixture of on-screen and off-screen target speech from noisy audio. We fused two multimodal clues to extract the target mixture in a computationally efficient manner. We also introduced the attention mechanism and proposed the muting strategy to improve the performance of our model further. We experimentally confirmed that our method estimated the target mixture more accurately and efficiently compared to the baseline method. Our future work includes the evaluation of our method using more realistic data. We also plan to extend the proposed method to extract non-speech off-screen target sounds, such as alarms and sirens.

REFERENCES

[1] T. Afouras, et al., "The conversation: Deep audio-visual speech enhancement," in *Interspeech*, 2018, pp. 3244-3248.
[2] M. Gogate, et al., "CochleaNet: A robust language-independent audio-visual model for real-time speech enhancement," *Information Fusion*, vol. 63, pp. 273–285, 2020.
[3] D. Michelsanti, et al., "An overview of deep-learning based audio-visual speech enhancement and separation," *IEEE/ACM TASLP*, vol. 29, pp. 1368–1396, 2021.
[4] K. Yang, et al., "Audio-visual speech codecs: Rethinking audio-visual speech enhancement by re-synthesis," in *IEEE/CVF CVPR*, 2022, pp. 8227–8237.
[5] A. Ephrat, et al., "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM TOG*, vol. 37, no. 4, pp. 1–11, 2018.
[6] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.
[7] L. Girin, et al., "Audio-visual enhancement of speech in noise," *The Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 3007–3020, 2001.
[8] T. Afouras, et al., "My lips are concealed: Audio-visual speech enhancement through obstructions," in *Interspeech*, 2019, pp. 4295–4299.
[9] H. Sato, et al., "Multimodal attention fusion for target speaker extraction," in *IEEE SLT*, 2021, pp. 778–784.
[10] Q. Wang, et al., "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Interspeech*, 2019, pp. 2728–2732.
[11] M. Ge, et al., "Multi-stage speaker extraction with utterance and frame-level reference signals," in *IEEE ICASSP*, 2021, pp. 6109-6113.
[12] T. Ochiai, et al., "Listen to what you want: Neural network-based universal sound selector," in *Interspeech*, 2020, pp. 1441–1445.
[13] S. Liu, et al., "N-HANS: A neural network-based toolkit for in-the-wild audio enhancement," *Multimedia Tools and Applications*, vol. 80, pp. 28365–28389, 2021.
[14] S. Pascual, et al., "SEGAN: Speech enhancement generative adversarial network," in *Interspeech*, 2017, pp. 3642–3646.
[15] Y. Hu, et al., "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Interspeech*, 2020, pp. 2472–2476.
[16] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM TASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.
[17] Z. Pan, et al., "Selective listening by synchronizing speech with lips," *IEEE/ACM TASLP*, vol. 30, pp. 1650–1664, 2022.
[18] E. Vincent, et al., "Performance measurement in blind audio source separation," *IEEE/ACM TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.
[19] S.-W. Chung, et al., "FaceFilter: Audio-visual speech separation using still images," in *Interspeech*, 2020, pp. 3481–3485.
[20] Y. Hao, et al., "Wase: Learning when to attend for speaker extraction in cocktail party environments," in *IEEE ICASSP*, 2021, pp. 6104–6108.
[21] J. S. Chung, et al., "VoxCeleb2: Deep speaker recognition," in *Interspeech*, 2018, pp. 1086–1090.
[22] J. Garofolo, et al., "CSR-I (WSJ0) Complete LDC93S6A," *Philadelphia: Linguistic Data Consortium*, 1993.
[23] J. F. Gemmeke, et al., "Audio Set: An ontology and human-labeled dataset for audio events," in *IEEE ICASSP*, 2017, pp. 776–780.
[24] Z. Pan, et al., "Muse: Multi-modal target speaker extraction with visual cues," in *IEEE ICASSP*, 2021, pp. 6678–6682.
[25] J. Wu, et al., "Time domain audio visual speech separation," in *IEEE ASRU*, 2019, pp. 667–673.
[26] T. Afouras, et al., "Deep audio-visual speech recognition," *IEEE TPAMI*, 2018.
[27] J. L. Roux, et al., "SDR–half-baked or well done?," in *IEEE ICASSP*, 2019, pp. 626–630.