INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Multimodal SpeakerBeam: Single channel target speech extraction with audio-visual speaker clues

Tsubasa Ochiai, Marc Delcroix, Keisuke Kinoshita, Atsunori Ogawa, Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
tsubasa.ochiai.ah@hco.ntt.co.jp

Abstract

Recently, with the advent of deep learning, there has been significant progress in the processing of speech mixtures. In particular, the use of neural networks has enabled target speech extraction, which extracts speech signal of a target speaker from a speech mixture by utilizing auxiliary clue representing the characteristics of the target speaker. For example, audio clues derived from an auxiliary utterance spoken by the target speaker have been used to characterize the target speaker. Audio clues should capture the fine-grained characteristic of the target speaker's voice (e.g., pitch). Alternatively, visual clues derived from a video of the target speaker's face speaking in the mixture have also been investigated. Visual clues should mainly capture the phonetic information derived from lip movements. In this paper, we propose a novel target speech extraction scheme that combines audio and visual clues about the target speaker to take advantage of the information provided by both modalities. We introduce an attention mechanism that emphasizes the most informative speaker clue at every time frame. Experiments on mixture of two speakers demonstrated that our proposed method using audio-visual speaker clues significantly improved the extraction performance compared with the conventional methods using either audio or visual speaker clues.

Index Terms: target speech extraction, speaker-aware mask estimation network, audio-visual speaker clues

1. Introduction

The recent development of deep learning has led to the active investigation of neural network-based single channel source separation approaches for the enhancement of speech signals corrupted by overlapping speakers. Most such studies focus on blind source separation (BSS) approaches such as deep clustering [1] or permutation invariant training [2], which separates an observed speech mixture into each of its sources without any prior information (i.e., by utilizing only the observed speech mixture). BSS approaches can work in a fully blind fashion, but they generally suffer from a global permutation ambiguity issue, i.e. the mapping between the speakers in the mixture and the outputs is arbitrary. To address this ambiguity issue, recent studies have also investigated target speech extraction approaches [3–8], which extract only a speech signal from a target speaker from an observed speech mixture by utilizing additional clues characterizing the target speaker. By utilizing the clue of the target speaker, the target speech extraction scheme does not suffer from the global permutation ambiguity issue by its nature, i.e., possible to track a specific speaker's voice across utterances. Moreover, it has the potential to achieve better speech quality than the BSS scheme by exploiting the additional information in the extraction stage [4, 6].

Target speech extraction research has focused on exploiting either audio [3–5] or visual speaker clues [6–8]. Both approaches have provided promising results as regards target speech extraction, and are shown to have different merits and demerits. Previous studies using audio speaker clues, e.g., SpeakerBeam [3, 4], assume that only the audio stream (for input mixture and speaker clue) is available to extract the target speaker, and the speaker clue, i.e., the target speaker's voice, can be prerecorded. [3–5] used a prerecorded utterance spoken by the target speaker as the speaker clue, to adapt the network behavior so that it extracted the target speaker from the observed speech mixture. An audio-based clue should capture the fine-grained characteristics of the target speaker's voice (e.g., pitch, timbre, etc.). It has been confirmed that with audio clues, the extraction performance degrades when the speakers in the mixture have similar voice characteristics to those of the target speaker [4].

On the other hand, previous studies using visual speaker clues, e.g., [6–8], assume that audio stream (for input mixture) and visual stream (for speaker clue) are available for extracting the target speaker. As the speaker clue, [6–8] utilized the cropped face region of a video recording of the target speaker in the mixture. A visual-based clue should mainly capture phonetic information from the target speaker's lip movement and may help us better handle mixtures that contain the voices with similar characteristics [6]. However, the quality of the visual clue could be affected by the target speaker's behavior at every time instant and in practice could suffer from face movement or occlusions, unlike the audio clue whose quality does not change once it has been prerecorded.

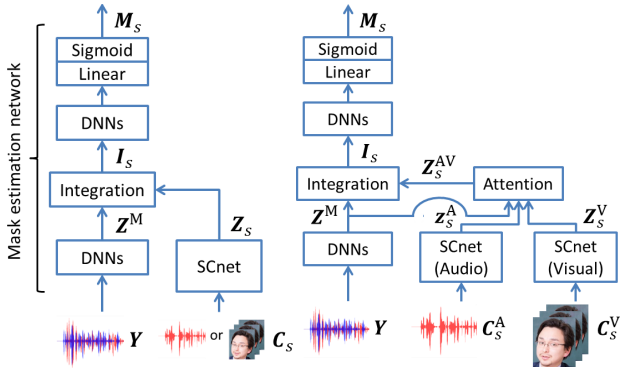
In this paper, to take advantage of both audio and visual speaker clues, and to increase robustness against the lack or corruption of either clue, we propose a novel target speech extraction scheme using multiple (audio-visual) speaker clues, which we call multimodal (audio-visual) SpeakerBeam. In contrast to the conventional scheme using a single modality as a clue of the target speaker, the proposed scheme exploits both audio and visual modalities for speaker clues. Motivated by the success of the attention-based audio-visual fusion in automatic speech recognition [9] and video description [10] tasks, this paper proposes an attention mechanism that integrates the audio and visual speaker clues. With the proposed attention scheme, we are able to exploit more informative speaker clue for the target speech extraction at every time frame. In addition, we propose a multitask learning-based training procedure [11] that simultaneously considers the extraction loss using audio-visual, audio-only, and visual-only speaker clues. It enables the proposed extraction system for audio-visual speaker clues to work even when either audio or visual speaker clues are unavailable.

2. Conventional target speech extraction

2.1. Framework

In this paper, we employ a mask-based approach [12] to extract the target speaker from an observed speech mixture. The speech signal of the target speaker s, i.e., $\hat{\mathbf{X}}_s$, is extracted by applying the time-frequency mask $\mathbf{M}_s \in \mathbb{R}^{T \times F}$ to the observed speech

http://dx.doi.org/10.21437/Interspeech.2019-1513



(a) Single speaker clue (b) Multiple speaker clues

Figure 1: *Overview of speaker-aware mask estimation network.*

mixture $\mathbf{Y} \in \mathbb{C}^{T \times F}$ in the short-time Fourier transform (STFT) domain, as follows:

$$\hat{\mathbf{X}}_s = \mathbf{M}_s \odot \mathbf{Y}, \tag{1}$$

where $\odot$ denotes an element-wise product, and T and F denote the number of time frames and frequency bins, respectively.

In the target speech extraction setup, we assume that an additional speaker clue $\mathbf{C}_s$ is available when extracting the speech signal of the target speaker from the mixture. The time-frequency mask of the target speaker $\mathbf{M}_s$ is estimated by the speaker-aware mask estimation network, as follows:

$$\mathbf{M}_s = \mathrm{DNN}(|\mathbf{Y}|, \mathbf{C}_s), \tag{2}$$

where $\mathrm{DNN}(\cdot)$ is the non-linear transformation of a deep neural network (DNN), and $|\mathbf{Y}|$ denotes the amplitude spectrum coefficients of $\mathbf{Y}$.

Figure 1-(a) shows a schematic diagram of a typical network architecture for speech extraction, which consists of a speaker clue extraction network (SCnet) and a mask estimation network that exploits speaker clues through an integration module. Given the speaker clue $\mathbf{C}_s$ as an input, the SCnet generates an intermediate representation of the speaker clue $\mathbf{Z}_s$. The integration module of the mask estimation network combines this representation $\mathbf{Z}_s$ with an intermediate feature of the observed speech mixture $\mathbf{Z}^{\mathrm{M}} = \{\mathbf{z}_t^{\mathrm{M}}; t = 1, 2, \cdots, T\}$ derived from the bottom layers of the mask estimation network. The output of the integration module $\mathbf{I}_s = \{\mathbf{i}_{st}; t = 1, 2, \cdots, T\}$ corresponds to an intermediate representation of the target speaker, which makes it possible to predict the target speaker's time-frequency mask $\mathbf{M}_s$ with the top layers of the mask estimation network. By integrating the additional speaker clue in the mask estimation network, the network behavior can be adapted for the extraction of the target speaker.

We can consider several ways to perform the above integration procedure. For example, [4] adopted element-wise product-based integration, such as $\mathbf{I}_s = \mathbf{Z}^{\mathrm{M}} \odot \mathbf{Z}_s$. In contrast, [6, 7] adopted concatenation-based integration, such as $\mathbf{I}_s = [\mathbf{Z}^{\mathrm{M}}, \mathbf{Z}_s]$, where $[\cdot]$ represents concatenation over the feature dimension. In the following, to be consistent with our previous work [4], we adopt element-wise product-based integration.

2.2. Speech extraction based on audio speaker clue

For an audio speaker clue [3–5], the additional information consists of a sequence of STFT-based amplitude spectrum features $\mathbf{C}_s^{\mathrm{A}} \in \mathbb{R}^{T^{\mathrm{A}} \times F}$ derived from a prerecorded utterance spoken by the target speaker, where T^{A} is the number of time frames for the audio speaker clue.

Given the audio speaker clue $\mathbf{C}_s^{\mathrm{A}}$ as an input, the integrated features $\mathbf{i}_{st}^{\mathrm{A}} \in \mathbb{R}^{1 \times H}$ are computed as follows:

$$\mathbf{z}_s^{\mathrm{A}} = \mathrm{Avg}(\mathrm{SCnet}^{\mathrm{A}}(\mathbf{C}_s^{\mathrm{A}})), \tag{3}$$

$$\mathbf{i}_{st}^{\mathrm{A}} = \mathbf{z}_t^{\mathrm{M}} \odot \mathbf{z}_s^{\mathrm{A}} \quad (t = 1, 2, \cdots, T), \tag{4}$$

where $\mathbf{z}_s^{\mathrm{A}} \in \mathbb{R}^{1 \times H}$ represents the extracted features for the audio clue, $\mathrm{SCnet}^{\mathrm{A}}(\cdot)$ is the feature extraction network for the audio clue, $\mathrm{Avg}(\cdot)$ is the average operation over the time axis, and H is the output dimension of $\mathrm{SCnet}^{\mathrm{A}}(\cdot)$. $\mathrm{Avg}(\mathrm{SCnet}^{\mathrm{A}}(\cdot))$ corresponds to a sequence summary network [13] that extracts the speech characteristics directly from the input utterance.

Note that with audio speaker clues, $\mathrm{SCnet}^{\mathrm{A}}(\cdot)$ maps the sequence of amplitude spectrum coefficients $\mathbf{C}_s^{\mathrm{A}}$ to a vector as shown in Eq. (3). The audio clue is thus *time-invariant*.

2.3. Speech extraction based on visual speaker clue

When using a visual speaker clue [6–8], the additional information consists of video-based features $\mathbf{C}_s^{\mathrm{V}} \in \mathbb{R}^{T^{\mathrm{V}} \times D}$ derived from the cropped face region of the target speaker. Following [6], this paper adopted face embedding features extracted with a pre-trained face recognition model, i.e., Facenet [14], as the video-based features. Here, T^{V} is the number of time frames for the visual speaker clue, and D is the dimension of face embeddings.

Given the visual speaker clue $\mathbf{C}_s^{\mathrm{V}}$ as an input, the integrated features $\mathbf{i}_{st}^{\mathrm{V}} \in \mathbb{R}^{1 \times H}$ are computed as follows:

$$\mathbf{Z}_s^{\mathrm{V}} = \mathrm{SCnet}^{\mathrm{V}}(\mathbf{C}_s^{\mathrm{V}}), \tag{5}$$

$$\mathbf{i}_{st}^{\mathrm{V}} = \mathbf{z}_t^{\mathrm{M}} \odot \mathbf{z}_{st}^{\mathrm{V}} \quad (t = 1, 2, \cdots, T), \tag{6}$$

where $\mathbf{Z}_s^{\mathrm{V}} = \{\mathbf{z}_{st}^{\mathrm{V}}; t = 1, 2, \cdots, T^{\mathrm{V}}\}$ is the extracted feature for the visual clue, and $\mathrm{SCnet}^{\mathrm{V}}(\cdot)$ is the feature extraction network for the visual clue.

Note that, in contrast to the audio speaker clue, the visual speaker clue is used as *time-variant* clue as in Eq. (5). As a common situation in the audio-visual processing, there exists a gap in the number of frames per second (fps) between audio and visual streams; e.g., video is at 25 fps (40 ms) while audio is at 50 fps (20 ms). It is necessary that every video frame of the input sequence corresponds to the audio frames. In this paper, we aligned them by repeating a video frame for several audio frames, e.g., $\tilde{\mathbf{Z}}^{\mathrm{V}} = \{\mathbf{z}_{s,1}^{\mathrm{V}}, \mathbf{z}_{s,1}^{\mathrm{V}}, \mathbf{z}_{s,2}^{\mathrm{V}}, \mathbf{z}_{s,2}^{\mathrm{V}}, \cdots\}$

3. Proposed multimodal SpeakerBeam

3.1. Attention-based fusion of audio-visual speaker clues

In the proposed method, we assume that the multiple speaker clues (i.e., audio and visual) are available when extracting the signal of the target speaker from the observed speech mixture. To utilize the advantages of both types of speaker clue, this paper proposes an attention-based fusion mechanism to integrate the audio and visual speaker clues. The role of the attention mechanism is to emphasize (i.e., softly select) more informative speaker clues for the extraction of the target speaker, at every time frame of the mixture.

Figure 1-(b) is a schematic diagram of the proposed network architecture, which extends the feature extraction process of the speaker clues by adding the attention-based fusion mechanism. Given multiple speaker clues, i.e., $\mathbf{C}_s^{\mathrm{A}}$ for audio clue and $\mathbf{C}_s^{\mathrm{V}}$ for visual clue, the feature extraction modules transform them into intermediate feature sequences $\mathbf{z}_s^{\mathrm{A}}$ and $\mathbf{Z}_s^{\mathrm{V}}$, respectively, as described in Eqs. (3) and (5). Then, the attention

mechanism combines these speaker clues $\mathbf{z}_s^{\mathrm{A}}$ and $\mathbf{Z}_s^{\mathrm{V}}$ into the intermediate feature sequences for the audio-visual speaker clue $\mathbf{Z}_s^{\mathrm{AV}} = \{\mathbf{z}_{st}^{\mathrm{AV}}; t = 1, 2, \cdots, T\}$. Finally, the network generates time-frequency masks $\mathbf{M}_s^{\mathrm{AV}}$ based on the audio-visual speaker clue $\mathbf{Z}_s^{\mathrm{AV}}$ in a similar way as described in Section 2.1.

Given both the audio and visual speaker clues $\mathbf{C}_s^{\mathrm{A}}$ and $\mathbf{C}_s^{\mathrm{V}}$ as an input, the integrated features $\mathbf{i}_{st}^{\mathrm{AV}} \in \mathbb{R}^{1\times H}$ based on the audio-visual speaker clue are computed as follows:

$$\mathbf{z}_{st}^{\mathrm{AV}} = \underbrace{\sum_{\psi\in\{\mathrm{A,V}\}} a_{st}^{\psi}\mathbf{z}_{st}^{\psi}}_{\text{Attention}} \quad (t = 1, 2, \cdots, T), \tag{7}$$

$$\mathbf{i}_{st} = \mathbf{z}_t^{\mathrm{M}} \odot \mathbf{z}_{st}^{\mathrm{AV}} \quad (t = 1, 2, \cdots, T), \tag{8}$$

where $\{a_{st}^{\psi}\}_{\psi\in\{\mathrm{A,V}\}}$ are the attention weights at time step t for the target speaker s.

We adopt the additive attention mechanism proposed in [15] to compute the attention weights. The attention weights $\{a_{st}^{\psi}\}_{\psi\in\{\mathrm{A,V}\}}$ are computed from the intermediate feature of the mixture $\mathbf{z}_t^{\mathrm{M}}$ and the speaker clues $\{\mathbf{z}_{st}^{\psi}\}_{\psi\in\{\mathrm{A,V}\}}$ as follows:

$$e_{st}^{\psi} = \mathbf{w}\tanh(\mathbf{W}\mathbf{z}_t^{\mathrm{M}} + \mathbf{V}\mathbf{z}_{st}^{\psi} + \mathbf{b}), \tag{9}$$

$$a_{st}^{\psi} = \frac{\exp(\epsilon e^{\psi})}{\sum_{\psi\in\{\mathrm{A,V}\}}\exp(\epsilon e^{\psi})}, \tag{10}$$

where $\mathbf{w}, \mathbf{W}, \mathbf{V}, \mathbf{b}$ are trainable weight and bias parameters, and ϵ is a sharpening factor [15]. Here, for the audio speaker clues, we used the time-invariant (global) clue for all of the time frames, i.e., $\mathbf{z}_{st}^{\mathrm{A}} = \mathbf{z}_s^{\mathrm{A}}$ $(t = 1, 2, \cdots, T)$.

3.2. Multitask learning-based training procedure

We assume that a set of input and target features $\{\mathbf{Y}, \mathbf{C}_i^{\mathrm{A}}, \mathbf{C}_i^{\mathrm{V}}, \mathbf{X}_i\}_{i=1}^{I}$ is available for training the model, where $\mathbf{X}_i \in \mathbb{C}^{T\times F}$ is the target speech signal of the i-th speaker in the mixture, and I denotes the number of speakers in the mixture.

We propose using multitask learning (MTL) to enable the proposed audio-visual extraction system to work even when either audio or speaker clue is unavailable. The multitask learning-based objective function L_{MTL} considers three situations, namely 1) audio-visual clues are available, 2) only audio clues are available, and 3) only visual clues are available, as follows:

$$L_{\mathrm{MTL}} = \alpha L_{\mathrm{AV}} + \beta L_{\mathrm{A}} + \gamma L_{\mathrm{V}}, \tag{11}$$

$$L_{\psi} = \frac{1}{I}\sum_{i=1}^{I} l(\mathbf{M}_i^{\psi} \odot |\mathbf{Y}|, |\mathbf{X}_i|), \tag{12}$$

where $\psi \in \{\mathrm{AV, A, V}\}$, a set of parameters $\{\alpha, \beta, \gamma\}$ are multitask weights, and $l(\mathbf{A}, \mathbf{B}) = \frac{1}{TF}\|\mathbf{A} - \mathbf{B}\|^2$ is the mean squared error (MSE) criterion. Here, $\mathbf{M}_i^{\mathrm{AV}}$, $\mathbf{M}_i^{\mathrm{A}}$, and $\mathbf{M}_i^{\mathrm{V}}$ indicate the time-frequency masks estimated based on the intermediate features $\mathbf{Z}_i^{\mathrm{AV}}$, $\mathbf{z}_i^{\mathrm{A}}$, and $\mathbf{Z}_i^{\mathrm{V}}$ in Eqs. (3), (5), and (7), respectively. Note that when audio or visual clues are used alone, i.e., $\psi = \{\mathrm{A}\}$ or $\psi = \{\mathrm{V}\}$, the attention weight for that clue a_{st}^{ψ} becomes 1 (see Eq. (10)).

4. Experiments

We compared our proposed extraction method using multiple speaker clues (SpeakerBeam-AV, SpeakerBeam-AV-MTL) with two conventional extraction methods using a single speaker clue (Baseline-A [4], Baseline-V [6, 7]). SpeakerBeam-AV and SpeakerBeam-AV-MTL correspond to the proposed method with both audio and visual clues, where we set the multitask weights at $\{\alpha = 1.0, \beta = 0.0, \gamma = 0.0\}$ and $\{\alpha = 0.8, \beta = 0.1, \gamma = 0.1\}$, respectively. Baseline-A and Baseline-V[1] correspond to the conventional method with audio or visual clues, respectively. Note that Baseline-A can be regarded as $\{\alpha = 0.0, \beta = 1.0, \gamma = 0.0\}$ and Baseline-V as $\{\alpha = 0.0, \beta = 0.0, \gamma = 1.0\}$.

4.1. Experimental conditions

4.1.1. Data

To evaluate the effectiveness of our proposed method, we created a simulation dataset of speech mixtures based on the Lip Reading Sentences 3 (LRS3-TED) audio-visual corpus [16]. Our dataset consisted of two-speaker mixtures generated by mixing utterances at a signal-to-noise ratio (SNR) between 0 and 5 dB, in a similar way to that employed with the widely-used WSJ0-2mix corpus [1]. Moreover, we also downsampled speech to 8 kHz to reduce the computational and memory costs.

The training set consisted of 50000 mixtures from 500 speakers. The development set consisted of 10000 mixtures from 300 speakers. The speakers used for the training and development sets were randomly selected from the *pre-train* and *train-val* sets in the LRS3-TED corpus. The test set consisted of 5000 mixtures from 295 speakers based on the *test* set in the LRS3-TED corpus. The number "295" corresponds to the number of speakers who have more than two utterances in the *test* set of the LRS3-TED corpus.

For the visual speaker clues, we used the video data corresponding to each speaker in the mixture. For the audio speaker clue, we randomly selected an utterance of the same speaker in the database that was not used to generate the speech mixture.

4.1.2. Settings

As the audio features, we used amplitude spectrograms computed using an STFT with a 64 ms window length and 20 ms window shift. As the visual features, we used Facenet-based features, which are extracted for every video frame (at 25 fps, i.e., 40 ms shift) using the software and pre-trained model provided in the GitHub repository [17].

For all the experiments, we used a 3-layer BLSTM network with 512 units. Each BLSTM layer was followed by a linear projection layer with 512 units to combine the forward and backward LSTM outputs. We employed one fully connected layer to output an amplitude mask estimated with a sigmoid activation function. For the integration layer, we adopted element-wise product-based integration[2]. The integration layer was inserted after the first BLSTM layer.

For the feature extraction network of the audio clue (i.e., $\mathrm{SCnet}^{\mathrm{A}}(\cdot)$ in Eq. (3)), we used a network with 2 fully connected layers with 200 units and ReLU activations, followed by 1 linear output layer with 512 units. On the other hand, for the feature extraction network of the visual clue (i.e., $\mathrm{SCnet}^{\mathrm{V}}(\cdot)$ in Eq. (5)), we used a network with 3 convolution layers with 256 channels (filter=7x1,5x1,5x1, shift=1x1,1x1,1x1), where spatial convolutions were performed over the temporal axis inspired by [6],

[1] When compared with [6, 7], Baseline-V employed a simpler network architecture, e.g., without a phase recovery framework. However, it allows a fair comparison with Baseline-A, while maintaining the fundamental characteristics needed to evaluate the effectiveness of the proposed multimodal (audio-visual) speaker clues.

[2] In our preliminary experiment, we observed that element-wise product-based integration and concatenation-based integration worked comparably well for both audio and visual clues.

Table 1: *SDR (dB) for evaluated methods with audio-only, visual-only, and audio-visual speaker clues.*

Method	Diff	Same	All
Mixture	0.5	0.5	0.5
Baseline-A	9.8	6.8	8.3
Baseline-V	9.4	7.1	8.3
SpeakerBeam-AV	10.7	9.1	9.9

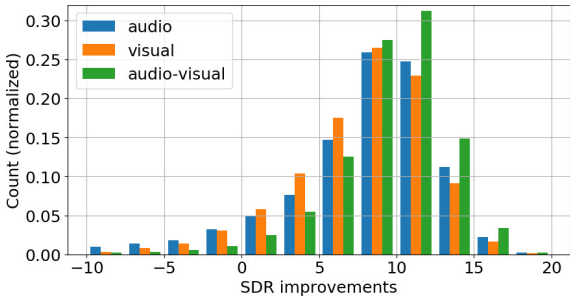


Figure 2: *Histograms of SDR improvements.*

followed by 1 linear output layer with 512 units. We adopted ReLU activation and batch normalization [18] for every convolution layer. We set the dimension of the attention inner product (i.e., the dimension of $\mathbf{w}$ in Eq. (9)) at 200, and the sharpening factor ϵ at 2.

We adopted the Adam algorithm [19] for optimization with an initial learning rate of 0.0001 and used gradient clipping [20]. We stopped the training procedure after 200 epochs.

We evaluated the results in terms of the signal-to-distortion ratio (SDR) computed with the BSS Eval toolbox [21]. All the experimental results were obtained by averaging the extraction performance of both speakers in the mixture.

4.2. Experimental results

4.2.1. Evaluation: Single clue vs. Multiple clues

Table 1 shows the SDR scores of unprocessed mixture, baselines, and the proposed SpeakerBeam-AV. "Diff" and "Same" denote the scores for mixtures of different genders and the same gender, respectively. "All" denotes the scores averaged over all mixtures.

From Table 1, we confirmed that the conventional Baseline-A (audio clue) and Baseline-V (visual clue) performed comparably in this experimental setup. Moreover, the gender-based results showed that Baseline-A worked slightly better with different genders, while Baseline-V worked slightly better for the same gender.

The proposed SpeakerBeam-AV (audio-visual clues) successfully outperformed the conventional Baseline-A and Baseline-V, which use a single speaker clue. Specifically, we confirmed that SpeakerBeam-AV significantly improved the extraction performance for the same gender.

4.2.2. Analysis of performance improvement

We investigated the performance improvement of the proposed method in more detail. Figure 2 shows histograms of the SDR improvement for Baseline-A (audio), Baseline-V (visual), and SpeakerBeam-AV (audio-visual), where each histogram bin represents a 2.5 dB interval and the vertical axis shows the (normalized) count of evaluated mixtures in each bin.

Focusing on the histogram bins below 2.5 dB, which indicate poor extraction performance (e.g., the system extracted the speech signal of the interfering speaker, not the target speaker), we can observe that the proposed SpeakerBeam-AV using multiple speaker clues significantly reduced the number of the utterances with such lower SDR scores (i.e., reduced the rate of the extraction failure) compared with the conventional Baseline-A and Baseline-V using single speaker clues. This indicates the more stable and robust behavior of the proposed SpeakerBeam-AV.

Table 2: *SDR (dB) for proposed methods without and with multitask learning.*

Method	Weights $\{\alpha,\ \beta,\ \gamma\}$	Clues AV	A	V
SpeakerBeam-AV	$\{1.0,\ 0.0,\ 0.0\}$	9.9	6.7	1.1
SpeakerBeam-AV-MTL	$\{0.8,\ 0.1,\ 0.1\}$	9.9	8.6	9.0

4.2.3. Evaluation of multitask learning effect

Table 2 shows SDR scores, averaged over all mixtures (i.e., "All" in Table 1), obtained with the proposed systems trained on single-task (SpeakerBeam-AV) and multitask (SpeakerBeam-AV-MTL) objectives. "Weights" denotes the multitask weights in Eq. (11). "Clues" denotes the speaker clue used in the extraction stage; audio (A), visual (V), and audio-visual (AV) clues.

We confirm that the performance of SpeakerBeam-AV degraded when using single speaker clues (especially, Clues = V). On the other hand, SpeakerBeam-AV-MTL could achieve better performance than Baseline-A and Baseline-V (see Table 1) even in such a situation, while maintaining the performance with both audio and visual clues (Clues = AV). This result demonstrates that the multitask objective is effective in enabling the proposed audio-visual extraction system to work even when either the audio or visual speaker clue is unavailable. In addition, the result suggests that the use of multiple speaker clues in the training stage is effective in improving the extraction performance of the system using a single speaker clue in the extraction stage.

5. Conclusion

This paper proposed a novel target speech extraction scheme that uses multiple (audio-visual) speaker clues. We introduced an attention-based mechanism to integrate the audio and visual clues and the multitask learning-based training procedure. The experimental results showed that our proposed multimodal SpeakerBeam using audio-visual clues significantly improved the extraction performance compared with that of conventional baselines using a single audio or visual clue. In addition, we observed that the proposed multitask learning scheme could improve the performance of the proposed audio-visual extraction system when either the audio or visual speaker clue was unavailable.

In this paper, we evaluated our proposed method in a controlled setup, where good quality audio and visual clues were both always available for all of the mixtures. Future work will include a more detailed investigation of the effectiveness of the attention-based mechanism in dealing with more informative speaker clues for every time frame, in a more challenging (realistic) setup, e.g., 1) when the audio speaker clues are noisy or short, or 2) when some of the visual clues are missing because of face movement or occlusion. In addition, we plan to conduct an evaluation using a larger dataset and investigate network architectures with more representation power.

6. References

[1] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31–35.

[2] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1901–1913, 2017.

[3] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5554–5558.

[4] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, S. Araki, and T. Nakatani, "Compact network for SpeakerBeam target speaker extraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, (to be published).

[5] J. Wang, J. Chen, D. Su, L. Chen, M. Yu, Y. Qian, and D. Yu, "Deep extractor network for target speaker recovery from single channel speech mixtures," in *Interspeech*, 2018, pp. 307–311.

[6] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, 2018.

[7] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Interspeech*, 2018, pp. 3244–3248.

[8] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *Interspeech*, 2018, pp. 1170–1174.

[9] P. Zhou, W. Yang, W. Chen, Y. Wang, and J. Jia, "Modality attention for end-to-end audio-visual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, (to be published. preprint found in arXiv.).

[10] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *IEEE international conference on computer vision (ICCV)*, 2017, pp. 4193–4202.

[11] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[12] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[13] K. Veselỳ, S. Watanabe, K. Žmolíková, M. Karafiát, L. Burget, and J. H. Černockỳ, "Sequence summarizing neural network for speaker adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5315–5319.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE conference on computer vision and pattern recognition (CVPR)*, 2015, pp. 815–823.

[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations (ICLR)*, 2015.

[16] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[17] https://github.com/davidsandberg/facenet.

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (ICMR)*, 2015, pp. 448–456.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[20] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning (ICML)*, 2013, pp. 1310–1318.

[21] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing (TASLP)*, vol. 14, no. 4, pp. 1462–1469, 2006.