INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Background-Sound Controllable Voice Source Separation

Deokjun Eom[1], *Woo Hyun Nam*[1], *Kyung-Rae Kim*[1]

[1]Samsung Research, Samsung Electronics, Republic of Korea

dj.eom@samsung.com, woohyun.nam@samsung.com, kr728.kim@samsung.com

Abstract

There have been various approaches to separate mixed voices. In the real world, input voices contain many different kinds of background sounds but existing methods have not considered the background sounds in model architectures. These approaches are difficult to control the background sounds directly and the voice separation results include the background sounds randomly. In this paper, we propose an extended voice separation framework, background-sound controllable voice source separation that can control the degrees of background sounds of voice separation outputs using a control parameter that ranges from 0 to 1 without additional mixing procedures. Several experiments show the controllability of background sounds on various real world datasets with preserving voice separation performances.

Index Terms: background-sound controllable, voice source separation, speech separation, deep learning.

1. Introduction

Voice source separation is to separate mixed voices from an original source. There are many approaches to utilize the voice source separation technology in the real life. For example, you may want to concentrate on your son or daughter's voice in a video that includes other voices or background sounds and these sounds can disturb your concentration. In this case, you want to remove other voices and background sounds, or just remove other voices and leave the background sounds due to naturalness of the audio source. Some background sounds are just noises that need to be removed, but some background sounds like music may not disturb your concentration. Therefore, background sounds need to be considered in voice source separation.

Deep neural networks are the most effective frameworks on voice source separation tasks and there have been various approaches to separate mixed voices based on deep neural networks. In [1, 2, 3, 4], the authors propose audio-only speech separation based on convolutional neural network (CNN), long short-term memory (LSTM), or Transformer [5] structures. Time-domain signals or short-time Fourier transform (STFT) is the input of these models and fixed number of mixed voices are separated. Permutation invariant training [6, 7, 8] solves label permutation problem of audio-only speech separation. These approaches choose the number of target speakers before training and always separate fixed number of voices. Another approaches in [9, 10, 11] use prior speeches of a target speaker to separate one target speaker voice from mixed audios.

The other recent works in [12, 13] are also deep neural network based models and use visual information of target speakers including lip motion, face image, or face detection videos. The architectures are based on U-Net [14] and CNN networks [13], or Transformer structures [5, 12].

However, existing models do not consider the background sounds contained in mixed sources and the voice separation results include the background sounds randomly. It is not predictable how much background sounds are in separation results. In addition, computing the background sounds after separating all mixed voices and subtracting the separated voices from an original mixed source can make artifact sounds and some voices can remain in computed background sounds. Another example to control the background sounds is to use speech enhancement models [15, 16] after separating target voices. However, processing outputs of models multiple times can also accumulate artifacts, so the final outputs may not be close to ground truths. These mean that we need to control the background sounds in a model and additional modules are required to extract background sound features.

In this paper, we propose an extended voice separation framework, background-sound controllable voice source separation that can control the degrees of background sounds of voice separation outputs using a control parameter $\alpha \in [0, 1]$. The proposed model does not require additional post-processing or audio mixing procedure to control the background sounds. An input of the proposed approach is STFT of mixed voices and a control parameter, and the output is a separated target voice with background sounds controlled by α.

The contributions of this paper are as follows. First, we propose a background sound feature extractor that extracts features of background sounds, and the features are used to determine the skip-connected features that are generated from each separation encoder block. Second, we devise a background sound controller with a control activation function that is four times of derivative of sigmoid function. The skip-connected features from separation encoder blocks are multiplied by outputs of the control activation function, and in this way, the features that are related to the background sounds are controlled by control parameter α and features from the background sound feature extractor.

The remainder of this paper is organized as follows. In section 2, we present the proposed method for background-sound controllable voice source separation. We then provide qualitative and quantitative experiments on real-world datasets to show controllability of background sounds in section 3. Finally, we draw the conclusion with a furture work in section 4.

2. Proposed Method

Given a mixed audio $\mathbf{m} = \sum_{i=1}^{N} \mathbf{v}_i + \mathbf{b} \in \mathbb{R}^T$, where N is the number of speakers, T is the length of the audio, $\mathbf{v}_i \in \mathbb{R}^T$ is a voice of speaker i, and $\mathbf{b} \in \mathbb{R}^T$ is a background sound. The goal of our approach is to separate a target speaker i and

10.21437/Interspeech.2023-185

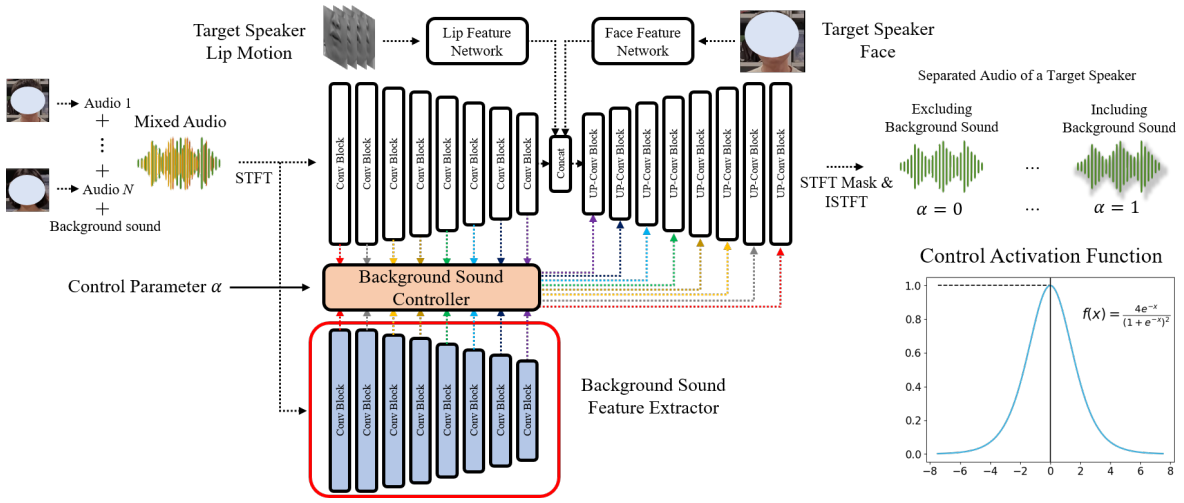

Figure 1: *The entire structure of our proposed model. We introduce a background sound feature extractor that analyzes representations of background sounds and a background sound controller that determines the magnitude of skip-connection features from separation encoder blocks. The detailed description of background sound controller is shown in Figure 2*

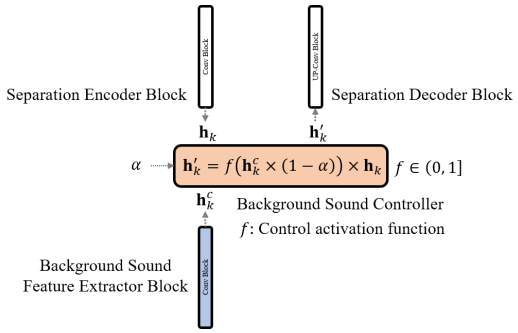

Figure 2: *Detailed description of background sound controller. A same color of dotted arrows in Figure 1 follows these computing steps.*

control the background sound $\mathbf{b}$ with the control parameter α simultaneously. Therefore, the final output of the model is $\mathbf{v}_i + \alpha\mathbf{b}$. In this framework, it is not necessary to extract $\mathbf{v}_i$ and $\mathbf{b}$ separately and mix two sources. The proposed model consists of a voice separation network and background sound control network. We use base structures of VisualVoice [13] as a voice separation network, and newly introduce a background sound control network in this paper.

2.1. Baseline Voice Separation Network

The voice separation network consists of separation encoder & decoder modules, lip feature network, and face feature network. First, we convert a time domain audio $\mathbf{m}$ to frequency domain $\mathbf{M} \in \mathbb{R}^{2\times K\times T'}$ using short-time Fourier transform (STFT), where K is the number of frequency bins and T' is the number of frames. The encoder has 8 convolution blocks and each block contains 2d-convolution, Batch Normalization [17], and LeakyReLU. The decoder is symmetric to the encoder, so the decoder has also 8 convolution blocks and each block contains upsampling, 2d-convolution, Batch Normalization, and ReLU [18]. In addition, there are skip connections between encoder blocks and corresponding decoder blocks, and the skip-connected features from encoder blocks are concatenated with the corresponding features of decoder blocks. In the middle of the U-Net structure, since we use lip-motion and face information [13, 19] to separate a target speaker, the lip motion features and face image features are concatenated with the output of the encoder.

Using one of the state of the art lip reading model [20], the lip motions are extracted from face detected videos. [20] uses 3D convolutions to get the region of lip motions. The lip motions are obtained before training. ShuffleNet v2 [21] and temporal convolutional network (TCN) [22] make the representations of the lip motions. Facial representations are obtained by ResNet-18 [23], and the lip motion & face features and encoder outputs are concatenated in the middle of the U-Net structure.

2.2. Proposed Background Sound Control Network

We propose a background sound feature extractor to control the magnitude of background sounds. The background sound feature extractor has a same structure with the separation encoder to analyze the background sound feature channels for each encoder block. Some channels of features of encoder blocks have a role in preserving background sounds and some channels have roles related to voice components. The background sound feature extractor can find whether corresponding channels are related to background sounds. The entire architecture of the proposed model is described in Figure 1.

In addition, we design a control activation function $f(x)$ shown in Equation 1.

$$f(x) = \frac{4e^{-x}}{(1+e^{-x})^2} = 4(1-\sigma(x))\sigma(x) \qquad (1)$$
$$\textbf{where} \quad \sigma(x) = \frac{1}{1+e^{-x}}$$

The properties of this function are as follows. First, the range of the function value is in $(0, 1]$. Second, the domain of this function is in the real line. Third, the maximum value is 1 at $x = 0$. Lastly, this function is symmetric about the y-axis. Let an output of k-th encoder block be $\mathbf{h}_k$, and an output of k-th block of background sound feature extractor be $\mathbf{h}^c_k$. Given a

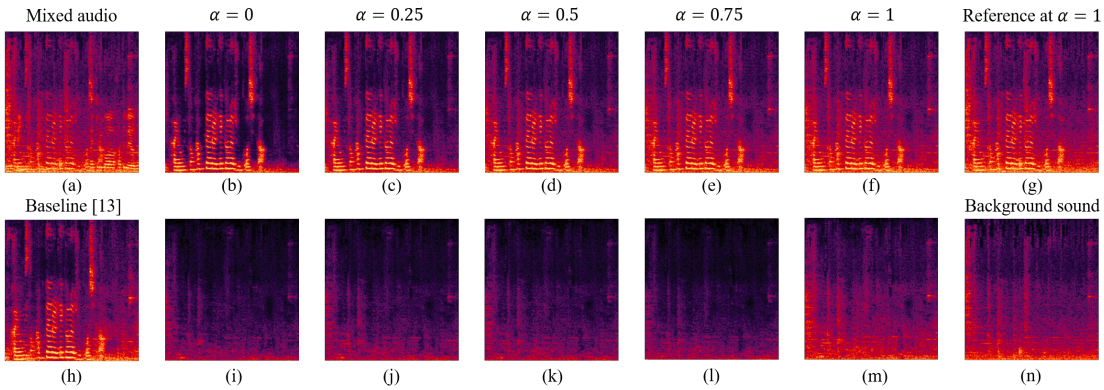

Figure 3: *Spectrogram visualizations of an example on Korean Utterance Dataset. (a) is a mixed audio that consists of voices and a background sound, (b)-(f) are voice separation results according to α, and (g) is a ground truth at $\alpha = 1$. (h) is a separation result on baseline [13] model. (i) is a difference between separated audios at $\alpha = 0.25$ and $\alpha = 0$, (j) is a difference between separated audios at $\alpha = 0.5$ and $\alpha = 0.25$, (k) is a difference between separated audios at $\alpha = 0.75$ and $\alpha = 0.5$, (l) is a difference between separated audios at $\alpha = 1$ and $\alpha = 0.75$, and (m) is a difference between separated audios at $\alpha = 1$ and $\alpha = 0$. (n) is a ground truth background sound mixed in a mixed audio.*

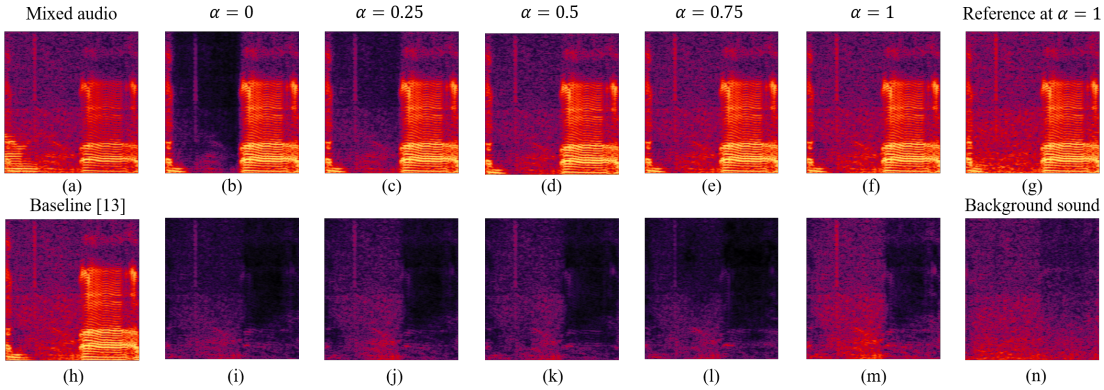

Figure 4: *Spectrogram visualizations of an example on VoxCeleb2. All explanations are same as Figure 3 except that we compute the ground truth background sound (n) by subtracting two voice outputs of speech enhancement model [15] from a mixed audio.*

control parameter $\alpha \in [0, 1]$, the skip-connected feature $\mathbf{h}'_k$ can be described as follows (background sound controller):

$$\mathbf{h}'_k = f(\mathbf{h}^c_k \times (1 - \alpha)) \times \mathbf{h}_k. \quad (2)$$

The visualization of this computation is shown in Figure 2. If $\alpha = 1$, $f(\mathbf{h}^c_k \times (1 - \alpha))$ is always 1, thus $\mathbf{h}_k$ is same as $\mathbf{h}'_k$. Therefore, background sounds are preserved in the voice separation output. If $\alpha = 0$, $f(\mathbf{h}^c_k \times (1 - \alpha))$ depends on $|\mathbf{h}^c_k|$. If $|\mathbf{h}^c_k|$ is large, $f(\mathbf{h}^c_k \times (1-\alpha))$ will be decreased to near zero and $\mathbf{h}'_k$ will be decreased to zero. It means that $\mathbf{h}_k$ is a feature of background sounds and if α decreases, the background sounds are reduced. On the other hand, if $|\mathbf{h}^c_k|$ is small, $|\mathbf{h}^c_k \times (1 - \alpha)|$ is also small even if α decreases. Thus, $f(\mathbf{h}^c_k \times (1-\alpha))$ is near 1 and $\mathbf{h}_k$ is sent to the decoder without reduction. It means that $\mathbf{h}_k$ is not the component of background sounds. With this structure, we can adaptively control the background sound features according to the control parameter α.

2.3. Training Strategy and Implementation Details

We learn the model on $\alpha \in \{0, 0.5, 1\}$, and the model is able to work on $\alpha \in [0, 1]$. When $\alpha = 1$, the target is a voice that includes background sounds, and when $\alpha = 0$, the target voice is a clean voice without background sounds. When $\alpha = 0.5$, we mix a clean voice and the same voice including background sounds with same ratio, and use this as a target. Additionally, we use 16kHz sampling rate, and in STFT, hop length is 160, window length is 400, and size of Fourier transform is 512. The length of a voice segment is 2.55 second. The proposed model predicts the STFT complex masks and the loss function for the mask is L2 distance. Additionally, we use audio-face feature matching loss [13] and speaker verification loss [24, 25]. In audio-face feature matching loss, audio and face features from a same speaker need to be similar in cosine similarity. In speaker verification loss, audio features from a same speaker need to be similar, but voice features from different speakers should be different in cosine similarity. We use ResNet-18 to extract audio features of separated audios.

3. Experiment

3.1. Dataset

- **Korean Utterance Dataset (KUD)** [26] consists of 30 speakers and has about 10 hours recorded videos. There are

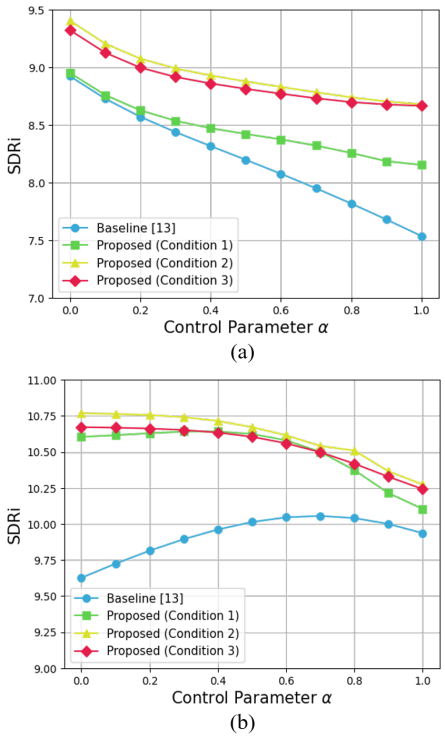

Figure 5: *SDRi results on Korean Utterance Dataset (a) and VoxCeleb2 (b). The reference audio for each α is $\mathbf{v}_i + \alpha\mathbf{b}$ when computing SDRi. We use different number of blocks of background sound feature extractor (2, 5, 8) for each condition (1, 2, 3).*

23 speakers on training data, 7 speakers on test data. The test dataset is unseen on training step. Speakers read Korean scripts randomly in a fixed room. Since the videos are recorded in a controlled environment, the recorded voices have less noises and are clean relatively. Therefore, when $\alpha = 1$, we add up background sounds to raw voices. When $\alpha = 0$, we use the dataset itself as targets.

- **VoxCeleb2** dataset [27] has 150,480 face recognition videos, 1,128,246 utterances, 2,442 hours, and over 6,000 celebrities. There are 5,994 speakers on training data, 59 speakers on validation data, and 59 speakers on test data. VoxCeleb2 dataset contains various background sounds. Therefore, when $\alpha = 1$, we use VoxCeleb2 dataset itself as targets, and when $\alpha = 0$, we use speech enhancement model [15] to generate clean voices.
- **Background Sound Dataset** we collected using a smartphone has 13 categories (traffic-street, café, office, etc.) about 17 hours. We use this dataset to make target labels on Korean Utterance Dataset when $\alpha = 1$.

3.2. Spectrogram Visualization

Figure 3 and 4 show separation results on Korean Utterance Dataset and VoxCeleb2 respectively. Two speakers are randomly mixed with same ratio, and in the case of Korean Utterance Dataset, each voice is mixed with a background sound. In (b)-(f) from Figure 3 and 4, we can observe significant changes when the control parameter is changed from 0 to 1. The spectrogram of $\alpha = 1$ contains most of the background sounds, and the background sounds are reduced when the control parameter goes to zero. To show the increase of background sounds from 0 to 1, we compute and visualize the differences between separated audios at $\alpha = 0.25 * (i + 1)$ and $\alpha = 0.25 * i$ for $i \in \{0, 1, 2, 3\}$ in (i)-(l). (m) results in both figures are predicted background sounds, and (m) and ground truth background sound (n) are close on both datasets. These experimental results show that we can design continuous neural network functions on control parameter α that controls the amount of specific audio elements like background sounds without performance degradation.

	KUD $\alpha = 0$	VoxCeleb2 $\alpha = 0.7$
VisualVoice [13]	8.92 ± 2.68	10.06 ± 2.48
Proposed (Condition 1)	8.95 ± 2.94	10.50 ± 2.38
Proposed (Condition 2)	9.40 ± 2.40	10.54 ± 2.41
Proposed (Condition 3)	9.32 ± 2.24	10.50 ± 2.25

Table 1: *SDRi evaluations for α where SDRi value of the baseline model is maximized for each dataset in Figure 5.*

3.3. Quantitative Experiment

Figure 5 describes the quantitative results on Korean Utterance Dataset (a) and VoxCeleb2 (b). A mixed audio consists of two speakers and a background sound. The x-axis is the control parameter α and the y-axis is Signal to Distortion Ratio improvement (SDRi) computed by the Python library [28]. For each α, we use different references, $\mathbf{v}_i + \alpha\mathbf{b}$, to compute the SDRi scores. Since we know $\mathbf{v}_i + \mathbf{b}$ and $\mathbf{v}_i$, we mix the two audio sources $\mathbf{v}_i(1-\alpha)+(\mathbf{v}_i+\mathbf{b})\alpha = \mathbf{v}_i+\alpha\mathbf{b}$, and use these sources as the references when computing SDRi. Figure 5 shows that the proposed model has controllability of background sounds on both datasets since the separation outputs of the proposed model are more similar to the ground truths than the baseline model predictions for all α. In Table 1, we compute SDRi for α where SDRi value of the baseline model is maximized for each dataset in Figure 5. The α values that we use in Table 1 are different on two datasets since the amount of background sounds in baseline outputs is unpredictable. The results indicate that the proposed model learns the representations of background sounds without loss of performances on voice separation tasks.

4. Conclusion and Future Work

In this paper, we propose an extended AI framework, background-sound controllable voice source separation that can control the amount of background sounds and separate a target voice simultaneously. Without the loss of the separation performances, the model can extract general features of background sounds and does not require additional audio mixing procedures to control background sounds. We show the controllability of background sounds qualitatively and quantitatively on various real world datasets. As a future work, we think it is meaningful to analyze the relationship between the control parameter and loudness (LKFS [29]) of the background sound.

5. Acknowledgement

We specially thank to Master Han-gil Moon in Mobile Experience Business Division, Samsung Electronics for supporting collaboration in this research.

6. References

[1] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[2] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[3] J. Chen, Q. Mao, and D. Liu, "Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation," in *Proc. Interspeech 2020*, 2020, pp. 2642–2646.

[4] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[6] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[7] 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*, vol. 25, no. 10, pp. 1901–1913, 2017.

[8] C. Xu, W. Rao, X. Xiao, E. S. Chng, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid lstm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6–10.

[9] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Proc. Interspeech 2019*, 2019, pp. 2728–2732.

[10] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černockỳ, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[11] X. Xiao, Z. Chen, T. Yoshioka, H. Erdogan, C. Liu, D. Dimitriadis, J. Droppo, and Y. Gong, "Single-channel speech extraction using speaker inventory and attention network," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 86–90.

[12] A. Rahimi, T. Afouras, and A. Zisserman, "Reading to listen at the cocktail party: Multi-modal speech separation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 493–10 502.

[13] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 15 490–15 500.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[15] A. Défossez, G. Synnaeve, and Y. Adi, "Real Time Speech Enhancement in the Waveform Domain," in *Proc. Interspeech 2020*, 2020, pp. 3291–3295.

[16] S. Welker, J. Richter, and T. Gerkmann, "Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain," in *Proc. Interspeech 2022*, 2022, pp. 2928–2932.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[18] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[19] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 530–541, 2020.

[20] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6319–6323.

[21] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[22] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 47–54.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[25] C. Zhang, K. Koishida, and J. H. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1633–1644, 2018.

[26] W. H. Nam, K.-R. Kim, J. Kim, and D. Eom, "Audio-visual-information-based speaker matching framework for selective hearing in a recorded video," in *Audio Engineering Society Convention 153*. Audio Engineering Society, 2022.

[27] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[28] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, "Mir_eval: A transparent implementation of common mir metrics." in *ISMIR*, 2014, pp. 367–372.

[29] ITU-R BS.1770-4, "Algorithms to measure audio programme loudness and true-peak audio level." International Telecommunications Union, Geneva, 2015.