Interspeech 2022
18-22 September 2022, Incheon, Korea

A Hybrid Continuity Loss to Reduce Over-Suppression for Time-domain Target Speaker Extraction

Zexu Pan[1], *Meng Ge*[2], *Haizhou Li*[3,1,4,5]

[1]Institute of Data Science, National University of Singapore, Singapore
[2]Key Laboratory of Cognitive Computing and Application, Tianjin University, China
[3]The Chinese University of Hong Kong, Shenzhen, China
[4]Machine Listening Lab, University of Bremen, Germany, [5]Kriston AI, China

pan_zexu@u.nus.edu, gemeng@tju.edu.cn, haizhou.li@nus.edu.sg

Abstract

The speaker extraction algorithm extracts the target speech from a mixture speech containing interference speech and background noise. The extraction process sometimes over-suppresses the extracted target speech, which not only creates artifacts during listening but also harms the performance of downstream automatic speech recognition algorithms. We propose a hybrid continuity loss function for time-domain speaker extraction algorithms to settle the over-suppression problem. On top of the waveform-level loss used for superior signal quality, i.e., SI-SDR, we introduce a multi-resolution delta spectrum loss in the frequency-domain, to ensure the continuity of an extracted speech signal, thus alleviating the over-suppression. We examine the hybrid continuity loss function using a time-domain audio-visual speaker extraction algorithm on the YouTube LRS2-BBC dataset. Experimental results show that the proposed loss function reduces the over-suppression and improves the word error rate of speech recognition on both clean and noisy two-speakers mixtures, without harming the reconstructed speech quality.

Index Terms: speaker extraction, automatic speech recognition, hybrid continuity loss function, over-suppression, time-domain, audio-visual

1. Introduction

The acoustic environment during real-world human-robot interaction can be described as a cocktail party [1], where the speech from a speaker of interest, i.e., the target speaker, is often corrupted by interference speakers and background noise. In such a scenario, speech separation or speaker extraction algorithms are usually needed to extract the clean speech signal of the target speaker [2, 3], which is a crucial step for downstream applications such as hearing aid development [4], active speaker detection [5], speech recognition [6, 7], and source localization [8].

The speech separation algorithm [9–14] separates a mixture speech signal into individual streams, each containing a clean speech signal of a speaker. Inside the network, multiple speaker streams compete and segregate, which usually requires the number of speakers to be known.

The speaker extraction algorithm takes a different approach, which only extracts the speech of the target speaker from the mixture. It does not need to know the number of speakers but requires an auxiliary reference to distinguish the target speaker from the others. The auxiliary reference can be a pre-recorded reference speech signal, in which the algorithm extracts a speech signal that has a similar voice signature to the reference speech signal [15–20]. The video recording of the target speaker also serves as such a reference, in which the algorithm extracts a speech signal that is temporally synchronized with the speaker's motion in the video [21–28].

Time-domain speaker extraction algorithm is preferable to the frequency-domain counterparts because it gets rid of the phase reconstruction problem in Short Time Fourier Transform (STFT) [10, 17]. However, it experiences the over-suppress problem in extracting the target signal, in which the time-frequency bins of the extracted speech magnitude spectrogram are attenuated to a lower value than expected [15]. This not only creates auditory artifacts, but also adversely affects downstream applications such as automatic speech recognition (ASR).

Addressing the over-suppression problem, VoiceFilter-Lite [15] took the frequency-domain approach with an asymmetric loss function, which penalizes the over-suppression errors more than the under-suppression errors on the magnitude spectrogram. The asymmetric loss function is designed to improve the ASR performance, but it remains a question whether this will adversely impact the speech reconstruction. In this paper, we seek to alleviate the over-suppression problem in time-domain speaker extraction algorithms, so as to improve the downstream ASR performance, at the same time, maintaining speech reconstruction quality.

In speech synthesis [29, 30], the multi-resolution STFT loss function has been used to construct a speech signal. Such loss function has also been adopted in speech enhancement [31] to remove noise signals. In addition, delta features have been widely used in speech processing algorithms, e.g., The Mel Frequency Cepstral Coefficients. In a frequency-domain speech separation study named cuPIT-Grid LSTM [32], the delta features have been introduced in the loss function to ensure the continuity of the speech.

Motivated by the above studies, we introduce a hybrid continuity loss function to train a time-domain system, that consists of a time-domain loss and a frequency-domain loss. We use the widely adopted scale-invariant signal-to-noise ratio (SI-SDR) [33] as the time-domain loss, which works well for time-domain systems in reconstructing raw a speech waveform.

This research is supported by The Science and Engineering Research Council, Agency for Science, Technology and Research (A*STAR), Singapore, through the National Robotics Program under Human-Robot Interaction Phase 1 (Grant No. 192 25 00054), by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen), and the Guangdong Provincial Key Laboratory of Big Data Computing under the Grant No. B10120210117-KP02, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen).

10.21437/Interspeech.2022-157

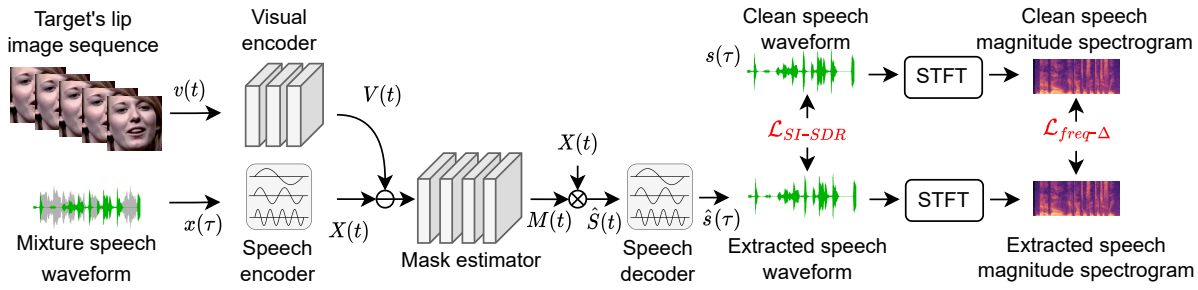

Figure 1: *The pipeline of a mask-based time-domain audio-visual speaker extraction network. $\ominus$ represents the concatenation of features along the channel dimension and $\otimes$ represents element-wise multiplication. We proposed a hybrid continuity loss function that consists of the time-domain loss $\mathcal{L}_{SI\text{-}SDR}$ and frequency-domain delta spectrum loss $\mathcal{L}_{freq\text{-}\Delta}$ at different resolutions, to alleviate the over-suppression problem.*

We also propose a multi-resolution delta spectrum loss in the frequency-domain, to ensure the temporal continuity of a spectrogram. The over-suppressed time-frequency bins or frames will be penalized based on the value of the preceding and following frames, thus alleviating the over-suppression. We examine the loss function on a state-of-the-art time-domain audio-visual speaker extraction algorithm [28], experimental results show that the hybrid continuity loss function outperforms the use of a single SI-SDR loss function, in improving the downstream ASR performance, while achieving comparable or better signal quality for clean and noisy 2-speaker mixtures simulated from the YouTube LRS2-BBC dataset [34, 35].

The rest of the paper is organized as follows. In Section 2, we introduce the speaker extraction framework and the proposed hybrid continuity loss. In Section 3, we present the experimental setup. In Section 4, we discuss the experimental results. Finally, we conclude the study in Section 5.

2. Proposed method

Let $x(\tau)$ be a mixture speech in the time-domain, consisting of the target speech $s(\tau)$, interference speech $b_i(\tau)$, and noise $n(\tau)$:

$$x(\tau) = s(\tau) + \sum_{i=1}^{I} b_i(\tau) + n(\tau) \tag{1}$$

where $i \in \{1, ..., I\}$ denotes the index of interference speakers. We would like to estimate $\hat{s}(\tau)$ to approximate $s(\tau)$.

2.1. Neural architecture

We use an audio-visual time-domain speaker extraction network, that has a similar architecture to the USEV network [28]. It has a speech encoder, a visual encoder, a mask estimator, and a speech decoder, which is illustrated in Fig. 1.

The speech encoder transforms the time-domain mixture speech $x(\tau)$ into a frame-based embedding $X(t)$ in the latent space [1] through a convolutional layer $Conv1D(\cdot)$, which behaves like a frequency analyzer [10]:

$$X(t) = ReLU(Conv1D(x(\tau), C, L, L/2)) \tag{2}$$

where $C = 256$, $L = 40$, and $L/2$ are the channel, kernel, and stride sizes respectively, $ReLU(\cdot)$ is rectified linear activation.

[1] In this paper, a variable with subscript τ represents it is a time-domain waveform, while a variable with subscript t represents it is a frame-based embedding.

The visual encoder encodes the target's lip image sequence $v(t)$ into viseme embeddings $V(t)$, which are temporally synchronized with the target's speech in $X(t)$. The $V(t)$ serves as the speaker cue, i.e., lip motion, to attract the target speaker in the following processes. It consists of a ResNet18 [22], and 5 layers of temporal convolutional neural network for adaptation [3, 21, 36]. The ResNet18 is pre-trained on the visual speech recognition task [35], and its weights are fixed during the speaker extraction training, to retain the pre-trained knowledge in extracting the lip motion. The visual embeddings $V(t)$ coming out from the visual encoder are up-sampled along the time dimension to match the temporal resolution of the mixture speech embeddings $X(t)$.

The mask estimator takes in the concatenated viseme embeddings $V(t)$ and mixture speech embeddings $X(t)$ as input, to estimate a mask $M(t)$ for the target speech in $X(t)$. The mask estimator has $R = 6$ repeated dual-path long short-term memory structures to capture the global dependencies of a speech signal [9]. The mask is element-wise multiplied with $X(t)$ to obtain the extracted speech embeddings $\hat{S}(t)$ in the latent space:

$$\hat{S}(t) = M(t) \otimes X(t) \tag{3}$$

The decoder renders the extracted speech embeddings $\hat{S}(t)$ back to the time-domain speech waveform $\hat{s}(\tau)$. It consists of a linear layer $Linear(\cdot)$ with an overlap-and-add operation $OnA(\cdot)$ [37].

$$\hat{s}(\tau) = OnA(Linear(\hat{S}(t), C, L), L/2) \tag{4}$$

2.2. Hybrid loss

In time-domain end-to-end speaker extraction network training, the negative SI-SDR loss function [33] has been widely used, which is formulated as follow:

$$\mathcal{L}_{SI\text{-}SDR} = -20 \log_{10} \frac{||\frac{<\hat{s},s>s}{||s||^2}||}{||\hat{s} - \frac{<\hat{s},s>s}{||s||^2}||} \tag{5}$$

we omit the subscript τ in the equations for brevity.

We propose a hybrid continuity loss function that consists of the time-domain SI-SDR loss, with a frequency-domain multi-resolution delta spectrum loss as follows:

$$\mathcal{L}_{hybrid} = \mathcal{L}_{SI\text{-}SDR} + \gamma \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}^{m}_{freq\text{-}\Delta} \tag{6}$$

where $M = 3$ is the number of delta spectrum loss $\mathcal{L}_{freq\text{-}\Delta}$ at different resolutions, with number of FFT bins $\in \{512, 1024,$

2048}, hop sizes ∈ {50, 120, 240}, and window lengths ∈ {240, 600, 1200} respectively. γ is a balancing weight that is set to 1 in this paper.

Each delta spectrum loss consists of a spectrum convergence loss and magnitude loss with delta features, i.e., differential and acceleration features, that is formulated as follow:

$$\mathcal{L}_{freq\text{-}\Delta} = \mathcal{L}_{sc\text{-}\Delta} + \mathcal{L}_{mag\text{-}\Delta} \tag{7}$$

$$\begin{aligned}\mathcal{L}_{sc\text{-}\Delta} = &\frac{||\ |STFT(s)| - |STFT(\hat{s})|\ ||_F}{||\ |STFT(s)|\ ||_F} \\ &+ \frac{||\ f_D(|STFT(s)|) - f_D(|STFT(\hat{s})|)\ ||_F}{||\ f_D(|STFT(s)|)\ ||_F} \\ &+ \frac{||\ f_A(|STFT(s)|) - f_A(|STFT(\hat{s})|)\ ||_F}{||\ f_A(|STFT(s)|)\ ||_F}\end{aligned} \tag{8}$$

$$\begin{aligned}\mathcal{L}_{mag\text{-}\Delta} = &\frac{1}{N}||\ log|STFT(s)| - log|STFT(\hat{s})|\ ||_1 \\ &+ \frac{1}{N}||\ f_D(log|STFT(s)|) - f_D(log|STFT(\hat{s})|)\ ||_1 \\ &+ \frac{1}{N}||\ f_A(log|STFT(s)|) - f_A(log|STFT(\hat{s})|)\ ||_1\end{aligned} \tag{9}$$

where $STFT(\cdot)$ represents the operation to obtain the magnitude spectrogram of a signal, N denotes the total number of time-frequency bins, $||\cdot||_F$ and $||\cdot||_1$ are the Frobenius and the L_1 norms respectively. The function $f_D(\cdot)$ to compute the differential feature is as follow:

$$f_D(v(t)) = \frac{\sum_{l=1}^{L} l \times (v(t+l) - v(t-l))}{\sum_{l=1}^{L} 2l^2} \tag{10}$$

where L is the order and sets to 2 in this paper. The function $f_A(\cdot)$ to compute the acceleration feature is the computation of $f_D(\cdot)$ twice.

With multiple resolutions and the delta features for the frequency-domain loss, we aim to enhance the continuity of the extracted speech magnitude spectrogram with a greater temporal receptive field, and ensure that the over-suppressed frames are seen and compensated by the preceding or following frames, thus reducing over-suppression. In the meantime, the extracted speech is also constrained by the SI-SDR loss, which directly optimizes the extracted signal quality. We expect the hybrid continuity loss function to perform well for both signal quality evaluations and ASR evaluations.

3. Experimental setup

3.1. Dataset

We simulate a 2-speaker mixture dataset named LRS2-mix dataset, from the Oxford-BBC Lip Reading Sentences 2 (LRS2) dataset [34, 35] to evaluate the speaker extraction and downstream ASR performances. The LRS2 dataset is an audio-visual dataset, that consists of 45,839, 1,082, and 1,243 clean speech utterances in the original train, validation, and test sets respectively. The speech signal is available at 16 kHz. The face track video of each speech utterance is provided and is available at 25 frames per second.

To simulate the LRS2-mix dataset, we use the original train, validation, and test sets of the LRS2 dataset to simulate 200,000, 5,000, and 3,000 mixture speech utterances in our train, validation, and test sets respectively. The interference speech is truncated or zero-padded to the length of the target speech first and mixed with the target speech at a random Signal-to-Noise ratio (SNR) set between 10dB and -10dB.

We also simulate a noisy version of the LRS2-mix dataset using babble noise from the WHAM! noise dataset [38]. For each mixture speech utterance of the LRS2-mix dataset in the train, validation, and test set, a noise signal is randomly sampled from the WHAM! train, validation, and test set respectively and added at a random SNR set between 15 dB to -5 dB. The utterances in the test set do not overlap with those in the train and validation sets, which allows us to perform speaker-independent evaluations.

3.2. Training

We implement the model with PyTorch[2]. We use the Adam optimizer with an initial learning rate of 0.001. The learning rate is halved if the best validation loss (BVL) does not improve for 6 consecutive epochs, and the training stops when the BVL does not improve for 10 consecutive epochs. During training, the utterances are truncated to 6 seconds to fit into the GPU memory, while the full utterance is used for evaluation.

3.3. Evaluation metric

3.3.1. Signal quality and intelligibility

We use the SI-SDR and signal-to-noise ratio (SDR) to measure the extracted speech quality as compared with the mixture speech. We use the Perceptual Evaluation of Speech Quality (PESQ) and Short Term Objective Intelligibility (STOI) to measure the perceptual quality and intelligibility of the extracted speech as compared with the mixture speech. The higher the better for the four metrics.

3.3.2. Automatic speech recognition

We connect the audio-only or audio-visual *Transformer Connectionist Temporal Classification* TM-CTC model [35] at the end of the speaker extraction model, to transcribe the extracted speech utterances. The TM-CTC models are pre-trained on the original LRS2 dataset [34, 35] that consists of clean speech only [3]. We use the Word Error Rate (WER) and Character Error Rate (CER) to evaluate the downstream ASR performance. The lower the better for the two metrics.

3.3.3. Over-suppression

We use the mean absolute error (MAE) to measure the difference of the magnitude spectrogram between the extracted speech utterance $\hat{s}(\tau)$ and the clean speech utterance $s(\tau)$. The over-suppression MAE and under-suppression MAE for an extracted speech utterance are:

$$MAE_{over} = \frac{1}{N}||\ ReLU(|STFT(s)| - |STFT(\hat{s})|)\ ||_1 \tag{11}$$

$$MAE_{under} = \frac{1}{N}||\ ReLU(|STFT(\hat{s})| - |STFT(s)|)\ ||_1 \tag{12}$$

where the STFT operation here has FFT bins, hop size, and window length of 512, 120, and 600 respectively. The $ReLU$ operation sets the time-frequency bins to zero if the latter has a negative value.

[2]The data generation and the training scripts are available at `https://github.com/zexupan/avse_hybrid_loss`.

[3]The ASR pre-trained models are taken from `https://github.com/lordmartian/deep_avsr`

Table 1: *We report the performance of the time-domain audio-visual speaker extraction network using SI-SDR loss function and our proposed hybrid continuity loss function on the clean and noisy LRS2-mix dataset.*

Noise condition	Model	SI-SDR	SDR	PESQ	STOI	Audio-visual ASR		Audio-only ASR		MAE	
						WER	CER	WER	CER	over	under
Clean	Mixture speech	0.79	1.10	1.985	0.708	59.2	36.5	80.9	54.7	-	-
	SI-SDR loss [27]	13.41	14.08	2.997	0.931	17.3	7.4	23.5	11.3	0.105	0.055
	Hybrid loss (Ours)	13.48	14.12	3.163	0.934	16.1	6.8	21.2	10.0	0.096	0.060
	Clean speech	88.12	288.88	4.500	1.000	11.4	4.4	12.5	4.9	-	-
Noisy	Mixture speech	-4.43	-3.81	1.467	0.559	56.5	32.7	89.5	60.6	-	-
	SI-SDR loss [27]	4.54	5.91	2.179	0.776	38.8	20.7	57.9	34.6	0.241	0.118
	Hybrid loss (Ours)	5.12	6.52	2.276	0.789	34.5	17.7	50.8	29.5	0.223	0.119
	Clean speech	88.12	288.88	4.500	1.000	11.4	4.4	12.5	4.9	-	-

Table 2: *Study on the multi-resolution delta spectrum loss.* Δ *here means the use of the delta features, i.e., differential and acceleration features. The systems (Sys) here are trained and evaluated on the clean LRS2-mix dataset.*

Sys	STFT loss	FFT size	Δ	SI-SDR	WER
1	$\mathcal{L}_{sc} + \mathcal{L}_{mag}$	{512,1024,2048}	✓	13.48	**16.1**
2	$\mathcal{L}_{sc}$			13.51	17.0
3	$\mathcal{L}_{mag}$			13.28	16.5
4	$\mathcal{L}_{sc} + \mathcal{L}_{mag}$	{512,1024,2048}	✗	13.46	16.3
5	$\mathcal{L}_{sc}$			13.42	17.1
6	$\mathcal{L}_{mag}$			13.44	16.6
7	$\mathcal{L}_{sc} + \mathcal{L}_{mag}$	512	✓	13.25	16.5
8		1024		13.42	17.1
9		2048		13.26	16.3

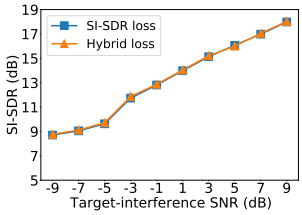

Figure 2: *The SI-SDR against the target-interference SNR.*

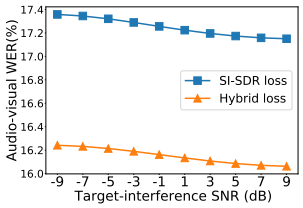

Figure 3: *The WER against the target-interference SNR.*

4. Results

4.1. Comparison with baseline

In Table 1, we compare the hybrid continuity loss function with the SI-SDR loss function for the time-domain audio-visual speaker extraction network that is shown in Fig. 1. We also present the results when the mixture speech and clean speech are used to calculate the metrics.

When the systems are trained and evaluated on the clean LRS2-mix dataset, the hybrid loss function performs similarly to SI-SDR loss for SI-SDR, SDR, and STOI evaluations, but outperforms the latter on PESQ evaluation. The hybrid loss has lower WER and CER for both audio-visual and audio-only ASR evaluations, credited to the 8.6% reduction of over-suppression MAE. When the systems are trained and evaluated on the noisy LRS2-mix dataset, the hybrid loss outperforms the SI-SDR loss for both the signal quality evaluations, i.e., SI-SDR, SDR, PESQ, STOI, and ASR evaluations, i.e., WER and CER. For both clean and noisy evaluations, the MAE for over-suppression decreases, but that for under-suppression increases, which further shows that the ASR algorithm tolerates more under- than over-suppression errors.

4.2. Ablation studies

Next, we study the contribution of individual components from the multi-resolution delta spectrum loss and present the results in Table 2. System 1 is trained with our proposed hybrid continuity loss function, while systems 2-9 are trained with the SI-SDR loss together with part of the multi-resolution delta spectrum loss. We report the SI-SDR and the audio-visual WER here. All systems achieve similar SI-SDR, which is similar to using the SI-SDR loss function alone, i.e., 13.41 in Table 1, this shows that the use of the frequency-domain loss does not harm the signal reconstruction from 2-speaker mixture speech.

System 2 only has the spectrum convergence loss in the hybrid continuity loss function, while system 3 only has that of the magnitude loss. Both systems are outperformed by system 1 on WER evaluations, which used our proposed loss. Both the spectrum convergence loss and the magnitude loss contribute to the reduction of WER. System 4 does not include the delta features, i.e., differential and acceleration features, the WER increases by 0.2 % compared with system 1, which shows that the delta features are useful. Systems 7-9 only use single-resolution for STFT in the hybrid continuity loss function, none of them outperform system 1 in terms of WER, which uses that of multi-resolution loss, which shows the importance of various resolutions.

We also present the SI-SDR and the audio-visual WER evaluations of the baseline and the proposed system on the clean LRS2-mix dataset with various target-interference SNR in Figures 2 and 3. The target-interference SNR is defined as the energy contrast between the target speaker and the interference speaker in the mixture speech in terms of SNR. As the input mixture becomes less noisy, i.e. higher SNR, the SI-SDR increases and the WER decreases. We observe that the improvement in WER is less affected by the target-interference SNR than the SI-SDR. The hybrid loss outperforms the SI-SDR loss in terms of WER evaluation and achieves a similar result on the SI-SDR evaluation.

5. Conclusion

In this work, we proposed a hybrid continuity loss function that consists of an SI-SDR loss and a multi-resolution delta spectrum loss, to alleviate the over-suppression errors in time-domain speaker extraction networks. This makes a step forward for downstream ASR algorithms.

6. References

[1] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[2] 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 Trans. Graph.*, vol. 37, no. 4, pp. 1–11, 2018.

[3] Z. Pan, R. Tao, C. Xu, and H. Li, "Selective listening by synchronizing speech with lips," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1650–1664, 2022.

[4] D. Wang, "Deep learning reinvents the hearing aid," *IEEE spectrum*, vol. 54, no. 3, pp. 32–37, 2017.

[5] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, "Is someone speaking? Exploring long-term temporal features for audio-visual active speaker detection," in *Proc. of the 29th ACM Int. Conf. on Multimedia*, 2021, pp. 3927–3935.

[6] X. Yue, G. Lee, E. Yılmaz, F. Deng, and H. Li, "End-to-end code-switching ASR for low-resourced language pairs," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 972–979.

[7] Q. Song, B. Sun, and S. Li, "Multimodal sparse transformer network for audio-visual speech recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[8] X. Qian, M. Madhavi, Z. Pan, J. Wang, and H. Li, "Multi-target DoA estimation with an audio-visual fusion mechanism," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 4280–4284.

[9] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 46–50.

[10] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[11] 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 Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[12] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 31–35.

[13] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *preprint arXiv:2002.08933*, 2020.

[14] Y. Liu and D. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2092–2102, 2019.

[15] Q. Wang, I. L. Moreno, M. Saglam, K. Wilson, A. Chiao, R. Liu, Y. He, W. Li, J. Pelecanos, M. Nika *et al.*, "VoiceFilter-Lite: Streaming targeted voice separation for on-device speech recognition," *Proc. INTERSPEECH*, pp. 2677–2681, 2020.

[16] C. Xu, W. Rao, E. S. Chng, and H. Li, "SpEx: Multi-scale time domain speaker extraction network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1370–1384, 2020.

[17] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A complete time domain speaker extraction network," in *Proc. INTERSPEECH*, 2020, pp. 1406–1410.

[18] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 86–90.

[19] J. Shi, J. Xu, Y. Fujita, S. Watanabe, and B. Xu, "Speaker-conditional chain model for speech separation and extraction," in *Proc. INTERSPEECH*, 2020, pp. 2707–2711.

[20] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, "Improving speaker discrimination of target speech extraction with time-domain SpeakerBeam," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 691–695.

[21] J. Wu, Y. Xu, S. Zhang, L. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 667–673.

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

[23] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, "Multimodal SpeakerBeam: Single channel target speech extraction with audio-visual speaker clues," in *Proc. INTERSPEECH*, 2019, pp. 2718–2722.

[24] Z. Pan, X. Qian, and H. Li, "Speaker extraction with co-speech gestures cue," *IEEE Signal Processing Letters*, 2022.

[25] K. Ito, M. Yamamoto, and K. Nagamatsu, "Audio-visual speech enhancement method conditioned in the lip motion and speaker-discriminative embeddings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6668–6672.

[26] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2021.

[27] Z. Pan, R. Tao, C. Xu, and H. Li, "MuSE: Multi-modal target speaker extraction with visual cues," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6678–6682.

[28] Z. Pan, M. Ge, and H. Li, "USEV: Universal speaker extraction with visual cue," *arXiv preprint arXiv:2109.14831*, 2021.

[29] R. Yamamoto, E. Song, and J.-M. Kim, "Probability density distillation with generative adversarial networks for high-quality parallel waveform generation," *Proc. INTERSPEECH*, pp. 699–703, 2019.

[30] ——, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6199–6203.

[31] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *Proc. INTERSPEECH*, pp. 3291–3295, 2020.

[32] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 6–10.

[33] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 626–630.

[34] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *IEEE conference on computer vision and pattern recognition*, 2017, pp. 3444–3453.

[35] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell*, 2018.

[36] Z. Pan, Z. Luo, J. Yang, and H. Li, "Multi-modal attention for speech emotion recognition," in *Proc. INTERSPEECH*, 2020, pp. 364–368.

[37] A. Oppenheim and R. Schafer, "Theory and application of digital signal processing," *Englewood Cliffs*, 1978.

[38] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "Wham!: Extending speech separation to noisy environments," in *Proc. INTERSPEECH*, 2019.