ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095266

TARGET SOUND EXTRACTION WITH VARIABLE CROSS-MODALITY CLUES

Chenda Li[1,2,†], *Yao Qian*[2], *Zhuo Chen*[2], *Dongmei Wang*[2],
Takuya Yoshioka[2], *Shujie Liu*[2], *Yanmin Qian*[1], *Michael Zeng*[2]

[1]MoE Key Lab of Artificial Intelligence, AI Institute
[1]X-LANCE Lab, Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Microsoft, Redmond, WA, USA

ABSTRACT

Automatic target sound extraction (TSE) is a machine learning approach to mimic the human auditory perception capability of attending to a sound source of interest from a mixture of sources. It often uses a model conditioned on a fixed form of target sound clues, such as a sound class label, which limits the ways in which users can interact with the model to specify the target sounds. To leverage variable number of clues cross modalities available in the inference phase, including a video, a sound event class, and a text caption, we propose a unified transformer-based TSE model architecture, where a multi-clue attention module integrates all the clues across the modalities. Since there is no off-the-shelf benchmark to evaluate our proposed approach, we build a dataset [1] based on public corpora, Audioset and AudioCaps. Experimental results for seen and unseen target-sound evaluation sets show that our proposed TSE model can effectively deal with a varying number of clues which improves the TSE performance and robustness against partially compromised clues.

Index Terms— Target sound extraction, cross-modality attention, multi-clue processing

1. INTRODUCTION

People can focus their auditory attention on the sound of their interest in a complex acoustic environment [1]. Researchers have attempted to endow machines with a similar capability by audio source separation, a process of separating all audio sources out of their mixture. Audio source separation includes speech separation [2–4], music separation [5–7], and universal sound separation [8–10].

In some cases, instead of separating all sound sources, we may only be interested in a specific source in the mixed signal. With target sound extraction (TSE), only the sound of interest is extracted from the audio mixture given a target clue. The clues for the TSE systems can be provided in various forms. Sound-related clues include a sound tag [10, 11], a reference speech signal and a target speaker embedding [12–16]. Visual information can also be used as the extraction clue in multi-modal target sound extraction [17–19]. Some recent works use natural language descriptions as the clues [20, 21]. In [22], the authors use 'concept' as the clue for target speech extraction, where the concept can be extracted from an image or a speech signal that is related to the target speech.

While there are various TSE systems each of which handles a specific clue form, there is still room for improvement in practical applications. First, a single clue may be insufficient to describe a specific target sound. Secondly, the single-clue TSE system is not robust against device failures. For example, the single-clue vision-based TSE system may become incapacitated when there is a camera failure. Lastly, the pieces of information provided by multiple types of clues may be complementary to each other to create a more comprehensive clue about the target sound, which could result in extraction performance improvement. Some recent works took advantage of multiple clues for target speech extraction [23–26] and showed the performance superiority to single-clue systems. However, most of these systems, except for [26], require all the clues to be available during the inference. Also, they were built for speech signals and did not cope with general non-speech sounds.

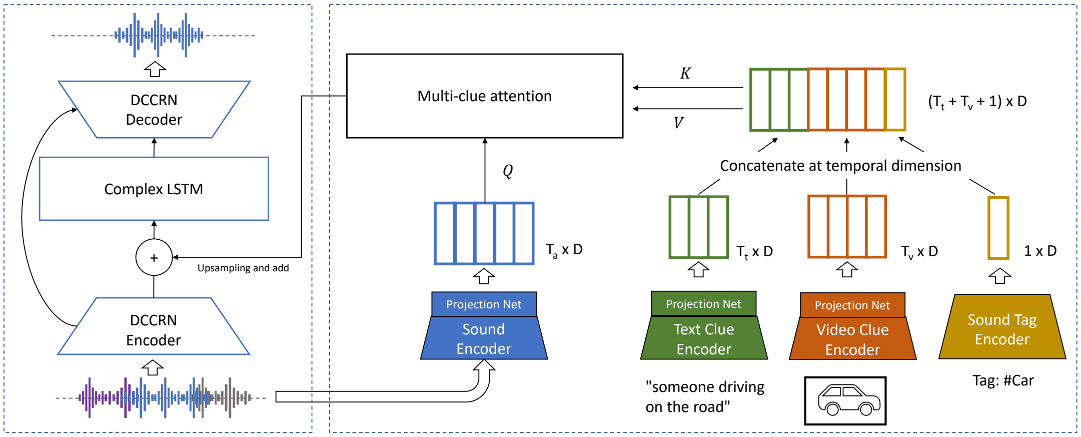
In this paper, we propose a unified TSE system that can extract the target sound by flexibly combining multiple clues from different modalities that are available at test time, including a sound event tag, a text description, and a video clip related to the target sound. Designing such a system gives rise to several challenges. First, clues of different modalities take different input forms, requiring a unified approach to processing them in the embedding space. Secondly, some of the clues (e.g. the sound event tag) provide static information about the sound to be extracted, while others (e.g. the video clip) provide dynamic information which changes over time. It is important to appropriately deal with the alignment between the clues and the input audio features. Thirdly, the multi-clue TSE system should be able to deal with various number of input clues. To solve these challenges, we design a clue processing network based on multi-clue attention. The basic idea is inspired by [26], which propose a unified Transformer-based model for speech extraction with text and video clues. In our TSE system, the observed sound mixture and all the input clues are firstly encoded into a unified embedding space with the corresponding modality encoders. Then, the encoded sound mixture can attend to all the different clues at the same time to synthesize a cross-modality clue. This process does not require all the clues to be present, and the alignment problem is handled with an attention mechanism. The contributions of this paper are as follows. (1) We propose a TSE model based on a multi-clue attention module to leverage a variable number of clues with different modalities. (2) The system robustness and the details of the attention mechanism are experimentally analyzed. (3) We build a multi-modal TSE dataset based on public corpora, Audioset [27] and AudioCaps [28].

2. TARGET SOUND EXTRACTION

The goal of target sound extraction (TSE) is to extract the sound of interest from an audio mixture given a set of one or more target clues. Let $\mathbf{y}$ denote the input audio mixture consisting of J sound sources

[†] The first author conducted this work during internship at Microsoft.

[1] Python scripts to generate this dataset can be found at `https://github.com/LiChenda/Multi-clue-TSE-data`.



(A) DCCRN sound extraction backbone

(B) Clue processing net

Fig. 1: Proposed multi-clue target sound extraction network.

$\mathbf{s}_1, \cdots, \mathbf{s}_J$, and suppose the j-th source to be the target sound. Then, the TSE mapping function can be formulated as

$$f_{TSE}(\mathbf{y}, \mathbf{c}_j) = \hat{\mathbf{s}}_j \rightarrow \mathbf{s}_j, \tag{1}$$

where $\hat{\mathbf{s}}_j$ is the estimated target sound and $\mathbf{c}_j$ is a target sound representation from the provided clue set. One of the simplest forms of the clue set contains only one sound event tag represented as a one-hot vector [10, 11], and it is used as our baseline system. In our proposed multi-clue system, $\mathbf{c}_j$ is produced by a multi-clue processing net as described in Section 3.

The backbone of our TSE system is based on a deep complex convolution recurrent network (DCCRN) [29]. It was originally proposed for speech enhancement using complex time-frequency spectra and also applied to target speech extraction [30]. As Fig. 1 (A) shows, the DCCRN consists of an encoder, an enhancement LSTM, and a decoder. The encoder and decoder consist of complex convolution layers and are connected with U-Net-like skip-connections [31, 32]. The enhancement LSTM between the encoder and the decoder processes the sum of the complex deep encoded features and the clue embedding features as follows:

$$\mathbf{F}_{rr} = \text{LSTM}_r(\mathbf{Y}_r + \mathbf{c}_j), \quad \mathbf{F}_{ir} = \text{LSTM}_r(\mathbf{Y}_i + \mathbf{c}_j), \tag{2}$$

$$\mathbf{F}_{ri} = \text{LSTM}_i(\mathbf{Y}_r + \mathbf{c}_j), \quad \mathbf{F}_{ii} = \text{LSTM}_i(\mathbf{Y}_i + \mathbf{c}_j), \tag{3}$$

$$\mathbf{F}_{out} = (\mathbf{F}_{rr} - \mathbf{F}_{ii}) + (\mathbf{F}_{ri} + \mathbf{F}_{ir})i, \tag{4}$$

where $\mathbf{Y}_r, \mathbf{Y}_i \in \mathbb{R}^{T\times D}$ are the real and imaginary parts of the complex deep features generated by the DCCRN encoder with T and D being the feature sequence length and dimension, respectively. $\mathbf{c}_j \in \mathbb{R}^{T\times D}$ is the encoded clue of the target sound, which is mapped from the one-hot tag vector with a linear layer and tiled to length T. $\text{LSTM}_r, \text{LSTM}_i$ are LSTMs for the real and imaginary features. $\mathbf{F}_{out}$ is the complex features that are fed to the DCCRN decoder. The DCCRN decoder maps $\mathbf{F}_{out}$ into the estimate target sound $\hat{\mathbf{s}}_j$.

The model is trained to minimize the following loss function:

$$\mathcal{L} = \mathcal{L}_{snr} + \lambda \mathcal{L}_{L1}, \tag{5}$$

$$\mathcal{L}_{snr} = -10 \log_{10} \left(\frac{||\mathbf{s}_j||_2^2}{||\mathbf{s}_j - \hat{\mathbf{s}}_j||_2^2} \right), \tag{6}$$

$$\mathcal{L}_{L1} = ||\mathbf{S}_j - \hat{\mathbf{S}}_j||_1, \tag{7}$$

where $\mathbf{S}_j$ and $\hat{\mathbf{S}}_j$ are the complex spectra of the target and estimated sources, respectively, and λ was set at 5 in our experiments.

3. MULTI-CLUE PROCESSING NET

Fig. 1 shows a diagram of the proposed multi-clue TSE model. It consists of two modules: a DCCRN TSE backbone (Fig. 1 (A)) and a multi-clue processing net (Fig. 1 (B)), where the backbone model is the same as the one described in the previous section. The clue processing net takes a variable number of clues as input and generates a time-aligned fused clue, which is fed to DCCRN's LSTM block as $\mathbf{c}_j$ in Eqs. (2) and (3). To deal with clues from different modalities, the clue processing net uses different modal encoders, including a sound encoder, a text clue encoder, a video clue encoder, and a sound tag encoder, which transform the corresponding clue inputs to a unified D-dimensional space. On top of the modal encoders, we use an attention module [33] for clue fusion.

Sound encoder: The sound encoder extracts D-dimensional sound embeddings from the input mixture signal. We adopt a pre-trained sound event detection (SED) model [34] as the sound encoder to capture the sound event discriminated representation. The original SED model is pre-trained for an audio classification task and outputs an SED probability mass over the audio classes for an entire input audio clip. To keep dynamic time-dependent information, we use the frame-wise hidden embeddings obtained before the temporal aggregation layer of the SED model. Then, we use a trainable modality projection net, to obtain a D-dimensional sound embedding sequence $\mathbf{Q} \in \mathbb{R}^{T_a\times D}$, where T_a is the sound embedding sequence length. The modality projection net comprises a layer-norm module and a fully connected layer with ReLU activation. Modality projection nets of this structure are also used for the text and video clues to map the text and video embeddings to the D-dimensional unified space.

Text clue encoder: The text clue is a natural language description about the target sound. The function of the text encoder is to transform the target sound sentence into deep embeddings that can be used as the TSE clue. We adopt a pre-trained DistilBERT [35] as our text encoder, where DistilBERT is a self-supervised learning (SSL) model trained on large-scale text data. Token-level hidden embeddings are extracted with the pre-trained DistilBERT and then mapped by a trainable modality projection net to obtain D-dimensional text

Table 1: SNRi (dB) for seen and unseen test sets. ✓: clue is available in inference stage.

Model	Clues used			Seen	Unseen
	tag	text	video		
DCCRN-tag-clue	✓			6.4	6.0
DCCRN-text-clue		✓		6.3	5.9
DCCRN-video-clue			✓	5.9	5.6
DCCRN-multi-clue	✓	✓	✓	**6.9**	**6.5**
	✓	✓		6.8	6.4
		✓	✓	6.5	6.4
	✓		✓	6.6	6.4
	✓			6.4	6.2
		✓		6.3	6.0
			✓	5.8	5.9

clue embedding $\mathbf{O} \in \mathbb{R}^{T_t \times D}$ in the unified space, where T_t is the number of word tokens in the sentence.

Video clue encoder: The video clue is based on a video clip related to the target sound. As with the text clue encoder, we use an SSL model based on Swin Transformer [36] that is pre-trained on large-scale image data as our video clue encoder. Each image frame in the video clip is processed by the pre-trained Swin Transformer, and its output is mapped into the unified embedding space with a learnable modality projection net. The video clue embedding sequence is denoted as $\mathbf{V} \in \mathbb{R}^{T_v \times D}$, where T_v is the number of the image frames in the video clip.

Sound tag encoder: The sound tag encoder is a simple linear layer, it takes a one-hot sound event tag as input and outputs an embedding vector $\mathbf{E} \in \mathbb{R}^{1 \times D}$.

Multi-Clue Attention: The sound tag encoder fuses the embeddings from the text clue, video clue, and sound tag encoders to generate a clue sequence alinged with the embeddings from the sound encoder. We achieve this by using source-target attention [33], where the queries are extracted from the sound encoder output $\mathbf{Q}$ while the key-value pairs are obtained from the concatenated embeddings. Specifically, we concatenate the embedded clues [2] of different modalities along the sequence dimension to obtain the concatenated multi-modal clue $\mathbf{U}$ in the unified embedding space, i.e.,

$$\mathbf{U} = \text{Concatenate}(\mathbf{O}; \mathbf{V}; \mathbf{E}) \in \mathbb{R}^{(T_t + T_v + 1) \times D}. \qquad (8)$$

By using the sound embedding sequence, $\mathbf{Q}$, and $\mathbf{U}$ as the query and key-value pairs in the attention module respectively, we can get fused clue $\mathbf{C}_u$ as follows:

$$\mathbf{C}_u = \text{MultiHeadAttention}(\mathbf{Q}, \mathbf{U}, \mathbf{U}) \in \mathbb{R}^{T_a \times D}, \qquad (9)$$

where MultiHeadAttention(Q, K, V) represents a multi-head attention [33] with query Q, key K, and value V. We can see that the fused clue $\mathbf{C}_u$ have the same length, T_a, as the sound embedding $\mathbf{Q}$. To match the sequence length (T_a) to that of the DCCRN encoder T, $\mathbf{C}_u$ is up-sampled before being added to $\mathbf{Y}$ in Eqs. (2) and (3).

4. EXPERIMENTS

4.1. Dataset

Audio simulation. We first simulate a dataset for the TSE task based on AudioSet [27]. AudioSet is a large-scale audio dataset drawn from YouTube videos, and it has 527 sound classes labeled by humans. Most of the data in AudioSet are 10-second video clips with sound track. AudioSet is a weakly labeled dataset. There is usually more than one sound event in a video clip with labeled tags, and the occurrence time of the sound events is not provided. To ensure each clip has only one audio source and correct label, the pre-processing method from [10] was employed in our simulation. For each audio clip, a pre-trained SED model [34] is applied to locate its sound event anchor by comparing the SED probability in 10s audio clip. The audio of 2s around the event anchor is then selected for each clip for later simulation. For the training set, we clipped 64k (about 35h) sound sources of 463 classes for simulation and simulated 124k (about 70h) audio mixture of two sound sources. The signal-to-noise ratio (SNR) of the target sound is randomly sampled between -2 to 2 dB during the simulation. For validation and testing, we simulated 0.5h and 1h data, respectively. And all the target sound classes in the validation and test set are seen in the training. Besides, we also simulated 0.7h data of unseen target sound classes for testing, in which most of the sound classes are musical instruments.

Table 2: Impacts of compromised clues on SNRi (dB). $*$: compromised clue.

Correct/compromised clues				SNRi
text	text$*$	video	video$*$	
	✓			5.8
	✓	✓		5.8
			✓	5.4
✓			✓	6.3
	✓		✓	6.0

Text clue. The AudioCaps [28] provides human-written natural language description for a subset of the AudioSet. However, after the single-source clipping for the target sound, the description of the original 10s audio is no longer precise for the clipped-out 2s audio. So, we adopt an audio caption generation model [37] to create pseudo natural descriptions as the text clues. The 2s target sound is sent to the audio caption generation model, and the model outputs a sentence that describes the target sound.

Visual clue. The data in the AudioSet has parallel video and audio. For the visual clue, we simply use the frames (with FPS 15) from the $2s$ video that aligned with the target sound.

One-hot tag clue. As mentioned above, the AudioSet only provided weak sound class labels without the occurrence time, and we clipped out 2s audio from the original data with a SED model. During the clipping, we transform the SED probability of the 2s audio into one-hot vectors as the clue of the target sound.

4.2. Model configuration and training

The FFT size, the window length, and the hop length for STFT were set to 512, 400, and 100, respectively, for 16 kHz input. The channel numbers of the DCCRN encoder were $\{32, 64, 128, 256, 256, 256\}$. The complex LSTM consisted of two (real and imaginary) two-layer bidirectional LSTMs with 512 hidden units. We used ESPNet-SE toolkit [38] for implementation. Open text[3] and vision[4] encoders were used for the text and video clues.

The proposed multi-clue TSE system was trained with two stages. In the first stage, we first trained the DCCRN backbone by using only the sound tag clue. In the second stage, starting from this

[2] When one of the clues, for example, the video clue is missing, the video encoder obtains no input and outputs nothing. Then, Eq. 8 will be $\mathbf{U} = \text{Concatenate}(\mathbf{O}; \mathbf{E}) \in \mathbb{R}^{(T_t+1) \times D}$.

[3] https://huggingface.co/distilroberta-base

[4] https://huggingface.co/microsoft/swin-large-patch4-window7-224

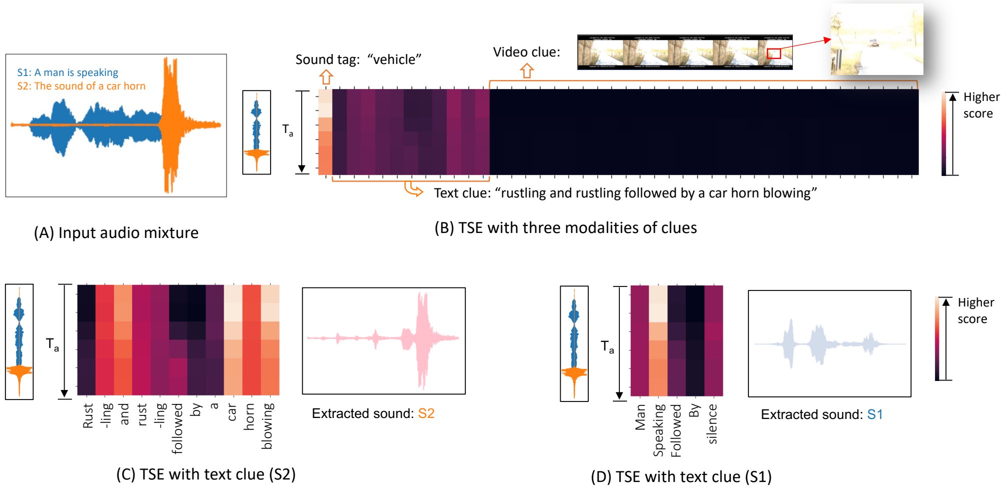


Fig. 2: Analysis of attention scores with different clues.

pre-trained model, we trained all the model parameters except for the pre-trained text and video encoders to minimize the extraction loss of Eq. (5). We adopted the two-stage approach because our preliminary experiment found that training the entire model from scratch in an end-to-end fashion based on Eq. (5) was hard to achieve optimal checkpoint. This could be because the clue processing net cannot produce stable clue embeddings in the early training steps. The two-stage training approach addresses this by starting the training with the simple sound tag-based clue. To enable fair comparison, the same initialization scheme was used for training the single-clue baselines. The training was performed by using an Adam optimizer with an initial learning rate of 0.5×10^{-4} and recuding it by a factor of 0.97 for every epoch until convergence.

4.3. Results

Comparison with baselines. Table 1 compares the proposed multi-clue TSE system with three single-clue TSE baselines for different clue-usage conditions in terms of SNR improvement (SNRi). The proposed TSE system achieved the best SNRi score for both *seen* and *unseen* test sets when it used all the three clues. Even with two clues, our multi-clue TSE outperformed all the single-clue baselines. When only one clue was provided, the proposed multi-clue TSE performed comparably with the single-cue baselines for the seen test set and marginally performed better for the unseen test set. These results show that the superiority of the proposed system in terms of both effectiveness and flexibility.

Robustness to compromised clues In real applications, some of the the input clues may sometimes become inaccurate. To test the robustness of the proposed model, we carried out experiments with text and video clues where one or two of them were artificially compromised. This was done as follows. For the text clue, one third of the words in the clue sentence were replaced with random words. For the video clue, a Gaussian noise of -2.5dB is added to the original video clips. Since there is no ambiguity in tag clues, we do not include the tag clues here. Table 2 shows the experimental results. While the performance degradation was observed when one or two clues were compromised, good SNR improvements were still observed for all conditions. By comparing the 1st row with the 2nd row, and the 3rd row with the 4th row, we found when the text clue was compromised, adding correct video clue helps, and vice versa. When both clues were compromised, it still showed better performance than only using one compromised clue.

4.4. Analysis of multi-clue attention

To analyze the attention mechanism by which the multi-clue net processes clues, we plot the attention weight matrix in figure. 2. The input audio is mixed by two sound events, a man's speech and the sound of a car horn. The available clues are the sound tag "vehicle", a natural description of the car horn, and a video in which a car is driving. Figure 2.b plots the attention scores in the multi-clue attention module. We can see that the highest scores are mainly in the tag and text clues. In this case, the video clues contributed little. That may be because the car in the video is too far from the camera. When we only use the text as the extraction clue, we can see both the sound of the car horn and the sound of speech can be extracted with different descriptions as figure 2.c and figure 2.d) show. And the words in the sentence related to the target sound get higher scores in the attention matrix.

5. CONCLUSION

In this work, we propose a multi-clue model for target sound extraction. The proposed model can freely combine different modal clues to extract target sounds. The experiments show that the proposed multi-clue TSE can leverage each single clue effectively to extract target sounds, achieving comparable performance with single-clue based systems. When combining clues from multiple modalities, the proposed model shows further improvements in both performance and robustness. These advantages make the proposed TSE system strong, robust, and flexible in real applications. With these observations, in future work, we would like to investigate further integration with more clues and their interaction with each other during TSE process.

6. REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical so-*

ciety of America, vol. 25, no. 5, pp. 975–979, 1953.

[2] J. R. Hershey, Z. Chen *et al.*, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE ICASSP*, Mar. 2016, pp. 31–35.

[3] M. Kolbæk, D. Yu *et al.*, "Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," *IEEE/ACM Trans. ASLP.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[4] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Trans. ASLP.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[5] Y. Liu, B. Thoshkahna *et al.*, "Voice and accompaniment separation in music using self-attention convolutional neural network," Mar. 2020.

[6] H. Liu, L. Xie *et al.*, "Channel-Wise Subband Input for Better Voice and Accompaniment Separation on High Resolution Music," in *Proc. ISCA Interspeech*. ISCA, Oct. 2020, pp. 1241–1245.

[7] A. Défossez, N. Usunier *et al.*, "Music Source Separation in the Waveform Domain," Apr. 2021.

[8] I. Kavalerov, S. Wisdom *et al.*, "Universal Sound Separation," in *Proc. IEEE ICASSP*, Oct. 2019, pp. 175–179.

[9] E. Tzinis, S. Wisdom *et al.*, "Improving Universal Sound Separation Using Sound Classification," in *Proc. IEEE ICASSP*, May 2020, pp. 96–100.

[10] Q. Kong, Y. Wang *et al.*, "Source Separation with Weakly Labelled Data: an Approach to Computational Auditory Scene Analysis," in *Proc. IEEE ICASSP*, May 2020, pp. 101–105.

[11] G. Li, X. Xu *et al.*, "Category-Adapted Sound Event Enhancement with Weakly Labeled Data," in *Proc. IEEE ICASSP*, May 2022, pp. 851–855.

[12] M. Delcroix, K. Zmolikova *et al.*, "Single Channel Target Speaker Extraction and Recognition with Speaker Beam," in *Proc. IEEE ICASSP*, Apr. 2018, pp. 5554–5558.

[13] Q. Wang, H. Muckenhirn *et al.*, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Proc. ISCA Interspeech*. ISCA, Sep. 2019, pp. 2728–2732.

[14] P. Wang, Z. Chen *et al.*, "Speech Separation Using Speaker Inventory," in *Proc. IEEE ASRU*, Dec. 2019, pp. 230–236.

[15] C. Xu, W. Rao *et al.*, "SpEx: Multi-Scale Time Domain Speaker Extraction Network," *IEEE/ACM Trans. ASLP.*, vol. 28, pp. 1370–1384, 2020.

[16] M. Delcroix, T. Ochiai *et al.*, "Improving Speaker Discrimination of Target Speech Extraction With Time-Domain Speakerbeam," in *Proc. IEEE ICASSP*, May 2020, pp. 691–695.

[17] T. Afouras, J. S. Chung *et al.*, "The Conversation: Deep Audio-Visual Speech Enhancement," in *Proc. ISCA Interspeech*. ISCA, Sep. 2018, pp. 3244–3248.

[18] 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)*, Jun. 2021, pp. 15 490–15 500.

[19] E. Tzinis, S. Wisdom *et al.*, "AudioScopeV2: Audio-Visual Attention Architectures for Calibrated Open-Domain On-Screen Sound Separation," Jul. 2022.

[20] K. Kilgour, B. Gfeller *et al.*, "Text-Driven Separation of Arbitrary Sounds," Apr. 2022.

[21] X. Liu, H. Liu *et al.*, "Separate What You Describe: Language-Queried Audio Source Separation," Mar. 2022.

[22] Y. Ohishi, M. Delcroix *et al.*, "ConceptBeam: Concept Driven Target Speech Extraction," Jul. 2022.

[23] C. Li and Y. Qian, "Listen, Watch and Understand at the Cocktail Party: Audio-Visual-Contextual Speech Separation," in *Proc. ISCA Interspeech*. ISCA, Oct. 2020, pp. 1426–1430.

[24] K. Tan, Y. Xu *et al.*, "Audio-Visual Speech Separation and Dereverberation With a Two-Stage Multimodal Network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 542–553, Mar. 2020.

[25] J. Li, M. Ge *et al.*, "VCSE: Time-Domain Visual-Contextual Speaker Extraction Network," in *Proc. ISCA Interspeech*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 906–910.

[26] A. Rahimi, T. Afouras *et al.*, "Reading To Listen at the Cocktail Party: Multi-Modal Speech Separation," 2022, pp. 10 493–10 502.

[27] J. F. Gemmeke, D. P. W. Ellis *et al.*, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP*, Mar. 2017, pp. 776–780.

[28] C. D. Kim, B. Kim *et al.*, "AudioCaps: Generating Captions for Audios in The Wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, Jun. 2019, pp. 119–132.

[29] Y. Hu, Y. Liu *et al.*, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. ISCA Interspeech*. ISCA, Oct. 2020, pp. 2472–2476.

[30] S. E. Eskimez, T. Yoshioka *et al.*, "Personalized speech enhancement: new models and comprehensive evaluation," in *Proc. IEEE ICASSP*, May 2022, pp. 356–360.

[31] O. Ronneberger, P. Fischer *et al.*, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer International Publishing, 2015, pp. 234–241.

[32] D. Stoller, S. Ewert *et al.*, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," in *Proceedings of the 19th ISMIR 2018*, 2018, pp. 334–340.

[33] A. Vaswani, N. Shazeer *et al.*, "Attention is All you Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[34] Q. Kong, Y. Cao *et al.*, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Trans. ASLP.*, vol. 28, pp. 2880–2894, 2020.

[35] V. Sanh, L. Debut *et al.*, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *CoRR*, vol. abs/1910.01108, 2019.

[36] Z. Liu, Y. Lin *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 9992–10 002.

[37] M. Wu, H. Dinkel *et al.*, "Audio Caption: Listen and Tell," in *Proc. IEEE ICASSP*, May 2019, pp. 830–834.

[38] C. Li, J. Shi *et al.*, "ESPnet-SE: End-To-End Speech Enhancement and Separation Toolkit Designed for ASR Integration," in *Proc. IEEE SLT*, Jan. 2021, pp. 785–792.