COMPSPOOF: A DATASET AND JOINT LEARNING FRAMEWORK FOR COMPONENT-LEVEL AUDIO ANTI-SPOOFING COUNTERMEASURES

Xueping Zhang[1]*, Liwei Jin*[2] *, Yechen Wang*[2]*, Linxi Li*[2]*, Ming Li*[1]

[1]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems,
Digital Innovation Research Center, Duke Kunshan University, Kunshan, China
[2]OfSpectrum, Inc., Los Angeles, USA

ABSTRACT

Component-level audio Spoofing (CompSpoof) targets a new form of audio manipulation where only specific components of a signal, such as speech or environmental sound, are forged or substituted while other components remain genuine. Existing anti-spoofing datasets and methods treat an utterance or a segment as entirely bona fide or entirely spoofed, and thus cannot accurately detect component-level spoofing. To address this, we construct a new dataset, CompSpoof, covering multiple combinations of bona fide and spoofed speech and environmental sound. We further propose a separation-enhanced joint learning framework that separates audio components apart and applies anti-spoofing models to each one. Joint learning is employed, preserving information relevant for detection. Extensive experiments demonstrate that our method outperforms the baseline, highlighting the necessity of separate components and the importance of detecting spoofing for each component separately. Datasets and code are available at: `https://github.com/XuepingZhang/CompSpoof`.

***Index Terms*—** Audio Anti-spoofing, Audio Deepfake Detection, Speech Separation, Component-Level Audio Anti-spoofing, Joint Learning

1. INTRODUCTION

Component-level audio anti-spoofing addresses a new type of audio manipulation where only specific components of a signal are forged or substituted while the rest remains authentic. Unlike conventional spoofing attacks [1, 2] that generate or convert an entire utterance or segment along the time axis, component spoofing operates at a finer granularity: for instance, the speech content may be substituted with a synthetic voice while keeping the genuine environment noise, or conversely, the original speech may be preserved while the environment sound is generated or substituted. Such component manipulations are more complex to detect because they can slip through systems made for whole-utterance or time-domain partial spoofing [3, 4], and also sound more real to human listeners.

Over the past decade, the ASVspoof [5, 6] and other related datasets [7, 8, 9, 10] have driven significant progress in audio anti-spoofing research. Current systems [11, 12, 13, 14] typically formulate spoof detection as binary classification between bona fide and spoofed utterances. Existing anti-spoofing methods [12, 15, 16, 17] have achieved strong results under these formulations. However, these approaches implicitly assume that an utterance is either entirely genuine or entirely spoofed. The ADD challenges [18, 19] and related datasets [18, 19, 20, 21] have recently highlighted an issue of partial spoofing, where only specific time spans within an utterance are fake. However, even this setting does not address the scenario of component-level spoofing. Existing methods are unable to evaluate the genuineness of separate audio components in a mixture, which leads to poor performance when spoofing affects only one component of the audio scene.

To address this gap, we introduce both a new component spoofing dataset and a tailored separation-enhanced joint learning framework. The dataset contains about 2,500 utterances formed by mixing bona fide and spoofed speech or environment audio from multiple sources. Each utterance belongs to one of five categories, covering all combinations of genuine and spoofed speech and environmental sound. Building on this dataset, we propose a separation-enhanced joint learning framework: a mixture detection model first identifies utterances that may contain synthetic or substituted content, after which each part is then passed to its own anti-spoofing model: one for speech and one for environmental sound. The outputs of these models are combined and mapped to five classes. To preserve more discriminative information for spoofing detection in the speech separation module, we jointly train the separation model with the anti-spoofing models. Our contributions can be summarized as follows.

- We first present the component-level audio anti-spoofing concept, and introduce the first component spoofing dataset, covering diverse combinations of genuine and spoofed speech/environment audio.
- We propose a joint learning framework that couples

Corresponding Author: Ming Li, ming.li369@dukekunshan.edu.cn

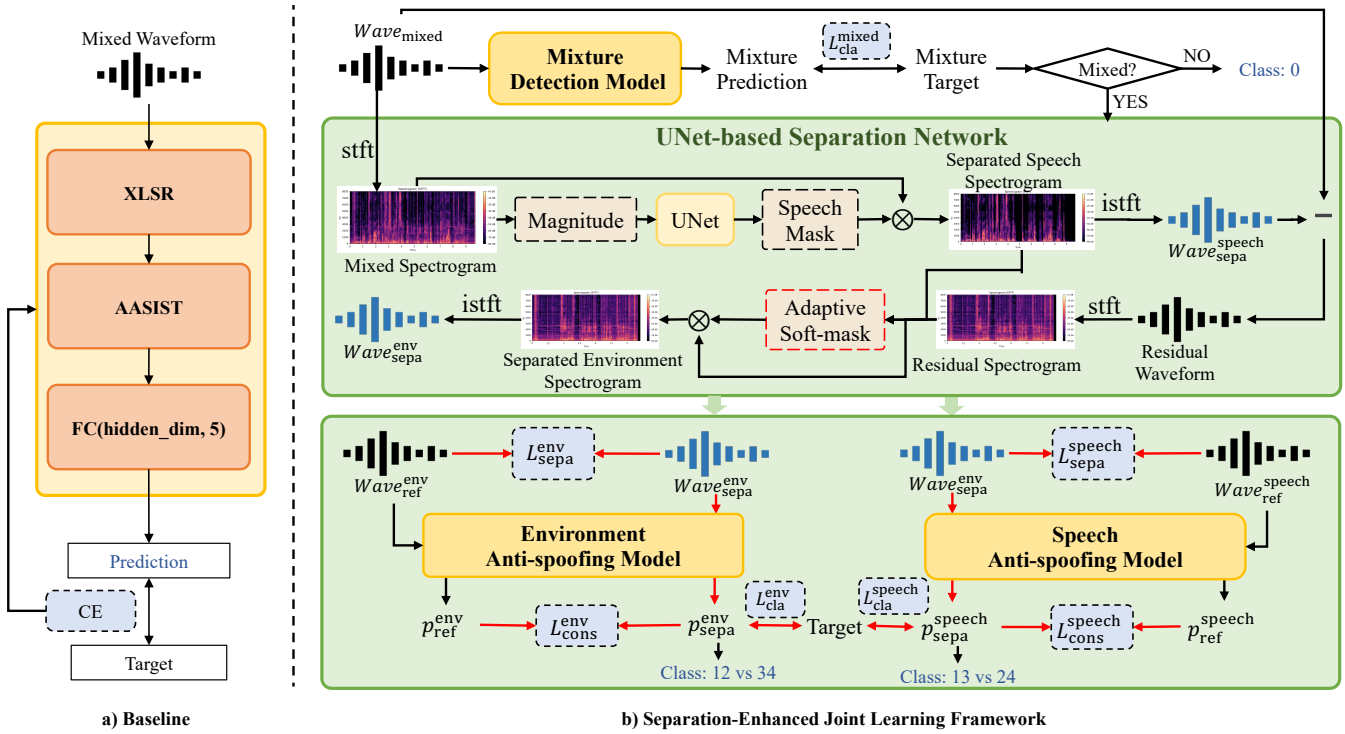

Fig. 1. Overview of the baseline and proposed separation-enhanced joint learning framework. " → " illustrates the joint learning data flow between the separation and anti-spoofing models.

separation and anti-spoofing, enabling separated signals to preserve spoof-relevant information.

- Extensive experiments demonstrate that our method outperforms the baseline, underscoring the importance of separating components and detecting spoofing for each.

2. COMPSPOOF DATASET

The CompSpoof dataset is designed for studying component-level anti-spoofing. The dataset comprises 2,500 audio samples, evenly distributed across five classes, with 500 samples per class. Table 1 summarizes the class definitions.

In the mixed part of this dataset, bona fide speech comes from ASVspoof5 [6] and CommonVoice [22], spoofed speech from ASVspoof5 [6] and SSTC [23], with bona fide environmental sounds from VGGSound [24] and spoofed environmental sounds from VCapAV [25]. Speech segments are chosen to contain clear voice activity, while environmental sounds are sampled from diverse scenarios such as indoor, street, and natural settings to ensure acoustic variety. In the original audio part of this dataset, we choose authentic audio utterances with both speech and simultaneously captured environmental audio signals from the VGGSound dataset [24]. The audio durations range from 5 to 21 seconds. More details can be found at: `https://xuepingzhang.github.io/CompSpoof-dataset/`.

During audio processing, all files are resampled to 16 kHz, with the shorter signal determining the final duration and longer ones truncated. To control the relative prominence of speech and environmental sound, the environmental sound is adjusted in amplitude to reach a predefined SNR [26] relative to the speech.

The dataset is partitioned into training, development, and evaluation sets using stratified sampling to maintain class balance, with a ratio of 70%, 10%, 20%.

3. SEPARATION-ENHANCED JOINT LEARNING FRAMEWORK

3.1. Baseline

Fig. 1 a) is the baseline framework; we adopt the XLSR-AASIST model [12], a widely used architecture for spoofing detection. Initially designed for binary classification (bona fide vs. spoofed), we extend it to a five-class classification task corresponding to the CompSpoof dataset. Although this direct extension is straightforward, the model does not explicitly disentangle the speech and environmental components, which may lead to confusion when only one component is spoofed. The limitation motivates the introduction of a separation-based framework.

Table 1. CompSpoof dataset class definitions

ID	Mixed	Speech	Environment	Class Label	Description
0	✓	Bona fide	Bona fide	original	Original bona fide speech and corresponding environment audio without mixing
1	✗	Bona fide	Bona fide	bonafide_bonafide	Bona fide speech mixed with another bona fide environmental audio
2	✗	Spoofed	Bona fide	spoof_bonafide	Spoof speech mixed with bona fide environmental audio
3	✗	Bona fide	Spoofed	bonafide_spoof	Bona fide speech mixed with spoof environmental audio
4	✗	Spoofed	Spoofed	spoof_spoof	Spoof speech mixed with spoof environmental audio

3.2. Separation-Enhanced Joint Learning Framework

Our method aims to explicitly separate speech and environmental sound components from an audio mixture and leverage them for robust anti-spoofing. As shown in Fig. 1(b), the framework consists of four models: a binary mixture detection model (implemented by XLSR-AASIST [12]), a UNet-based separation network, two dedicated anti-spoofing models for speech and environmental sound components (both implemented by XLSR-AASIST), and a joint learning mechanism that integrates their outputs. The details are introduced below.

UNet-based Separation Network: To explicitly separate speech and environmental sound, we design a UNet-based separation network that operates in the STFT domain. Given an input mixed waveform, the network first computes its STFT to obtain the complex spectrogram. The speech component is estimated by predicting a complex mask via the UNet [27], which is then applied to the mixture spectrogram. The inverse STFT (ISTFT) reconstructs the speech waveform from the masked spectrogram.

Since environmental sounds are highly diverse, obtaining a reliable environmental sound is a challenging task. We therefore compute the environmental sound in the STFT domain using an adaptive soft-mask [28]. Firstly, the remaining residual is computed by subtracting the separated speech waveform from the mixed waveform. Let $S(f,t)$ and $R(f,t)$ denote the magnitudes of the separated speech and residual spectrograms, respectively. We first compute a dynamic scaling factor α to balance the speech and residual magnitudes as Eq. 1.

$$\alpha = \frac{\text{mean}(|R(f,t)|)}{\text{mean}(|S(f,t)|) + \epsilon}, \tag{1}$$

where ϵ is a small constant for numerical stability. The environmental sound mask $M_{\text{env}}(f,t)$ is then defined as Eq. 2

$$M_{\text{env}}(f,t) = 1 - \tanh\Big(\frac{|S(f,t)|}{|R(f,t)| + \epsilon} \cdot \alpha\Big), \tag{2}$$

The soft-mask serves to suppress speech leakage in the residual, preventing residual speech from being misclassified as environmental sound. Moreover, the separated environment waveform is obtained via ISTFT. Finally, the network is trained using the Mean Squared Error (MSE) between the separated and reference waveforms for both speech and environmental sound.

Joint Learning: Training the separation and anti-spoofing models independently may cause the separation network to discard information that is important for detecting spoofed components. To address this, we adopt a joint learning strategy, where the separation network and the anti-spoofing models are trained together. Joint learning ensures that the separated signals retain features relevant to anti-spoofing.

In our framework, after obtaining the separated speech $W_{\text{sepa}}^{\text{speech}}$ and environmental sound $W_{\text{sepa}}^{\text{env}}$. Both the separated and the reference waveforms are then fed into the corresponding anti-spoofing models. The outputs from the separated components are compared to their target labels using cross-entropy loss $L_{\text{cls}}^{\text{speech}}$ and $L_{\text{cls}}^{\text{env}}$. In addition, a consistency loss L_{cons} is computed as the KL-divergence [29] between the predictions on the separated components and those on the reference waveform, encouraging the anti-spoofing outputs to be coherent, as shown in Eq. 3.

$$\begin{aligned} L_{\text{cons}} &= L_{\text{cons}}^{env} + L_{\text{cons}}^{speech} \\ &= \text{KL}(p_{\text{ref}}^{\text{env}} \,\|\, p_{\text{sepa}}^{\text{env}}) + \text{KL}(p_{\text{ref}}^{\text{speech}} \,\|\, p_{\text{sepa}}^{\text{speech}}), \end{aligned} \tag{3}$$

where $p_{\text{ref}}^{\text{speech}}$ and $p_{\text{ref}}^{\text{env}}$ are the softmax outputs from the reference components in the original mixture, and $p_{\text{sepa}}^{\text{speech}}$ and $p_{\text{sepa}}^{\text{env}}$ are from the separated speech and environmental sound.

Finally, the overall joint loss L_{joint} combines the MSE separation loss L_{sepa}, the mixture detection loss $L_{\text{cls}}^{\text{mixed}}$, the component-wise classification losses $L_{\text{cls}}^{\text{speech}}$ and $L_{\text{cls}}^{\text{env}}$, and the consistency loss L_{cons} as shown in Eq. 4

$$L_{\text{joint}} = \kappa * L_{\text{sepa}} + L_{\text{cls}}^{\text{mixed}} + L_{\text{cls}}^{\text{speech}} + L_{\text{cls}}^{\text{env}} + L_{\text{cons}}, \tag{4}$$

where κ is a constant. Joint training ensures the separation preserves spoof-relevant features while anti-spoofing models learn from both separated components and the reference waveform.

Inference: During inference, the mixed waveform is first passed through the mixture detection model to obtain a binary decision (c0 vs c1234). The separation model then processes the signal to generate speech and background components, which are individually evaluated by the speech detector and environment detector, yielding their own binary decisions (c13 vs c24 and c12 vs c34). These three decisions are then combined and mapped to one of the five target classes. The above procedure produces segment-level predictions. Segment-level predictions for all chunks of an audio

Table 2. Classification performance (Precision / Recall / F1) for baseline, Separation-Enhanced Framework (SEF), and Separation-Enhanced Framework with Joint Learning (SEF+JL) on dev and eval sets. The "Class" column shows IDs; the specific categories and their descriptions are provided in Table 1.

Method	Class	Dev	eval
Baseline	0	1.000 / 1.000 / 1.000	0.962 / 1.000 / 0.980
	1	0.746 / 0.820 / 0.781	0.827 / 0.860 / 0.843
	2	0.811 / 0.860 / 0.835	0.705 / 0.790 / 0.745
	3	0.778 / 0.700 / 0.737	0.860 / 0.800 / 0.829
	4	0.872 / 0.820 / 0.845	0.793 / 0.690 / 0.738
	ALL	0.841 / 0.840 / 0.840	0.829 / 0.828 / 0.827
SEF	0	1.000 / 1.000 / 1.000	0.990 / 1.000 / 0.995
	1	1.000 / 0.340 / 0.508	0.825 / 0.330 / 0.471
	2	0.710 / 0.440 / 0.543	0.646 / 0.420 / 0.509
	3	0.610 / 0.940 / 0.740	0.588 / 0.800 / 0.678
	4	0.613 / 0.920 / 0.736	0.561 / 0.889 / 0.688
	ALL	0.787 / 0.728 / 0.705	0.722 / 0.688 / 0.668
SEF +JL	0	1.000 / 1.000 / 1.000	0.980 / 1.000 / 0.990
	1	0.894 / 0.840 / 0.866	0.908 / 0.890 / 0.899
	2	0.860 / 0.980 / 0.916	0.903 / 0.840 / 0.871
	3	0.849 / 0.900 / 0.874	0.909 / 0.900 / 0.905
	4	0.977 / 0.840 / 0.903	0.841 / 0.909 / 0.874
	ALL	**0.916 / 0.912 / 0.912**	**0.908 / 0.907 / 0.908**

file are combined using majority voting to determine the final file-level label.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup

Preprocessing: All speech samples are normalized before being fed into the models. For the baseline methods, audio preprocessing follows the same procedure as in [12]. For the separation-based methods, audio is chunked with a window size of 4 seconds and a hop of 2 seconds. We performed Short-Time Fourier Transform (STFT) on the 16 kHz audio with a hop length of 16 ms and a window size of 64 ms.

Training: All models are trained on the same training and validation splits using Adam [30], with learning rates of 1×10^{-3} for the separation model and 1×10^{-5} for the anti-spoofing models. In the joint framework, models are trained independently for the first 4 epochs, then jointly from epoch 5. In the L_{total}, $\kappa = 10$.

Evaluation: We evaluate both the separation-based and baseline models on the dev and eval sets of CompSpoof, using file-level Precision, Recall, and F1 as the evaluation metrics.

4.2. Experimental results and analysis

Comparative experiments: Table 2 shows that SEF+JL consistently outperforms both the baseline and SEF, especially for mixed-content classes where speech and environment components differ. For example, in the spoof_bonafide class (Class ID = 2), F1 increases from 0.835 (baseline) to 0.916 (SEF+JL) on the development set. The improvement reflects that joint learning stabilizes classification across components and enhances robustness in challenging conditions.

In contrast, SEF without joint learning exhibits significant instability. While perfect performance is achieved on simple bona fide (Class ID = 0), F1 scores for mixed audio (Class ID = 1, 2, 3, 4) can drop to 0.508 or lower, indicating that separation alone may distort downstream representations without the guidance of joint learning. This emphasizes that joint optimization is critical for effectively leveraging separation in anti-spoofing detection.

Table 3. Segment-level detection performance (Precision / Recall / F1) on separated audio with/without Joint Learning (JL) on CompSpoof eval set.

Model	JL	Precision / Recall / F1
Speech Anti-spoofing	✓	0.860 / 0.875 / 0.863
	✗	0.777 / 0.764 / 0.720
Environment Anti-spoofing	✓	0.846 / 0.863 / 0.849
	✗	0.732 / 0.742 / 0.718

Segment-Level Model Analysis: Table 3 presents the segment-level performance of each detection model on both separated and original signals, with and without Joint Learning (JL). Here, Segment-level metrics reflect predictions on individual audio chunks prior to aggregation.

Table 3 shows that joint learning significantly improves the performance of anti-spoofing. For the speech anti-spoofing, F1 rises from 0.720 to 0.863, and for the environment anti-spoofing, F1 increases from 0.718 to 0.849. These improvements indicate that joint learning enhances the quality of separated representations and provides better supervision for downstream classification. Environment anti-spoofing consistently performs worse than speech anti-spoofing, indicating that the XLSR-AASIST-based environment anti-spoofing model may not be suited for this task.

5. CONCLUSIONS

We presented a component-level audio anti-spoofing method, tackling audio component manipulations where only speech or environmental sound is forged. To support this, we constructed CompSpoof, the first dataset covering all combinations of component-wise bona fide and spoofed speech or environmental sound. We proposed a separation-enhanced joint learning framework that separates audio components and applies dedicated anti-spoofing models while preserving spoof-relevant information. Experimental results demonstrate that our method outperforms baselines, highlighting the effectiveness of component separation and joint learning.

6. ACKNOWLEDGMENT

This research is funded by DKU foundation project "Emerging AI Technologies for Natural Language Processing". Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

7. REFERENCES

[1] Yogesh Kumar, Chamkaur Singh, "A deep learning approaches in text-to-speech system: a systematic review and recent research perspective," *Multimedia Tools and Applications*, vol. 82, no. 10, pp. 15171–15197, 2023.

[2] Tomasz Walczyna, Zbigniew Piotrowski, "Overview of voice conversion methods based on deep learning," *Applied sciences*, vol. 13, no. 5, pp. 3100, 2023.

[3] Zexin Cai, Ming Li, "Integrating frame-level boundary detection and deepfake detection for locating manipulated regions in partially spoofed audio forgery attacks," *Computer Speech & Language*, vol. 85, pp. 101597, 2024.

[4] Lin Zhang, Xin Wang, Erica Cooper, et al., "The partialspoof database and countermeasures for the detection of short fake speech segments embedded in an utterance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 813–825, 2022.

[5] Xuechen Liu, Xin Wang, Md Sahidullah, et al., "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[6] Xin Wang, Héctor Delgado, Hemlata Tak, et al., "Asvspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale," in *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof)*, 2024, pp. 1–8.

[7] Ricardo Reimao, Vassilios Tzerpos, "For: A dataset for synthetic speech detection," in *Proc. Speech Technology and Human-Computer Dialogue*, 2019, pp. 1–10.

[8] Nicolas Müller, Pavel Czempin, Franziska Diekmann, et al., "Does audio deepfake detection generalize?," in *Proc. Interspeech*, 2022, pp. 2783–2787.

[9] Combei David, Stan Adriana, Oneata Dan, et al., "Unmasking real-world audio deepfakes: A data-centric approach," in *Proc. Interspeech*, 2025, pp. 5343–5347.

[10] Han Yin, Yang Xiao, Rohan Kumar Das, et al., "EnvSDD: Benchmarking Environmental Sound Deepfake Detection," in *Interspeech 2025*, 2025, pp. 201–205.

[11] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, et al., "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. ICASSP*, 2022, pp. 6367–6371.

[12] Hemlata Tak, Massimiliano Todisco, Xin Wang, et al., "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *Proc. Odyssey*, 2022.

[13] Xiaohuan Chen, Wenhuan Lu, Ruiteng Zhang, et al., "Continual unsupervised domain adaptation for audio deepfake detection," in *Proc. ICASSP*, 2025, pp. 1–5.

[14] Neta Glazer, David Chernin, Idan Achituve, et al., "Few-Shot Speech Deepfake Detection Adaptation with Gaussian Processes," in *Proc. Interspeech*, 2025, pp. 2240–2244.

[15] Yang Xiao, Rohan Kumar Das, "Xlsr-mamba: A dual-column bidirectional state space model for spoofing attack detection," *IEEE Signal Processing Letters*, 2025.

[16] Tianchi Liu, Duc-Tuan Truong, Rohan Kumar Das, et al., "Nes2net: A lightweight nested architecture for foundation model driven speech anti-spoofing," *arXiv preprint arXiv:2504.05657*, 2025.

[17] Wen Huang, Xuechen Liu, Xin Wang, et al., "From sharpness to better generalization for speech deepfake detection," in *Proc. Interspeech*, 2025, pp. 5338–5342.

[18] Jiangyan Yi, Ruibo Fu, Jianhua Tao, et al., "Add 2022: the first audio deep synthesis detection challenge," in *Proc. ICASSP*, 2022, pp. 9216–9220.

[19] Jiangyan Yi, Jianhua Tao, Ruibo Fu, et al., "Add 2023: the second audio deepfake detection challenge," in *CEUR Workshop Proceedings*, 2023, vol. 3597, pp. 125–130.

[20] Jiangyan Yi, Ye Bai, Jianhua Tao, et al., "Half-truth: A partially fake audio detection dataset," in *Proc. Interspeech 2021*, 2021, pp. 1654–1658.

[21] Bowen Zhang, Terence Sim, "Localizing fake segments in speech," in *Proc. Pattern Recognition*, 2022, pp. 3224–3230.

[22] Rosana Ardila, Megan Branson, Kelly Davis, et al., "Common voice: A massively-multilingual speech corpus," in *Proc. Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[23] Ze Li, Yuke Lin, Tian Yao, et al., "The database and benchmark for the source speaker tracing challenge 2024," in *IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 1254–1261.

[24] Honglie Chen, Weidi Xie, Andrea Vedaldi, et al., "Vggsound: A large-scale audio-visual dataset," in *Proc. ICASSP*, 2020, pp. 721–725.

[25] Yuxi Wang, Yikang Wang, Qishan Zhang, et al., "VCapAV: A Video-Caption Based Audio-Visual Deepfake Detection Dataset," in *Proc. Interspeech*, 2025, pp. 3908–3912.

[26] Don H Johnson, "Signal-to-noise ratio," *Scholarpedia*, vol. 1, no. 12, pp. 2088, 2006.

[27] Daniel Stoller, Simon Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[28] Aarthi M Reddy, Bhiksha Raj, "Soft mask methods for single-channel speaker separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1766–1776, 2007.

[29] John R Hershey, Peder A Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *Proc. ICASSP*, 2007, vol. 4, pp. IV–317.

[30] DP Kingma, "Adam: a method for stochastic optimization," in *Proc. Learn Represent*, 2014.