applied sciences

MDPI

Article

Target Speaker Extraction with Cross-Correlation for Complex Spectra and Dual Post-Refinements

Sangwook Han, Seonggyu Lee and Jong Won Shin *

Department of the Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, Republic of Korea; swhan9873@gm.gist.ac.kr (S.H.); lsqjin2022@gm.gist.ac.kr (S.L.)
* Correspondence: jwshin@gist.ac.kr

Abstract

Target speaker extraction (TSE) aims to isolate speech spoken by a target speaker out of a mixture using speaker information in an enrollment utterance. Recently, several methods have been proposed that exploit the relationship between the enrollment utterance and the input mixture using cross-attention, without extracting speaker embeddings from the enrollment. Previous approaches applied the cross-attention to the encoded representations or to the real and imaginary parts of the compressed spectrograms separately, which may not have a physical meaning. In this paper, we propose a two-stage TSE method with a physically interpretable modified cross-attention block and a dual post-refinement structure. In the first stage, the attention weights to fuse the enrollment and mixture are derived from the cross-correlation between the complex spectra for the two signals in a form analogous to the phase-sensitive mask. The fused features along with the mixture features were subsequently fed into a speech extraction network to obtain a coarsely extracted target speech. The second stage consists of two parallel branches, where one branch refines the first-stage output using the enrollment in a similar way to the first stage, and the other utilizes the mixture to complement possibly attenuated target speech. In addition, the low-dimensional speaker embeddings extracted from the enrollment and the first-stage output are incorporated into the second stage to exploit the speaker discriminability. Experimental results show that the proposed method consistently outperformed existing TSE methods on the Libri2Mix dataset under both clean and noisy conditions, in terms of speech quality, speech intelligibility, and signal distortion measures.

Keywords: target speaker extraction; cross-correlation; phase-sensitive mask; dual post-refinement; speaker embeddings

Academic Editor: Douglas O'Shaughnessy

Received: 1 June 2026
Revised: 17 June 2026
Accepted: 19 June 2026
Published: 26 June 2026

1. Introduction

Target speaker extraction (TSE) is the task of isolating the target speech signal from a mixture using auxiliary information that specifies the target speaker, such as a pre-recorded enrollment utterance [1–27], spatial cues indicating the direction of the target speaker [23,28,29], or visual cues such as lip movements [23].

TSE differs from speech separation (SS) [30–36], which aims to reconstruct all speech sources in a mixture. It also differs from speech enhancement (SE) [37–44], which suppresses background noise and reverberation to improve speech quality and intelligibility. In contrast, TSE explicitly uses target speaker information to selectively extract the desired speech from a mixture. It is inspired by the well-known human auditory phenomenon known as the cocktail-party problem, in which humans can effortlessly concentrate on a

single speaker in the presence of background noises and multi-talker interference. However, the target speech can be partially masked by interfering speakers in multi-talker scenarios, and an initial extraction may still contain residual interference or attenuated target-speech components. These challenges highlight the importance of effectively utilizing target-speaker information and suggest that a two-stage refinement strategy can be useful when residual interference or target-speech attenuation remains after the initial extraction.

Based on these considerations, existing TSE approaches can be categorized according to how they utilize the auxiliary target speaker information. The most straightforward way to construct a TSE system using an enrollment utterance is to apply a speaker verification (SV) [45–49] model to obtain a low-dimensional speaker representation and use it as an additional information in the speech extraction network similar to the one for SS [2–16] as depicted in Figure 1a. Representative examples include SpEx [3] and SpEx+ [4], which extended Conv-TasNet [30] to enrollment-based TSE using multi-scale speech encoders, MC-SpEx [8], which enhanced multi-scale speaker conditioning, and X-SepFormer [10] and X-TF-GridNet [12], which incorporated speaker embeddings into advanced separation backbones such as SepFormer [33] and TF-GridNet [34]. Other studies have further improved the use of enrollment information through self-supervised disentangled speaker representations [11] or adaptive speaker embedding fusion [12]. While these methods demonstrated good performance by exploiting speaker discriminability through speaker classification losses, the low-dimensional speaker representations may not fully preserve the rich acoustic and speaker-related characteristics contained in the enrollment speech and provides limited interpretability regarding how the enrollment speech is matched with the mixture speech.

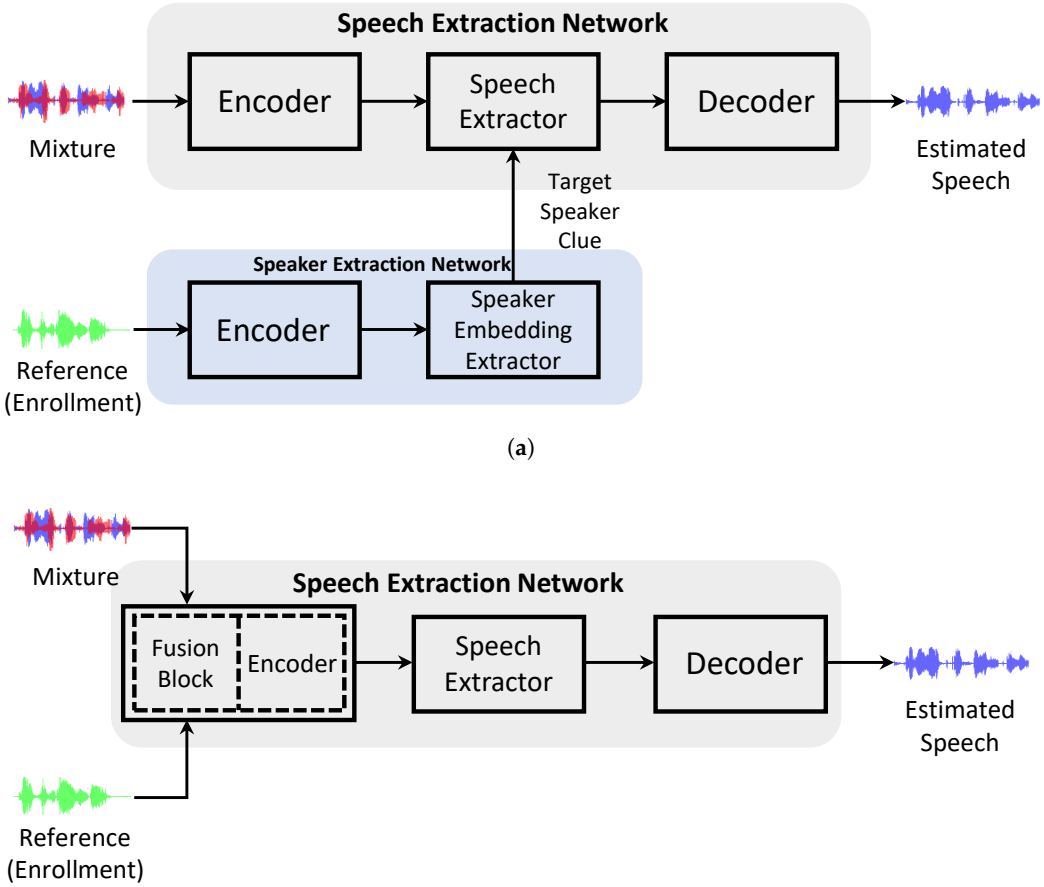

Figure 1. Block diagrams of target speaker extraction (TSE) with (**a**) a speaker embedding extracted from the enrollment speech, and (**b**) a fusion block to utilize the enrollment utterance itself. The dashed line in (**b**) indicates that the order of two blocks can be interchanged.

On the other hand, to fully exploit all information in the given enrollment speech, there have recently been attempts to utilize the enrollment speech itself in relation to the input mixture without relying on speaker embeddings [17–22], as illustrated in Figure 1b. In SMMA-Net [18], spectrogram matching was introduced to align the lengths of the enrollment, and mixture signals and then a mutual attention mechanism was used to fuse them. CIENet [19] adopted an attention-based interaction fusion mechanism in the pre-processing stage, where the real and imaginary components of the mixture and enrollment spectra were processed independently to obtain attention weights. CIENet-C2F [20] further extended CIENet by introducing a post-refinement stage, while DCF-Net [22] enhanced the interaction block with a mutually gated dual-stream fusion mechanism. USEF-TSE [21] provided a universal framework that can be applied to both time-domain and time-frequency-domain approaches, using cross-multi-head attention (CMHA).

Although previous approaches utilizing the enrollment utterance itself demonstrated superior performance, the cross-attention applied to the real and imaginary parts of the mixture and enrollment spectra separately or CMHA for encoded representations may limit the interpretability of how the enrollment speech is related with the mixture speech. Moreover, these methods mainly focus on the fusion between the enrollment and mixture signals, while the joint utilization of speaker-discriminative information is not considered. There was an attempt to use both the enrollment utterance itself and the speaker-discriminative embeddings in [27], but the performance without the speaker embeddings was better.

To address the limited interpretability of existing enrollment fusion strategies and the insufficient use of speaker discriminability, we propose a two-stage TSE with a cross-correlation-based fusion block and speaker embedding-based dual post-refinement. The proposed fusion block derives attention weights from the cross-correlation between the magnitude-compressed complex spectra of the enrollment and mixture speech, providing a more interpretable way to model their relationship. Specifically, the real-valued attention weights are obtained either from the magnitude of the complex-valued cross-correlation or from its phase-sensitive mask (PSM)-like modification [50]. Then, the fused features are concatenated with the mixture spectra and processed by a speech extractor to produce a coarsely extracted target speech. The coarsely extracted speech is subsequently refined in the second stage using two parallel branches. One branch processes the concatenation of the mixture spectra and coarsely extracted speech spectra to restore attenuated target speech in the first stage. The other branch applies a similar processing to the first stage using the enrollment and the first-stage output to further refine the target speech. In addition, low-dimensional speaker embeddings extracted from both the enrollment and the first-stage output are also utilized in the second stage through adapter modules to exploit speaker discriminability. Our experimental results on the Libri2Mix dataset demonstrated that the proposed method outperformed existing TSE approaches under both clean and noisy conditions in terms of the perceptual evaluation of speech quality (PESQ) scores [51], extended short-time objective intelligibility (ESTOI) [52], and scale-invariant signal-to-distortion ratio (SI-SDR) [53].

The remainder of this paper is organized as follows. In Section 2, we present the proposed TSE utilizing cross-correlation between complex spectra of the enrollment and mixture and dual-branch post-refinement. Experimental configurations are detailed in Section 3. Section 4 presents the experimental results and analyses. Finally, the conclusions are presented in Section 6.

2. Tse with Cross-Correlation for Complex Spectra and Dual Post-Refinements

To utilize the enrollment utterance itself in relation to the mixture input in a more interpretable manner and also make use of speaker discriminability through low-dimensional speaker embeddings, we propose CROCS, a two-stage target speaker extraction utilizing CROss-correlation for Complex Spectra of the enrollment and mixture and dual post-refinements. Overall block diagram of the CROCS is illustrated in Figure 2. In the first stage, the magnitude-compressed complex spectra of the enrollment and mixture are fused in the cross-correlation block and then processed by a speech extractor network similar to the MP-SENet [37,38] along with the mixture spectrogram. The second stage consists of two parallel branches that process the first-stage output with the mixture input and the enrollment, respectively. The low-dimensional speaker embeddings extracted from the enrollment speech and the first-stage output are also utilized in the two branches through the adapter modules. Detailed descriptions of the proposed system and contributions are given in the following subsections.

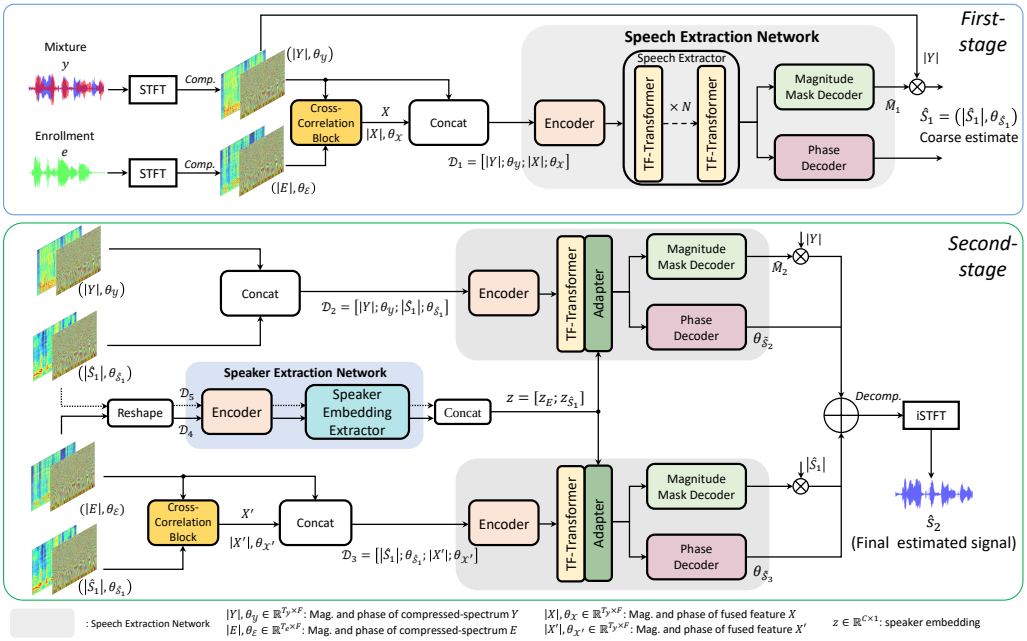

Figure 2. Overall structure of the proposed TSE system, CROCS. '*Comp.*' and '*Decomp.*' denote the power-law compression and decompression of magnitudes, respectively.

2.1. First Stage with Cross-Correlation Block

2.1.1. Previous Methods to Fuse Enrollment and Mixture

The interaction block of the CIENet [19,20] essentially applies the cross-attention to the real and imaginary parts of the magnitude-compressed complex spectra for the enrollment and mixture separately. From the mixture y and the enrollment e, each element of the magnitude-compressed spectrograms $Y \in \mathbb{C}^{T_y \times F}$ and $E \in \mathbb{C}^{T_e \times F}$ is obtained as

$$Y_{kf} = |\mathcal{Y}(k,f)|^c e^{j\theta_{\mathcal{Y}}(k,f)} = Y_{r_{kf}} + jY_{i_{kf}} \tag{1}$$

$$E_{kf} = |\mathcal{E}(k,f)|^c e^{j\theta_{\mathcal{E}}(k,f)} = E_{r_{kf}} + jE_{i_{kf}} \tag{2}$$

where $\mathcal{Y}(k,f)$ and $\mathcal{E}(k,f)$ are the complex spectra of y and e for the f-th frequency bin in the k-th frames, respectively, with the phases $\theta_{\mathcal{Y}}(k,f)$ and $\theta_{\mathcal{E}}(k,f)$. c is a compression factor, and Y_r, $Y_i \in \mathbb{R}^{T_y \times F}$ and E_r, $E_i \in \mathbb{R}^{T_e \times F}$ are the real and imaginary components of Y

and E, respectively. T_y, T_e, and F are the number of frames for the mixture and enrollment and the number of frequency bins, respectively.

The interaction block takes Y_r, Y_i, E_r and E_i as its input and produces the weighted enrollment features for real and imaginary parts, $X_r \in \mathbb{R}^{T_y \times F}$ and $X_i \in \mathbb{R}^{T_y \times F}$, i.e., $X_r = A_r E_r$ and $X_i = A_i E_i$ with the weight matrices $A_r, A_i \in \mathbb{R}^{T_y \times T_e}$. A_r and A_i are given as the normalized inner products between the real and imaginary parts of Y and E:

$$A_r = \text{softmax}\left(Y_r E_r^T\right) \tag{3}$$

$$A_i = \text{softmax}\left(Y_i E_i^T\right) \tag{4}$$

where softmax$(\cdot)$ is a softmax function applied to the last dimension of the argument. This procedure is essentially the same as applying the cross-attention to the real and imaginary parts of Y and E separately. However, determining the attention weights for the real or imaginary part based on the cross-correlation of only the real or imaginary parts of Y and E does not have any theoretical background, as it does not represent the relationship between the complex spectra of the enrollment and mixture. Meanwhile, in the SMMA-Net [18] and USEF-TFGridNet [21], $\mathcal{Y}$ and $\mathcal{E}$ are encoded with 2D convolutions and then fused through mutual attention [18] and CMHA [21], respectively. While the adoption of the convolution encoders may allow learnable transformation to produce features to be used for cross-attention, the interpretability of the TF domain features cannot be kept intact.

2.1.2. First Stage with Proposed Cross-Correlation Block

The attention weight matrix is originally designed to be a compatibility function that measures how well the *query* matches the *key*, which is applied to the *value* [54]. It would be desirable that the attention weight matrix A applied to E represents how the frames of the enrollment correlate with those of the mixture. To this end, the cross-correlation between the spectra for a certain frame of the enrollment and a frame of the mixture would be an intuitive measure on how closely they are related, as shown in Figure 3. The cross-correlation between the spectrum for the j-th frame of the mixture and that for the k-th frame of the enrollment becomes the (j,k)-th component of the matrix YE^H, i.e.,

$$CC(Y_j, E_k) = \left[YE^H\right]_{jk} = \sum_{f=1}^{F} |\mathcal{Y}(j,f)|^c |\mathcal{E}(k,f)|^c e^{j(\theta_{\mathcal{Y}}(j,f) - \theta_{\mathcal{E}}(k,f))} \tag{5}$$

where H denotes the Hermitian transpose, which is complex-valued and thus cannot be directly used as a real-valued attention weight. One straightforward way to obtain real-valued weights is to exploit the magnitude of the cross-correlation between Y_j and E_k, i.e.,

$$[A_M]_{jk} = \text{softmax}\left(\left|CC(Y_j, E_k)\right|/\tau\right) = \frac{\exp\left(\left|CC(Y_j, E_k)\right|/\tau\right)}{\sum_{k'=1}^{T_e} \exp\left(\left|CC(Y_j, E_{k'})\right|/\tau\right)} \tag{6}$$

where the softmax function is applied in analogy to the scaled dot-product attention [54], τ is a temperature parameter that adjusts the concentration of the attention weights, and the subscript M represents the magnitude of the cross-correlation.

The attention weights A_{mag} based on the cross-correlation between the magnitude spectra, $CC_{mag}(Y_j, E_k) = \sum_{f=1}^{F} |\mathcal{Y}(j,f)|^c |\mathcal{E}(k,f)|^c$, can also be considered, but A_{mag} tends to have high values for a few frames in the enrollment which have high energies among the frames with similar pitch values. It leads to a loss of information, as the output of this block will be used instead of E in the speech extraction network.

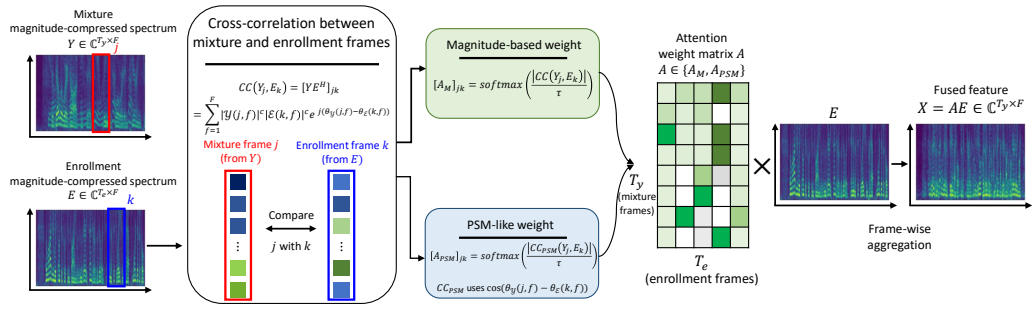

Figure 3. Illustration of the proposed cross-correlation-based fusion block.

An alternative way to derive real-valued weights is to consider the phase difference between the enrollment and mixture in each frequency bin explicitly in a similar form to the PSM [50] used for the SE. In the PSM, the phase difference between the clean and noisy speeches contributes to the target magnitude gain via a cosine function. Likewise, we can penalize the cross-correlation between spectral magnitudes with the cosine of the phase difference between the mixture and enrollment. The resulting PSM-like modification of the cross-correlation between the j-th frame of the mixture and the k-th frame of the enrollment is given by

$$CC_{PSM}(Y_j, E_k) = \sum_{f=1}^{F} |\mathcal{Y}(j,f)|^c |\mathcal{E}(k,f)|^c \cos\big(\theta_{\mathcal{Y}}(j,f) - \theta_{\mathcal{E}}(k,f)\big), \tag{7}$$

which coincides with the real part of the $CC(Y_j, E_k)$. This allows the correlation score to become larger when the mixture and enrollment both have similar spectral magnitudes by penalizing phase difference. Compared with magnitude-only correlation $CC_{mag}(Y_j, E_k)$, the PSM-like formulation incorporates the phase difference through the cosine term, making the attention weights physically more interpretable as a measure of both spectral magnitude similarity and phase alignment. As a result, the (j,k)-th component of the attention weight matrix A_{PSM} becomes

$$[A_{PSM}]_{jk} = \text{softmax}\big(CC_{PSM}(j,k)/\tau\big). \tag{8}$$

Subsequently, the weight matrix $A \in \{A_M, A_{PSM}\}$ is applied to E to produce the fused feature X, i.e., the weighted enrollment feature, as

$$X = AE \in \mathbb{C}^{T_y \times F}. \tag{9}$$

This operation can be interpreted as a frame-wise aggregation of the enrollment spectra to match each frame of Y as closely as possible, under the assumption that the interfering speech components are not represented in the enrollment spectra.

The magnitude and phase of the fused feature X, $|X|$ and $\theta_{\mathcal{X}}$, are concatenated with those for Y, $|Y|$ and $\theta_{\mathcal{Y}}$, to construct the input tensor for the speech extraction network, $\mathcal{D}_1 = \big[|Y|;\ \theta_{\mathcal{Y}};\ |X|;\ \theta_{\mathcal{X}}\big] \in \mathbb{R}^{4 \times T_y \times F}$. As for the speech extraction network, we adopt the MP-SENet [38] architecture, except that the input channel dimension is changed to 4 and the global layer normalization (gLN) [30] is applied at the end of the encoder. $\mathcal{D}_1$ is transformed into a latent TF representation through the encoder, which is then processed by a stack of N TF-transformer blocks with C channels to capture time and frequency dependencies stage by stage. Subsequently, the magnitude mask decoder and phase decoder produce the estimates of the mask for the compressed magnitude spectrum and wrapped phase spectrum of the clean speech, $\hat{M}_1$ and $\theta_{\hat{\mathcal{S}}_1}$, respectively, where $\hat{M}_1$ is applied to the mixture magnitude $|Y_{kf}|$ to construct the compressed spectrum of the clean speech

as $\hat{S}_1(k,f) = \hat{M}_1(k,f)|Y_{kf}|e^{j\theta_{\hat{S}_1}(k,f)} = |\hat{\mathcal{S}}_1(k,f)|^c e^{j\theta_{\hat{S}_1}(k,f)}$. A coarse estimate of the target speech in the form of $|\hat{\mathcal{S}}_1|^c$ and $\theta_{\hat{S}_1}$ is used in the second stage to obtain a better estimate of the target speech.

2.2. Dual Post-Refinement Utilizing Speaker Embeddings

Instead of processing $\mathcal{D}_1$ in Section 2.1.2 using many TF-transformer blocks, we employ a two-stage structure of which the second stage consists of two parallel branches to refine the first-stage output using the mixture input and the enrollment, respectively. Low-dimensional speaker embeddings extracted from the enrollment and the first-stage output are also adopted to exploit speaker discriminability. Detailed descriptions for the dual post-refinement structure and the adoption of the speaker embeddings are given below.

2.2.1. Dual Post-Refinement

The second stage consists of two parallel branches as depicted in Figure 2. The top branch processes the concatenation of the first-stage output and mixture, $\mathcal{D}_2 = [|Y|;\ \theta_y;\ |\hat{S}_1|;\ \theta_{\hat{S}_1}]$, with a speech extraction network consisting of an encoder, a single TF-transformer block followed by an adapter to incorporate the speaker embedding, and the magnitude masks and phase decoders to produce $\hat{M}_2$ and $\theta_{\tilde{S}_2}$. On the other hand, the cross-correlation block, which takes $\hat{S}_1$ and E instead of Y and E as input, is used in the bottom branch to produce the fused feature X', which is fed into the speech extraction network along with $\hat{S}_1$. The speech extraction network that takes $\mathcal{D}_3 = [|\hat{S}_1|;\ \theta_{\hat{S}_1}; |X'|;\ \theta_{\mathcal{X}'};\ ]$ as input has an identical structure to that in the top branch. The top branch is expected to restore possibly lost components of the target speaker in the first stage by utilizing the mixture input, while the bottom one applies the same processing as the first stage to further refine the first-stage output.

2.2.2. Extraction and Utilization of Speaker Embeddings

Although the TSE approaches utilizing the relationship between the enrollment utterance and the input mixture directly can fully utilize all the details in the enrollment utterance and have shown impressive performances, the speaker discriminability, which is the main training objective in speaker recognition, is not enforced in most of them. In this work, we utilize speaker-discriminative low-dimensional embeddings in the second stage to complement the cross-correlation-based fusion of the excerpts, as shown in Figure 2. As the enrollment E is a target speaker's voice sample without interferences and $\hat{S}_1$ contains the target speech in the input mixture Y, both speaker embeddings from E and $\hat{S}_1$, z_E and $z_{\hat{S}_1}$, are used to condition the speech extraction processes in the second stage. Instead of using a conventional speaker recognition architecture, we adopt the ConvNeXt-v2 [55] as the speaker embedding extractor, which showed better performance than Thin-ResNet34 [47], ECAPA-TDNN [46], and the U^2-Net in the X-TF-GridNet [12] when applied to the CROCS. Specifically, the speaker extraction network first encodes $\mathcal{D}_4 = [|E|;\theta_{\mathcal{E}}]$ or $\mathcal{D}_5 = [|\hat{S}_1|;\theta_{\hat{S}_1}]$ using an encoder which has the same structure with the one used for the speech extraction network after reshaping them from $2 \times T \times F$ to $4 \times (T/2) \times F$, $T \in \{T_e, T_y\}$. The encoded representations are fed into the speaker embedding extractor, which has a structure of the ConvNeXt-v2 [55] to produce z_E and $z_{\hat{S}_1}$ through a temporal average pooling (TAP) layer, respectively. It is noted that the encoder used for the speaker extraction network does not share weights with the one used for the speech extraction network. Finally, two speaker embeddings, z_E and $z_{\hat{S}_1} \in \mathbb{R}^{\frac{C}{2}\times 1}$ are concatenated along the channel axis to form the conditioning variable $z = [z_E; z_{\hat{S}_1}] \in \mathbb{R}^{C\times 1}$ in which C is the number of channels in the MP-SENet-based speech extractor.

As an adapter module to fuse z into the speech extraction network, we adopt the adaptive embedding aggregation (AEA) fusion block in the X-TF-GridNet [12]. The AEA fusion block is composed of a lightweight multi-head attention block and a 1×1 2D convolution block, in which the speaker embedding z is utilized as the *query*, and the intermediate features of the TF-transformer in each branch become the *key* and *value* of the cross-multi-head attention. This adapter module is placed after the TF-transformer, and the output of it is used in the magnitude mask and phase decoders as shown in Figure 2. Finally, the outputs of the speech extraction network in two branches, $\hat{M}_2, \theta_{\tilde{S}_2}$ and $\hat{M}_3, \theta_{\tilde{S}_3}$, are used to reconstruct the magnitude-compressed spectrum of the target speech as

$$\begin{aligned} \hat{S}_2(k,f) &= \left(\hat{M}_2(k,f)|Y_{kf}| + \hat{M}_3(k,f)|\hat{S}_1(k,f)|\right) e^{j(\theta_{\tilde{S}_2}(k,f)+\theta_{\tilde{S}_3}(k,f))} \\ &= |\hat{\mathcal{S}}_2(k,f)|^c e^{j\theta_{\hat{S}_2}(k,f)}. \end{aligned} \tag{10}$$

in which $\hat{\mathcal{S}}_2(k,f)$ denotes the complex-valued spectrum reconstructed from the estimated magnitude and phase in the second stage. The time-domain signal, $\hat{s}_2$, can be obtained by applying the inverse short-time Fourier transform (iSTFT) to the magnitude-decompressed spectrum $\hat{\mathcal{S}}_2$, i.e.,

$$\hat{s}_2 = \text{iSTFT}(\hat{\mathcal{S}}_2). \tag{11}$$

2.3. Training Objectives

The proposed TSE system, CROCS, is optimized with a multi-task learning that combines the reconstruction objectives and speaker classification objective. The overall loss function is given as a linear combination of the negative SI-SDRs for the first- and second-stage outputs and a cross-entropy loss for the speaker classification:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{si-sdr}}(s, \hat{s}_1) + \lambda_2 \mathcal{L}_{\text{si-sdr}}(s, \hat{s}_2) + \lambda_3 \mathcal{L}_{\text{ce}}(z, \ell_{\text{spk}}) \tag{12}$$

in which λ_1, λ_2, and λ_3 represent the weighting factors for the first-stage extraction loss, second-stage refinement loss, and speaker classification loss, respectively, s is the target speech, and $\hat{s}_1$ is the reconstructed time-domain signal obtained in the first stage, i.e., $\hat{s}_1 = \text{iSTFT}(\hat{\mathcal{S}}_1)$. The negative SI-SDR loss, $\mathcal{L}_{\text{si-sdr}}$, is defined as

$$\mathcal{L}_{\text{si-sdr}}(s, \hat{s}) = -10 \log_{10} \frac{\|\alpha s\|_2^2}{\|\hat{s} - \alpha s\|_2^2}, \tag{13}$$

where $\alpha = \langle \hat{s}, s \rangle / \langle s, s \rangle$ is the optimal scaling factor and $\langle \cdot, \cdot \rangle$ denotes the inner product. The cross-entropy loss to enforce speaker discriminability is computed from the output of a linear classifier applying one fully connected layer to z, $f(z)$, which is given by

$$\mathcal{L}_{\text{ce}}(z, \ell_{\text{spk}}) = -\sum_{i=1}^{J} \ell_{\text{spk},i} \log \left(\frac{\exp(f_i(z))}{\sum_{j=1}^{J} \exp(f_j(z))} \right), \tag{14}$$

where J is the number of speakers, $\ell_{\text{spk},i} \in \{0, 1\}$ is the i-th element of the one-hot encoded speaker label, and $f_i(z)$ denotes the i-th element of $f(z)$.

3. Experimental Setup

3.1. Datasets

To validate the proposed methods, we conducted experiments on the Libri2Mix dataset [56] under both clean and noisy conditions. Libri2Mix consists of two-speaker

mixtures generated from LibriSpeech [57], and the noisy condition additionally includes environmental noise from the WHAM! dataset [58]. For both conditions, we used the `train-100`, `dev`, and `test` subsets for training, validation, and evaluation, respectively. The `train-100` subset contains 13,900 mixtures from 251 speakers, while each of the `dev` and `test` subsets contains 3000 mixtures from 40 non-overlapping speakers. During training, each speaker in a mixture was considered as the target speaker in turn, and the enrollment utterance was randomly selected from utterances not used in the mixture. For validation and evaluation, we followed the enrollment setting used in prior works (https://github.com/BUTSpeechFIT/speakerbeam, accessed on 17 June 2026). We used the *min* version, where each mixture length is matched to the shorter utterance [56]. To allow comparison with previous studies, we report results at both 8 kHz and 16 kHz sampling rates.

To further evaluate the performance in the presence of background noise and reverberations, we have conducted additional experiments on the WHAMR! [59] dataset, which extended the WSJ0-2mix dataset [60] by introducing noises and reverberations. The reverberation time (RT60) for each mixture was selected from the range $[0.2, 1]$ s, and the distance between the speakers and the microphone spanned from 0.66 to 2 m. The level of the second speech in each mixture was between $[-5, 5]$ dB from that of the first speech, and the signal-to-noise ratio (SNR) between the louder speech and noise was randomly selected from $[-6, 3]$ dB. The dataset consisted of 20,000, 5000, and 3000 mixtures for training, validation, and test, respectively. During training and performance assessment, the direct-path signal for the target speech at the first microphone was considered as the target signal. The sampling rate for all audio was 8 kHz, and the *min* version was used.

3.2. Implementation Details

We used the Hann window with a length of 20 ms and a hop size of 10 ms. The 256- and 512-point STFTs were applied for the sampling rate of the 8 kHz and 16 kHz, respectively, which made F 129 and 257, respectively. The compression factor c was set to 0.3. Prior to training, both mixture and enrollment signals were normalized to have a unit variance. During training, mixtures were randomly cropped into 4-second-long segments, while the enrollment signals were kept intact. For validation and evaluation, both mixture and enrollment signals were not truncated. For the speech extraction network, the number of channels C and the number of attention heads M of the MP-SENet (https://github.com/yxlu-0102/MP-SENet, accessed on 17 June 2026) were set to 80 and 4, respectively. The number of the TF-transformer blocks in the speech extractor for the first stage was set to 3. For the speaker embedding extractor, we used a ConvNeXt-v2 model consisting of 4 stages with [2, 2, 3, 2] blocks, and the number of nodes in the last fully connected layer was set to match the number of speakers in the training dataset. The total number of parameters in the speaker extraction network was 0.288 M. The adapter module was configured in the same manner as in [12], except that the learnable weights for the multihead attention block output were initialized to 1 in training.

All models were trained from scratch for 150 epochs with a batch size of 2. The speaker extraction network was not pre-trained. We used the Adam [61] optimizer with an initial learning rate of 10^{-3}. The learning rate was multiplied by 0.98 for every two epochs during the first 100 epochs and multiplied by 0.9 in the remaining 50 epochs. The weight factors for losses, λ_1, λ_2, and λ_3 in (12), were set to 0.4, 0.6, and 0.5, respectively. The temperature τ for the softmax operation in the cross-correlation block was set to 4. In addition, gradient clipping with a maximum l_2 norm of 5 was applied. We selected the model that achieved the best performance on the `dev` set for evaluation. We did not apply any dynamic mixing or data augmentation techniques. All experiments were carried out on two NVIDIA A100

GPUs using a PyTorch 2.5.1 [62] deep learning framework. The performance measures used to compare the TSE systems include the narrowband PESQ (PESQ-NB) scores [51], wideband PESQ scores (PESQ-WB) [63], source-to-distortion ratio (SDR) [64], ESTOI [52], and SI-SDR [53].

For the computational complexity, the number of model parameters is reported in millions (M), and the computational cost is reported in Giga multiply–accumulate operations per second (GMAC/s) computed using ptflops 0.7.5 (https://github.com/sovrasov/flops-counter.pytorch, accessed on 17 June 2026) with a batch size of 1, where both the mixture and enrollment utterances were measured in 1-second segments. It should be noted that the computational cost can increase with longer enrollment utterances because the attention weight matrix has a size of $(T_y \times T_e)$. However, after the fused feature is obtained, the subsequent speech extraction network mainly depends on the mixture length. The source code, the trained model, and audio samples can be accessed at our GitHub page (https://github.com/aryanorb/CROCS-TSE, accessed on 17 June 2026).

4. Results and Analysis

4.1. Analysis of Each Contribution of CROCS

Firstly, we performed an ablation study to evaluate the effects of the contributions of the proposed system, CROCS, including the cross-correlation block, two-stage structure, incorporation of the low-dimensional speaker embedding, and dual post-refinements. Table 1 shows the performance of various configurations of the TSE system on the Libri2Mix-clean dataset sampled at 8 kHz, along with the number of parameters and the computational complexity. The first three rows compare the effect of different fusion methods to combine the enrollment and mixture using only the first stage. In this single-stage configuration, the number of the TF-transformer blocks was set to 5, which is the same as the total number of the TF-transformer blocks in the CROCS (3 in the first stage, 1 for each branch of the second stage). The proposed fusion methods using magnitude-based A_M and spectral magnitude with phase difference A_{PSM} outperformed the interaction fusion block of the CIENet [20], which applied cross-attention for the real and imaginary parts separately using A_r and A_i in all three metrics. Between the two proposed methods, the PSM-like modification of the cross-correlation showed the best performance, with an improvement of the PESQ score of 0.08, the ESTOI of 1.30%, and the SI-SDR of 0.97 dB over the fusion in the CIENet [20]. We have also examined the effect of incorporating the low-dimensional speaker embeddings in the single-stage setting, with the adapter modules in all TF-transformer blocks (#4) and the last two blocks (#5). The introduction of the speaker embedding in the single-stage setting did not improve the performance of (#3) regardless of the number of adapter modules. It is noted that only z_E from the enrollment utterance was used in the single-stage setting (#4) and (#5), and $(\lambda_1, \lambda_2, \lambda_3)$ in Equation (12) were $(1, 0, 0.5)$.

Table 1. Experimental results on Libri2Mix-clean dataset sampled at 8 kHz for various configurations of the TSE system.

#ID	Fusion in 1st Stage	Fusion in 2nd Stage	Two-Stage	Speaker Embedding	Dual Post-Refinements	TF-Transformer Blocks	Adapter Blocks	# Params. (M)	GMAC/s	PESQ	ESTOI (%)	SI-SDR (dB)
1	A_r&A_i [20]	None	✗	✗	✗	{5, - }	-	4.10	34.34	3.45	87.90	16.17
2	A_M	None								3.46	88.33	16.52
3	A_{PSM}	None								3.53	89.20	17.14
4	A_{PSM}	None	✗	✓	✗	{5, - }	5	5.39	37.81	3.41	87.84	16.18
5							2	5.28	37.30	3.49	88.67	16.69
6	A_r&A_i [20]	Concat $(Y\&\widehat{S}_1)$	✓ †	✗	✗	{3, 2}	-	5.35	45.01	3.50	88.49	16.90
7	A_{PSM}	Concat $(Y\&\widehat{S}_1)$	✓	✗	✗	{3, 2}	-	5.35	45.01	3.53	89.24	17.22
8		$A_{PSM}(\widehat{S}_1\&E)$							45.04	3.51	88.99	16.90
9	A_{PSM}	Concat $(Y\&\widehat{S}_1)$	✓	✓	✗	{3, 2}	2	5.72	46.01	3.53	89.48	17.54
10		$A_{PSM}(\widehat{S}_1\&E)$							46.04	3.41	87.77	16.28
11	A_{PSM}	Both	✓	✗	✓	{3, {1, 1}}	{1, 1}	6.61	55.71	3.58	89.82	17.62
12	A_{PSM}	Both	✓	✓	✓	{3, {1, 1}}	{1, 1}	6.97	56.70	**3.60**	**89.92**	**17.67**

† indicate that the coarse-stage output is added to the fine-stage output via a residual connection as in the CIENet-C2F [20]. ✓ indicates that the corresponding component is used, whereas ✗ indicates that it is not used, Bold values indicate the best performance in each column.

The next section (#6−#10) is on the experiments using the TSE systems consisting of two stages in which the second stage has only one of the two branches of the CROCS. The number of the TF-transformer blocks in the second stage was set to 2, matching the total number of the TF-transformer blocks in the TSE system. The configuration in (#6) is similar to the CIENet-C2F [20], with a second stage taking the concatenation of Y and $\hat{S}_1$, and an additional residual connection from the first-stage output to the final output. The performance of (#6) was better than (#1) with a single stage, but worse than (#7) with the proposed cross-correlation block in all the metrics. The system with the other branch taking the PSM-like modification of the cross-correlation between $\hat{S}_1$ and E, (#8), showed worse performance than (#7). Systems (#9) and (#10) were built upon (#7) and (#8) using speaker embeddings $z = [z_E; z_{\hat{S}_1}]$ through the adapter modules, which brought about a slight improvement in the ESTOI and SI-SDR. $\lambda_1, \lambda_2, \lambda_3$ values for (#6−#8) and (#9, #10) were $(0.4, 0.6, 0)$ and $(0.4, 0.6, 0.5)$, respectively. The final two rows of Table 1 are with the proposed cross-correlation block and dual post-refinement structure, without and with speaker embeddings. These two systems outperformed (#1−#10), and the proposed CROCS in (#12) marked the best performance. This improvement may be attributed to the distinct roles of the two post-refinement branches, where the top branch restores attenuated target-speaker components and the bottom branch further refines the first-stage output using the same processing as in the first stage. These experimental results also showed that all components of the contributions including the introduction of the proposed cross-correlation block (from (#1) to (#3)), two-stage architecture (from (#3) to (#7)), dual post-refinements structure (from (#7) to (#11)), and the introduction of speaker embeddings (from (#11) to (#12)) contributed to the performance improvement. It is noted that the adoption of the speaker embeddings required only 0.365 M parameters and 0.99 GMAC/s, although it only slightly improved the performance. This marginal improvement may be attributed to the lightweight design of the speaker embedding branch, which may have limited its ability to capture highly discriminative speaker characteristics.

Figure 4 presents a sample-wise analysis of the SI-SDR improvements over the baseline system (#1) for several systems in Table 1. The histogram indicates that the performance gains from each contribution are not limited to a small subset of samples, but are reflected as an overall shift in the performance distribution.

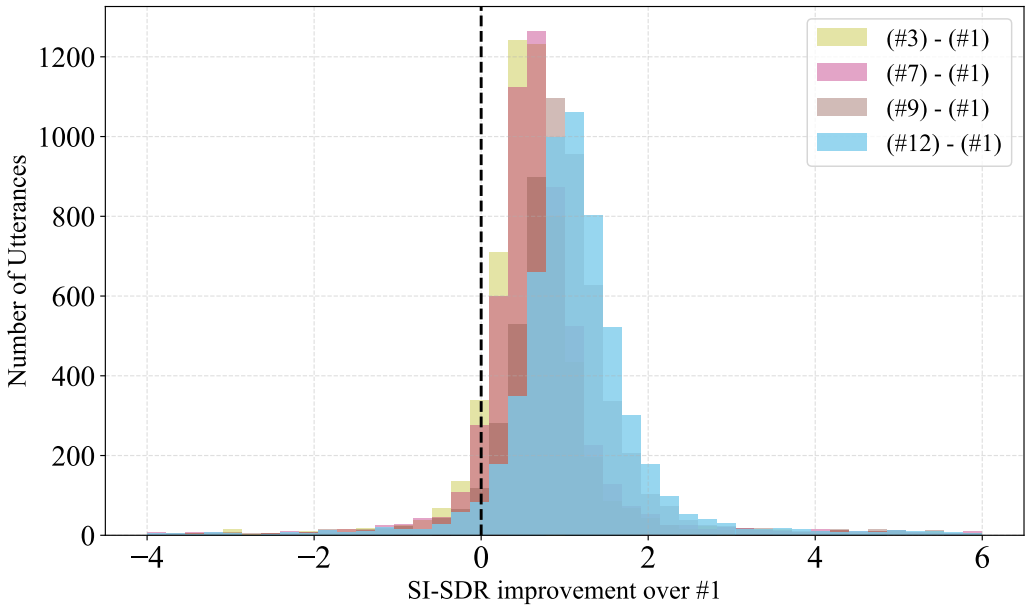

Figure 4. Histogram for SI-SDR improvements over system (#1) in terms of the sample-wise analyses for the systems in Table 1.

To provide more insight into the proposed cross-correlation block, Figure 5 visualizes the magnitude spectra (a) an enrollment speech, (b) a mixture, (c) a target speech, (d) a fused signal using the interaction fusion block in [20], (e) a fused signal using the proposed cross-correlation block with A_{PSM}, and (f) a speech signal estimated be the CROCS ($\hat{s}_2$). The fused signals consist of essentially a linear combination of the frames in the enrollment speech to match the mixture signal. Comparing Figure 5d,e, the fused signal using the proposed cross-correlation block showed a clearer harmonic structure and resembles the target speech more than the fused signal with the interaction fusion block which treated the real and imaginary parts separately. It is noted that essentially the fused feature in a frame is a linear combination of the most similar spectra among the enrollment frames, and thus was not silent in interference-only regions such as the region near 2.0 s. As the fused features X were used instead of E as auxiliary information, it did not degrade the output signal, as can be seen in Figure 5f.

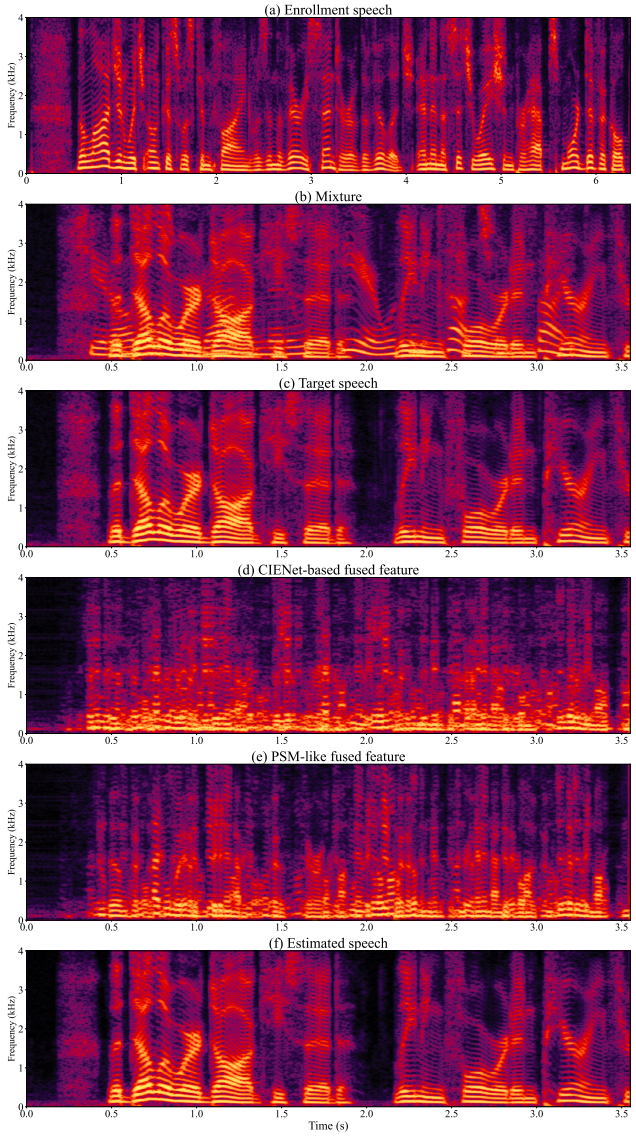

Figure 5. Spectrograms of the (**a**) enrollment speech, (**b**) mixture, (**c**) target speech, (**d**) a fused feature using the interaction fusion block in [20], (**e**) a fused feature X using A_{PSM}, and (**f**) a speech signal estimated be the CROCS ($\hat{s}_2$). The A_{PSM}-based fused feature in (**e**) exhibits clearer harmonic structures than the CIENet-based fused feature in (**d**), making it similar to the target speech in (**c**).

4.2. Comparison with Other TSE Systems

Table 2 summarizes the performance on the Libri2Mix dataset sampled at 8 kHz without noise. On top of the performance reported in the papers, we included the results for the SpEx+ [4] re-trained using the authors' code on GitHub (https://github.com/gemengtju/SpEx_Plus, accessed on 17 June 2026) and the re-implementation of the CIENet-C2F-mDPTNet [20] in the table. We have also re-trained the X-TF-GridNet [12] with the official GitHub code (https://github.com/HaoFengyuan/X-TF-GridNet, accessed on 17 June 2026) for a sample-wise analysis, and obtained similar performance to that in the paper, as can be seen in Table 2. The CIENet-C2F-mDPTNet [20] with 29.03 GMAC/s outperformed all the previously proposed methods except the X-TF-GridNet requiring 68.32 GMAC/s. The X-TF-GridNet showed the best performance among the existing methods in all three measures with large margins, and the proposed CROCS marked even higher average performances with the PESQ of 3.60, the ESTOI of 89.92%, and the SI-SDR of 17.67 dB on the Libri2Mix dataset at 8 kHz. Compared with X-TF-GridNet, CROCS reduced the number of parameters and GMAC/s by approximately 10.5% and 17.0%, respectively, while achieving statistically significant performance improvements. Specifically, CROCS improved SI-SDR by 1.413 dB with a 95% confidence interval (CI) of [1.186, 1.640] and PESQ by 0.080 with a 95% CI of [0.068, 0.092]. In addition, CROCS improved ESTOI by 2.98%. Moreover, the ratios of the test samples with the negative SI-SDR which would correspond to the speaker confusion were 4.65% for the X-TF-GridNet and 2.08% for the proposed CROCS. These results imply that the speaker confusion, which can be interpreted as a complete failure of the TSE, happened more rarely for the proposed CROCS compared with the X-TF-GridNet.

Table 2. Performance comparison with previous TSE approaches on the Libri2Mix-clean dataset sampled at 8 kHz with the *min* mode.

Method	# Params. (M)	GMAC/s	PESQ	ESTOI (%)	SI-SDR (dB)
Mixture	-	-	1.60	53.84	0.0005
sDPCCN (2022) [6]	-	-	2.73	78.90	11.65
TD-SpeakerBeam (2020) [2]	-	-	2.75	-	12.86
TD-SpeakerBeam + PL + PF (2022) [7]	-	-	2.86	-	13.88
Diff-TSE-MT (2023) [9]	-	-	3.08	80.00	11.28
SpEx+ † (2020) [4]	11.18	3.76	2.58	78.59	11.66
MC-SpEx (2023) [8]	-	-	3.19	84.90	14.61
CIENet-C2F-mDPTNet ‡ (2024) [20]	3.01	29.03	3.21	84.85	14.78
X-TF-GridNet (2024) [12]	7.79	68.32	3.53	86.94	16.20
X-TF-GridNet †	7.79	68.32	3.51	87.07	16.25
CROCS	6.97	56.70	**3.60**	**89.92**	**17.67**

† denotes the re-training results using the official Github code. ‡ denotes the re-implemented results obtained under the same dataset split, enrollment selection protocol, input sampling rate, optimization procedure, and evaluation metrics as CROCS. Bold values indicate the best performance in each column.

We have further analyzed the performance of the proposed CROCS and the re-trained models for the SpEx+ [4] and X-TF-GridNet [12] in Table 2, depending on the genders of the two speakers in each mixture. The performances for the same-gender mixtures and the different-gender mixtures are shown in Table 3. Compared with X-TF-GridNet, CROCS achieved relative improvements of 4.1%, 10.1%, and 15.5% in PESQ, SDR, and SI-SDR, respectively, for the same gender mixtures, while the corresponding relative improvements for the different gender mixtures were 0.8%, 2.2%, and 3.1%. It can be clearly seen that the performance gap between the same-gender mixture and the different-gender mixtures was significantly reduced for the proposed CROCS compared with the previous methods,

which implies that the performance for the harder cases was improved even more than the average performance improvement.

Table 3. Performances depending on whether two speakers in a mixture were of the same or different genders on the Libri2Mix-clean dataset sampled at 8 kHz with the *min* mode.

Method	PESQ		SDR (dB)		SI-SDR (dB)	
	Same	Different	Same	Different	Same	Different
Mixture	1.60	1.59	0.1843	0.1742	0.0034	−0.0023
SpEx+ [4]	2.49	2.67	11.39	13.52	10.53	12.78
X-TF-GridNet [12]	3.42	3.61	16.31	18.49	14.82	17.67
CROCS	**3.56**	**3.64**	**17.95**	**18.90**	**17.12**	**18.21**

Bold values indicate the best performance in each column.

The experiential results on the Libri2Mix dataset without noise sampled at 16 kHz are shown in Table 4. Most of the previous works did not report the PESQ scores and the ESTOI and focused on the SDR and SI-SDR. The proposed CROCS achieved relative gains of 30.6%, 15.5%, and 13.3% over the best previously reported results in PESQ-WB, SDR, and SI-SDR, respectively.

Table 4. Performance comparison with previous TSE approaches on the Libri2Mix-clean dataset sampled at 16 kHz with the *min* mode

Method	PESQ- NB	PESQ- WB	ESTOI (%)	SDR (dB)	SI-SDR (dB)
Mixture	1.50	1.15	53.74	0.0011	0.0006
TD-SpeakerBeam (2020) [2,14]	-	-	-	13.69	13.03
Li et al. (2024) [15]	-	-	-	15.23	14.57
SSL-TD-SpeakerBeam (2024) [14]	-	2.45	-	15.26	14.65
SHuBERT-BSRNN (2024) [16]	-	-	-	16.07	15.39
ECAPA-BSRNN (2025) [27]	-	-	-	-	15.91
CROCS	**3.64**	**3.20**	**90.88**	**18.56**	**18.02**

Bold values indicate the best performance in each column.

We have also evaluated the performance of the CROCS and the re-trained versions of the SpEx+ [4] and the X-TF-GridNet [12] on the Libri2Mix-noisy dataset consisting of mixtures of two speeches and noise sampled at 8 kHz. The results are shown in Table 5, in which we can verify that the proposed CROCS demonstrated a stronger performance than the compared methods. Specifically, CROCS achieved relative improvements of 3.5%, 6.8%, and 19.5% over X-TF-GridNet in PESQ, ESTOI, and SI-SDR, respectively.

Table 5. Performance with previous TSE systems on the Libri2Mix-noisy dataset sampled at 8 kHz with the *min* mode.

Training Set	Method	PESQ	ESTOI (%)	SI-SDR (dB)
Libri2Mix Noisy dataset	Mixture	1.43	40.05	−1.99
	SpEx+ † [4]	2.07	64.27	7.92
	X-TF-GridNet † [12]	2.56	71.89	9.80
	CROCS	**2.65**	**76.75**	**11.71**

† denote the re-training results using the official Github code. Bold values indicate the best performance in each column.

To further evaluate the performance of the proposed CROCS under noisy and reverberant environments with a wider variety of previously proposed methods, we carried out an experiment on the WHAMR! dataset [59]. Table 6 presents the performance of recent

TSE models in terms of the PESQ scores, SDR, and SI-SDR, where the best and second-best results are highlighted in bold and underlined, respectively. In this experiment, we set the window length and a hop size for the CROCS to 16 ms and 8 ms, respectively, as in the USEF-TFGridNet [21], MCFS-TFGridNet [26], and LExt-TFGridNetV2 [24]. For this dataset, the CROCS showed the second-best performance with much lower computational complexity than the best model, LExt-TFGridNetV2 [24], although the computational complexity of the CROCS was much higher than that of the models that have shown slightly lower performance, the CIENet-C2F-mDPTNet [20] and MCFS-TFGridNet [26].

Table 6. Performance comparison with previous TSE approaches on the WHAMR! dataset sampled at 8 kHz with the *min* mode.

Method	GMAC/s	PESQ	SDR (dB)	SI-SDR (dB)
Mixture	-	1.41	−3.5	−6.1
X-TF-GridNet (2024) [12]	68.32	-	10.3	8.5
CIENet-C2F-mDPRNN (2024) [20]	22.96	2.60	11.7	10.6
CIENet-C2F-mDPTNet (2024) [20]	25.88	2.72	12.5	11.4
USEF-Sepformer (2025) [21]	45.80	-	6.8	5.1
USEF-TFGridNet (2025) [21]	125.30	-	11.4	10.0
DCF-Net (2025) [22]	20.49	-	11.0	9.7
MCFS-TFGridNet (2026) [26]	22.16	2.79	12.2	11.0
LExt-TFGridNetV2 (2025) [24]	183.25 †	**2.94**	**13.2**	**12.2**
CROCS ‡	72.73	<u>2.80</u>	<u>12.6</u>	<u>11.5</u>

† denote the GMAC/S calculated using the configuration settings in [24]. ‡ denote the results with a window size of 16 ms and a hop size of 8 ms. Bold values indicate the best performance in each column, and underlined values indicate the second-best performance in each column.

5. Limitations

Although CROCS showed good performance on the Libri2Mix and the WHAMR! datasets, several limitations remain for practical deployment. Most experiments assume relatively clean enrollment conditions, whereas real-world enrollment utterances may be affected by background noise, reverberation, channel distortion, or short speech duration, which can degrade the performance of TSE systems. Furthermore, CROCS showed a larger performance gap over the previous TSE methods under clean mixture conditions, whereas the performance gap became smaller on the WHAMR! dataset, which includes noisy and reverberant mixture conditions.

6. Conclusions

In this paper, we propose a two-stage target speaker extraction system we call CROCS with a novel cross-correlation-based fusion block and a dual post-refinement structure. In the first stage, the cross-correlation between the complex spectra for the enrollment and mixture was used to obtain attention weights for the enrollment in an analogous form to the phase-sensitive mask. Then, the fused features and the mixture features are processed by a speech extraction network to produce a coarsely extracted target speech. The second stage consists of two parallel branches, one of which processes the first-stage output in the same manner to the cross-correlation-based fusion in the first stage and the other takes the concatenation of the mixture and the first-stage output as input to complement possibly attenuated target speech. In addition, the low-dimensional speaker embeddings extracted from the enrollment and the speech estimated in the first stage are incorporated in the speech extraction networks in the second stage through adapter modules to exploit the speaker discriminability. Experimental results showed that the proposed CROCS achieved superior performance to the existing TSE methods on the Libri2Mix datasets without and

with noise, especially for the same gender mixtures. The CROCS was also shown to be competitive with recently proposed TSE systems for the WHAMR! dataset with noise and reverberations.

Author Contributions: Conceptualization, S.H. and J.W.S.; methodology, S.H., S.L. and J.W.S.; validation, S.H. and J.W.S.; formal analysis, S.H., S.L. and J.W.S.; investigation, S.H.; resources, S.H.; data curation, S.H.; writing—original draft preparation, S.H.; writing—review and editing, S.H., S.L. and J.W.S.; supervision, J.W.S.; project administration, J.W.S.; funding acquisition, J.W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by the IITP (Institute of Information & Coummunications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2026-RS-2021-II211835, 50%) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. IRIS-2025-25399269, 50%).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets mentioned in this paper can be downloaded using the following links: LibriMix https://github.com/JorisCos/LibriMix (accessed on 17 June 2026), WHAMR! http://wham.whisper.ai/WHAMR_README.html (accessed on 17 June 2026), and DNS Challenge https://github.com/microsoft/DNS-Challenge (accessed on 17 June 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zmolikova, K.; Delcroix, M.; Ochiai, T.; Kinoshita, K.; Černockỳ, J.; Yu, D. Neural target speech extraction: An overview. *IEEE Signal Process. Mag.* **2023**, *40*, 8–29. [CrossRef]
2. Delcroix, M.; Ochiai, T.; Zmolikova, K.; Kinoshita, K.; Tawara, N.; Nakatani, T.; Araki, S. Improving speaker discrimination of target speech extraction with time-domain speakerbeam. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 691–695.
3. Xu, C.; Rao, W.; Chng, E.S.; Li, H. Spex: Multi-scale time domain speaker extraction network. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1370–1384. [CrossRef]
4. Ge, M.; Xu, C.; Wang, L.; Chng, E.S.; Dang, J.; Li, H. Spex+: A complete time domain speaker extraction network. In *Proceedings Interspeech*; Curran Associates, Inc.: Red Hook, NY, USA, 2020.
5. Wang, W.; Xu, C.; Ge, M.; Li, H. Neural Speaker Extraction with Speaker-Speech Cross-Attention Network. In *Proceedings Interspeech*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; pp. 3535–3539.
6. Han, J.; Long, Y.; Burget, L.; Černockỳ, J. DPCCN: Densely-connected pyramid complex convolutional network for robust speech separation and extraction. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 22–27 May 2022; pp. 7292–7296.
7. Zhao, Z.; Yang, D.; Gu, R.; Zhang, H.; Zou, Y. Target confusion in end-to-end speaker extraction: Analysis and approaches. In *Proceedings Interspeech*; Curran Associates, Inc.: Red Hook, NY, USA, 2022.
8. Chen, J.; Rao, W.; Wang, Z.; Lin, J.; Ju, Y.; He, S.; Wang, Y.; Wu, Z. Mc-spex: Towards effective speaker extraction with multi-scale interfusion and conditional speaker modulation. In *Proceedings Interspeech*; Curran Associates, Inc.: Red Hook, NY, USA, 2023.
9. Kamo, N.; Delcroix, M.; Nakatani, T. Target speech extraction with conditional diffusion model. In *Proceedings Interspeech*; Curran Associates, Inc.: Red Hook, NY, USA, 2023.
10. Liu, K.; Du, Z.; Wan, X.; Zhou, H. X-sepformer: End-to-end speaker extraction network with explicit optimization on speaker confusion. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
11. Mu, Z.; Yang, X.; Sun, S.; Yang, Q. Self-supervised disentangled representation learning for robust target speech extraction. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; pp. 18815–18823.
12. Hao, F.; Li, X.; Zheng, C. X-tf-gridnet: A time–Frequency domain target speaker extraction network with adaptive speaker embedding fusion. *Inf. Fusion* **2024**, *112*, 102550.

13. Ju, Y.; Chen, J.; Zhang, S.; He, S.; Rao, W.; Zhu, W.; Wang, Y.; Yu, T.; Shang, S. Tea-pse 3.0: Tencent-ethereal-audio-lab personalized speech enhancement system for icassp 2023 dns-challenge. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 4–10 June 2023; pp. 1–2.
14. Peng, J.; Delcroix, M.; Ochiai, T.; Plchot, O.; Araki, S.; Černockỳ, J. Target speech extraction with pre-trained self-supervised learning models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Seoul, Republic of Korea, 14–19 April 2024; pp. 10421–10425.
15. Li, J.; Zhang, K.; Wang, S.; Li, H.; Mak, M.W.; Lee, K.A. On the effectiveness of enrollment speech augmentation for target speaker extraction. In Proceedings of the IEEE Spoken Language Technology Workshop, Macao, China, 2–5 December 2024; pp. 325–332.
16. Lin, J.; Ge, M.; Wang, W.; Li, H.; Feng, M. Selective HuBERT: Self-supervised pre-training for target speaker in clean and mixture speech. *IEEE Signal Process. Lett.* **2024**, *31*, 1014–1018. [CrossRef]
17. Zeng, B.; Suo, H.; Wan, Y.; Li, M. Sef-net: Speaker embedding free target speaker extraction network. In *Proceedings Interspeech*; Curran Associates, Inc.: Red Hook, NY, USA, 2023; pp. 3452–3456.
18. Hu, Y.; Xu, H.; Guo, Z.; Huang, H.; He, L. SMMA-Net: An Audio Clue-Based Target Speaker Extraction Network with Spectrogram Matching and Mutual Attention. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Seoul, Republic of Korea, 14–19 April 2024; pp. 1496–1500.
19. Yang, X.; Bao, C.; Zhou, J.; Chen, X. Target Speaker Extraction by Directly Exploiting Contextual Information in the Time-Frequency Domain. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Seoul, Republic of Korea, 14–19 April 2024; pp. 10476–10480.
20. Yang, X.; Bao, C.; Chen, X. Coarse-to-fine target speaker extraction based on contextual information exploitation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 3795–3810. [CrossRef]
21. Zeng, B.; Li, M. Usef-tse: Universal speaker embedding free target speaker extraction. *IEEE Trans. Audio Speech Lang. Process.* **2025**, *33*, 2110–2124. [CrossRef]
22. Xue, K.; Fan, R.; Sun, C.; Zhao, P.; An, J. DCF-Net: Efficient Target Speaker Extraction by Leveraging Mixture and Enrollment Interactions. *IEEE Signal Process. Lett.* **2025**, *32*, 3240–3244. [CrossRef]
23. Gu, R.; Zhang, S.X.; Xu, Y.; Chen, L.; Zou, Y.; Yu, D. Multi-modal multi-channel target speech separation. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 530–541. [CrossRef]
24. Shen, P.; Chen, K.; He, S.; Chen, P.; Yuan, S.; Kong, H.; Zhang, X.; Wang, Z.Q. Listen to Extract: Onset-Prompted Target Speaker Extraction. *IEEE Trans. Audio Speech Lang. Process.* **2025**, *33*, 4832–4843. [CrossRef]
25. Yang, L.; Liu, W.; Tan, L.; Yang, J.; Moon, H.G. Target speaker extraction with ultra-short reference speech by ve-ve framework. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 4–9 June 2023; pp. 1–5.
26. Kim, H.; Shin, J.W. Target Speaker Extraction using Multi-Stage Cross-Attention and Frequency-wise State Initialization. *IEEE Signal Process. Lett.* **2026,** *33*, 773–777. [CrossRef]
27. Zhang, K.; Li, J.; Wang, S.; Wei, Y.; Wang, Y.; Wang, Y.; Li, H. Multi-level speaker representation for target speaker extraction. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Hyderabad, India, 6–11 April 2025; pp. 1–5.
28. Gu, R.; Chen, L.; Zhang, S.X.; Zheng, J.; Xu, Y.; Yu, M.; Su, D.; Zou, Y.; Yu, D. Neural Spatial Filter: Target Speaker Speech Separation Assisted with Directional Information. In *Proceedings Interspeech*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 4290–4294.
29. Ge, M.; Xu, C.; Wang, L.; Chng, E.S.; Dang, J.; Li, H. L-spex: Localized target speaker extraction. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 22–27 May 2022; pp. 7287–7291.
30. Luo, Y.; Mesgarani, N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1256–1266. [CrossRef] [PubMed]
31. Luo, Y.; Chen, Z.; Yoshioka, T. Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 46–50.
32. Chen, J.; Mao, Q.; Liu, D. Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation. In *Proceedings Interspeech*; Curran Associates, Inc.: Red Hook, NY, USA, 2020.
33. Subakan, C.; Ravanelli, M.; Cornell, S.; Bronzi, M.; Zhong, J. Attention is all you need in speech separation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, 6–11 June 2021; pp. 21–25.
34. Wang, Z.Q.; Cornell, S.; Choi, S.; Lee, Y.; Kim, B.Y.; Watanabe, S. TF-GridNet: Integrating full-and sub-band modeling for speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 3221–3236. [CrossRef]
35. Byun, J.; Shin, J.W. Monaural speech separation using speaker embedding from preliminary separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 2753–2763. [CrossRef]

36. Kim, H.; Shin, J.W. On Training Speech Separation Models with Various Numbers of Speakers. *IEEE Signal Process. Lett.* **2023**, *30*, 1202–1206. [CrossRef]
37. Lu, Y.X.; Ai, Y.; Ling, Z.H. MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra. In *Proceedings Interspeech*; Curran Associates, Inc.: Red Hook, NY, USA, 2023.
38. Lu, Y.X.; Ai, Y.; Ling, Z.H. Explicit estimation of magnitude and phase spectra in parallel for high-quality speech enhancement. *Neural Netw.* **2025**, *189*, 107562. [CrossRef] [PubMed]
39. Kim, M.; Shin, J.W. Improved speech enhancement considering speech PSD uncertainty. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 1939–1951. [CrossRef]
40. Hwang, S.; Kim, M.; Shin, J.W. Dual microphone speech enhancement based on statistical modeling of interchannel phase difference. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2865–2874. [CrossRef]
41. Richter, J.; Welker, S.; Lemercier, J.M.; Lay, B.; Gerkmann, T. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 2351–2364. [CrossRef]
42. Lemercier, J.M.; Richter, J.; Welker, S.; Gerkmann, T. Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 2724–2737. [CrossRef]
43. Lee, S.; Cheong, S.; Han, S.; Shin, J.W. FlowSE: Flow Matching-based Speech Enhancement. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Hyderabad, India, 6–11 April 2025; pp. 1–5.
44. Cheong, S.; Kim, M.; Shin, J.W. Integrated DNN-Based Parameter Estimation for Multichannel Speech Enhancement. *IEEE Signal Process. Lett.* **2025**, *32*, 3320–3324. [CrossRef]
45. Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. X-vectors: Robust dnn embeddings for speaker recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, 15–20 April 2018; pp. 5329–5333.
46. Desplanques, B.; Thienpondt, J.; Demuynck, K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proceedings Interspeech*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; pp. 3830–3834.
47. Chung, J.S.; Huh, J.; Mun, S.; Lee, M.; Heo, H.S.; Choe, S.; Ham, C.; Jung, S.; Lee, B.J.; Han, I. In Defence of Metric Learning for Speaker Recognition. In *Proceedings Interspeech*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; pp. 2977–2981.
48. Han, S.; Ahn, Y.; Kang, K.; Shin, J.W. Short-segment speaker verification using ecapa-tdnn with multi-resolution encoder. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 4–9 June 2023; pp. 1–5.
49. Han, S.; Ahn, Y.; Shin, J.W. Noise-Robust Speaker Verification with Attenuated Speech Restoration and Consistency Training. *IEEE Trans. Audio Speech Lang. Process.* **2025**, *33*, 1987–1997. [CrossRef]
50. Erdogan, H.; Hershey, J.R.; Watanabe, S.; Le Roux, J. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, South Brisbane, Australia, 19–24 April 2015; pp. 708–712.
51. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Salt Lake City, UT, USA, 7–11 May 2001; pp. 749–752.
52. Jensen, J.; Taal, C.H. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 2009–2022. [CrossRef]
53. Le Roux, J.; Wisdom, S.; Erdogan, H.; Hershey, J.R. SDR–half-baked or well done? In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019; pp. 626–630.
54. Waswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
55. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 16133–16142.
56. Cosentino, J.; Pariente, M.; Cornell, S.; Deleforge, A.; Vincent, E. Librimix: An open-source dataset for generalizable speech separation. In *Proceedings Interspeech*; Curran Associates, Inc.: Red Hook, NY, USA, 2020.
57. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, South Brisbane, Australia, 19–24 April 2015; pp. 5206–5210.
58. Wichern, G.; Antognini, J.; Flynn, M.; Zhu, L.R.; McQuinn, E.; Crow, D.; Manilow, E.; Roux, J.L. Wham!: Extending speech separation to noisy environments. *arXiv* **2019**, arXiv:1907.01160.
59. Maciejewski, M.; Wichern, G.; McQuinn, E.; Le Roux, J. WHAMR!: Noisy and reverberant single-channel speech separation. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 696–700.

60. Hershey, J.R.; Chen, Z.; Le Roux, J.; Watanabe, S. Deep clustering: Discriminative embeddings for segmentation and separation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 20–25 March 2016; pp. 31–35.
61. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
62. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv* **2019**, arXiv:1912.01703.
63. Rec, I. P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs. *Int. Telecommun. Union* **2005**, *41*, 48–60.
64. Vincent, E.; Gribonval, R.; Févotte, C. Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1462–1469. [CrossRef]