

Discriminative–Generative Target Speaker Extraction with Decoder-Only Language Models

Bang Zeng, Student Member, IEEE, Beilong Tang, Wang Xiang, Ming Li*, Senior Member, IEEE

Abstract—Target speaker extraction (TSE) aims to recover the speech signal of a desired speaker from a mixed audio recording, given a short enrollment utterance. Most existing TSE approaches are based on discriminative modeling paradigms. Although effective at suppressing interfering speakers, these methods often struggle to produce speech with high perceptual quality and naturalness. To address this limitation, we first propose **LauraTSE**, a generative TSE model built upon an auto-regressive decoder-only language model. However, purely generative approaches may suffer from hallucinations, content drift, and limited controllability, which may undermine their reliability in complex acoustic scenarios. To overcome these challenges, we further introduce a discriminative–generative TSE framework. In this framework, a discriminative front-end is employed to robustly extract the target speaker’s speech, yielding stable and controllable intermediate representations. A generative back-end then operates in the neural audio codec representation space to reconstruct fine-grained speech details and enhance perceptual quality. This two-stage design effectively combines the robustness and controllability of discriminative models with the superior naturalness and quality enhancement capabilities of generative models. Moreover, we systematically investigate collaborative training strategies for the proposed framework, including freezing or fine-tuning the front-end, incorporating an auxiliary SI-SDR loss, and exploring both auto-regressive and non-auto-regressive inference mechanisms. Experimental results demonstrate that the proposed framework achieves a more favorable trade-off among speech quality, intelligibility, and speaker consistency.

Index Terms—Target speaker extraction, Auto-regressive decoder-only language model, Discriminative–generative, Speech quality, Intelligibility.

I. INTRODUCTION

HUMANS are capable of selectively attending to a target speech signal in complex acoustic environments, a phenomenon known as the cocktail party effect [1], [2]. This remarkable ability has inspired extensive research on speech separation. Early approaches, such as non-negative matrix factorization (NMF) [3], [4] and computational auditory scene analysis (CASA) [5]–[7], primarily rely on spectro-temporal masking strategies and are known to degrade in highly complex acoustic conditions. With the advent of deep learning, neural network-based methods, including deep clustering [8]–[10], deep attractor networks (DANet) [11]–[13], and permutation invariant training (PIT) [14], [15], have substantially

Bang Zeng, Wang Xiang and Ming Li are with the School of Computer Science, Wuhan University, Wuhan 430072, China, and also with Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Digital Innovation Research Center, Duke Kunshan University, Kunshan 215316, China. Beilong Tang is with the North Carolina State University (e-mail: bangzeng@whu.edu.cn; btang5@ncsu.edu; 2025102110031@whu.edu.cn; ming.li369@dukekunshan.edu.cn).

* Corresponding author.

improved speech separation performance. In parallel, time-domain models such as TasNet [16] and its variants further enhance perceptual quality by enabling more accurate phase reconstruction. More recently, advanced architectures in both time and time–frequency domains have continued to push the performance boundaries of speech separation. Nevertheless, most existing speech separation methods [17]–[28] aim to separate all speakers in a mixture and typically require prior knowledge of the number of sources. These assumptions are often impractical in real-world scenarios.

In contrast, target speaker extraction (TSE) [29]–[41] focuses on extracting a desired speaker from a mixture using auxiliary speaker information, offering a more flexible and application-oriented solution. Recently, TSE has emerged as an effective paradigm to address the limitations of conventional speech separation, particularly in scenarios where the number of speakers in a mixture is unknown. By leveraging reference speech, TSE models aim to extract only the target speaker from complex acoustic mixtures, making them more suitable for real-world applications. As illustrated in Fig. 1, a typical TSE framework follows an encoder–separator–decoder architecture, which can be implemented either in the time–frequency domain using STFT/iSTFT or directly in the time domain via convolutional operations. Most existing TSE approaches employ a speaker embedding extractor to derive a compact representation of the target speaker from the reference utterance, which is then used to guide the separation process. However, such embedding extractors are commonly optimized for speaker recognition rather than TSE, and may discard fine-grained information contained in the reference speech. This mismatch can limit the effectiveness of embedding-based TSE methods. Consequently, speaker-embedding-free TSE approaches [42]–[45] have been proposed to exploit reference speech representations directly, enabling more precise and efficient target speaker extraction.

However, most existing TSE methods adopt discriminative modeling paradigms, which directly learn a deterministic mapping from mixture and reference information to the target signal. While such models exhibit strong robustness and controllability in suppressing interfering speakers, they inherently suffer from several limitations. First, discriminative models are typically optimized using signal-level distortion metrics that poorly align with human auditory perception, leading to limited speech naturalness and perceptual quality. Second, these models have limited capability to recover fine-grained speech details that are missing or distorted during the separation process [46]. In contrast, generative models adopt a probabilistic modeling perspective and aim to learn the joint

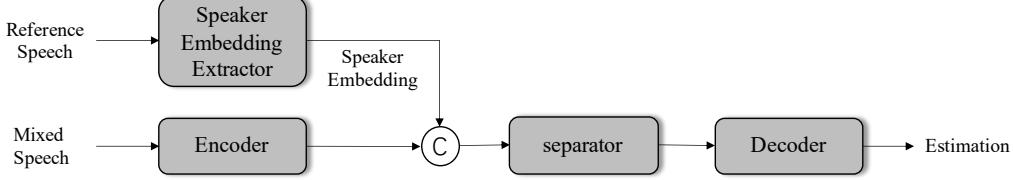


Fig. 1. The diagram of a typical target speaker extraction method. The speaker embedding extractor is typically a pre-trained speaker recognition model. 'C' denotes the concatenation.

or conditional distributions among the mixed speech, the clean target speech, and the enrollment information. By explicitly modeling the underlying generative process of target speech, generative approaches can produce multiple plausible speech estimates under the same input conditions, rather than being constrained to a single deterministic solution. This property often results in improved perceptual quality compared to discriminative methods [47]–[50].

In recent years, various generative frameworks, including diffusion models [47] and variational autoencoders (VAEs) [51] have been increasingly explored for the TSE task. Among these studies, TSELM [52] leverages discrete semantic units extracted by WavLM [53] and employs an encoder-only language model for target speaker extraction. AnyEnhance [49] adopts a masking-based language model and constructs a unified framework that supports multiple speech processing tasks, including TSE. These works provide preliminary evidence of the potential of generative modeling for TSE. Nevertheless, as an important class of generative models, auto-regressive (AR) decoder-only language models remain relatively underexplored in the context of TSE. Although SpeechX [54] develops a multi-task speech processing system based on an AR decoder-only LM, it does not directly address a key question: in a single-task TSE setting, can a compact AR decoder-only language model provide sufficient modeling capacity for effective target speaker extraction?

We recently proposed LauraTSE [55], a generative TSE model based on an auto-regressive (AR) decoder-only language model. LauraTSE comprises a compact AR decoder-only LM that predicts coarse-grained target speech representations conditioned on continuous representations of the mixed and reference speech, together with a lightweight encoder-only LM designed to recover fine-grained acoustic details. Extensive experimental results demonstrate that LauraTSE can produce speech with improved perceptual quality. Nevertheless, generative models are often sensitive to the design and discretization of input representations. When discrete representations fail to preserve fine-grained, speaker-related acoustic characteristics, even models with strong generative capacity may struggle to reconstruct the target speaker's speech accurately. Moreover, LauraTSE remains prone to errors such as hallucinations, which raise concerns regarding model stability and reliability.

To address these issues, this study proposes a discriminative–generative two-stage framework for TSE. In the first stage, a discriminative module is employed to robustly extract target speaker–related information while effectively suppress-

ing interfering sources, leveraging its strong discrimination capability. In the second stage, a generative module is introduced to perform distribution-level modeling and high-quality reconstruction based on the discriminative front-end's output, thereby further enhancing speech quality. By integrating the complementary strengths of discriminative and generative paradigms, the proposed two-stage framework provides a more robust and effective solution for improving the perceptual quality of target speaker extraction.

This work extends our previous study on LauraTSE [55]. The main contributions of this article are summarized as follows:

- We develop an AR decoder-only language model, LauraTSE [55], for the TSE task. By leveraging continuous acoustic features and neural audio codec representations as a bridging interface, LauraTSE enables end-to-end generative modeling for TSE. Extensive experimental results across multiple objective metrics demonstrate that the proposed approach achieves improved speech quality and intelligibility compared with conventional discriminative methods.
- We propose a discriminative–generative two-stage TSE framework, in which USEF-TFGridNet [45] serves as the discriminative front-end and LauraTSE [55] acts as the generative back-end, forming a complete system termed USEF-Laura-TSE. Through comprehensive experimental analysis, we investigate the impact of the discriminative front-end on the reconstruction quality of the generative back-end, and the feedback of the generative back-end in suppressing residual interference and artifacts introduced by the front-end, thereby revealing the interdependence between the two stages.
- Building upon the proposed discriminative–generative architecture, we further investigate both auto-regressive and non-auto-regressive inference strategies. Without modifying the training procedure, a non-autoregressive inference scheme is introduced by treating the discriminative front-end outputs as pseudo-labels, enabling a more favorable trade-off between speech quality and intelligibility.

II. RELATED WORKS

A. Discriminative Approaches for Target Speaker Extraction

Discriminative target speaker extraction (TSE) methods have achieved substantial progress in recent years and can generally be categorized into time-domain and time-frequency (T-F) domain approaches. Early T-F domain methods estimate speaker-dependent masks on short-time Fourier transform

(STFT) representations. In contrast, time-domain architectures operate directly on raw waveforms, avoiding explicit phase reconstruction and thereby improving perceptual quality. Representative models such as TasNet [16] and Conv-TasNet [19] employ convolutional encoder–decoder structures to learn waveform-level representations. More advanced architectures, including DPRNN [17], SepFormer [23], and transformer-based models [22], further enhance extraction performance by explicitly modeling long-range temporal dependencies and global contextual information. For the TSE task, target speaker information is typically incorporated through speaker embeddings extracted from reference speech, which are used to guide the extraction process [29], [30], [56]. In such embedding-based frameworks, speaker encoders [57]–[60] are integrated with separation networks via feature concatenation, conditioning, or attention mechanisms, and various architectural designs have been proposed to improve robustness in feature extraction and cross-stream fusion. More recently, speaker-embedding-free TSE approaches [42]–[44], [61], [62] have been introduced to bypass fixed-dimensional speaker embeddings. Instead, these methods directly exploit frame-level acoustic features and model the contextual interactions between reference and mixed speech using attention-based mechanisms. By preserving fine-grained temporal and spectral information, speaker-embedding-free approaches can effectively mitigate the information loss and representation mismatch commonly introduced by speaker embeddings [63], [64].

Despite these advances, most discriminative TSE models rely on deterministic mappings optimized with signal-level distortion objectives, which limits their ability to capture the intrinsic uncertainty and multimodality of speech signals. Consequently, they often struggle to recover fine-grained speech details and achieve high perceptual naturalness, particularly under challenging acoustic conditions. These limitations motivate the exploration of generative modeling approaches, which offer greater flexibility in modeling speech distributions and provide new opportunities for perceptual quality enhancement in target speaker extraction.

B. Generative Approaches for Target Speaker Extraction

Generative approaches for target speaker extraction (TSE) can be broadly categorized into continuous and discrete modeling paradigms. Continuous generative models, such as diffusion models [65]–[70] and variational autoencoders (VAEs) [71], [72], directly model the probability distribution of speech signals and generate target speech through iterative denoising or latent-variable reconstruction. These methods exhibit strong modeling capacity and high reconstruction fidelity. However, their substantial computational cost and inference latency often limit practical deployment, particularly in real-time and edge-device scenarios. More recently, increasing attention has been directed toward discrete representation-based generative approaches leveraging large language models (LLMs) [73]–[75]. In such frameworks, speech signals are first converted into discrete token sequences using neural audio codecs, then generated conditionally with LLMs, and finally reconstructed into waveforms via codec decoders. Benefiting

from powerful contextual modeling and sequence generation capabilities, LLM-based approaches have demonstrated promising performance across various speech processing tasks, including speech enhancement, separation, and target speaker extraction. Among different LLM architectures, decoder-only models are particularly well-suited for generative speech modeling due to their auto-regressive formulation and their flexibility for multi-task and multi-modal extensions. Representative systems show that combining codec-based discrete representations with decoder-only LLMs can substantially improve perceptual speech quality and robustness [54], [76], [77].

Despite these advantages, LLM-based generative TSE approaches still face several challenges. The reliance on discrete token prediction may lead to error accumulation and stability issues, while the large model size and associated computational overhead limit inference efficiency. Moreover, purely generative reconstruction does not always guarantee stable and reliable performance under diverse acoustic conditions, particularly when the input representations fail to preserve fine-grained, speaker-related information. These observations motivate the exploration of discriminative-generative two-stage frameworks for target speaker extraction. In such frameworks, a discriminative front-end provides reliable target alignment and effective interference suppression, while a generative back-end performs distribution-level modeling to enhance speech quality further. This two-stage design offers a practical compromise by combining the stability and efficiency of discriminative models with the perceptual advantages of generative modeling, thereby enabling more robust and high-quality target speaker extraction.

III. DISCRIMINATIVE-GENERATIVE TARGET SPEAKER EXTRACTION

In this section, we first introduce LauraTSE in detail. We then present the proposed discriminative–generative two-stage framework, followed by a comprehensive description of its architecture and design principles. Finally, to validate the effectiveness of the two-stage framework, we construct a complete system, USEF-Laura-TSE, that employs USEF-TFGridNet [45] as the discriminative front-end and LauraTSE as the generative back-end.

A. LauraTSE

In this study, we propose LauraTSE, a target-speaker extraction method based on an auto-regressive (AR) decoder-only language model built on the LauraGPT [77] backbone. LauraTSE takes the log-mel spectrogram features of both the target speaker’s enrollment speech and the mixed speech as inputs, and employs the residual vector quantization (RVQ) layers of a neural audio codec to discretize audio representations, enabling high-quality modeling and reconstruction of the target speaker’s speech. The overall architecture of LauraTSE is illustrated in Fig 2. LauraTSE consists of two key components. The first is an AR decoder-only language model that predicts the discrete representations of the target speech corresponding to the first several codec encoding layers. The second is a one-step encoder-only language model that jointly

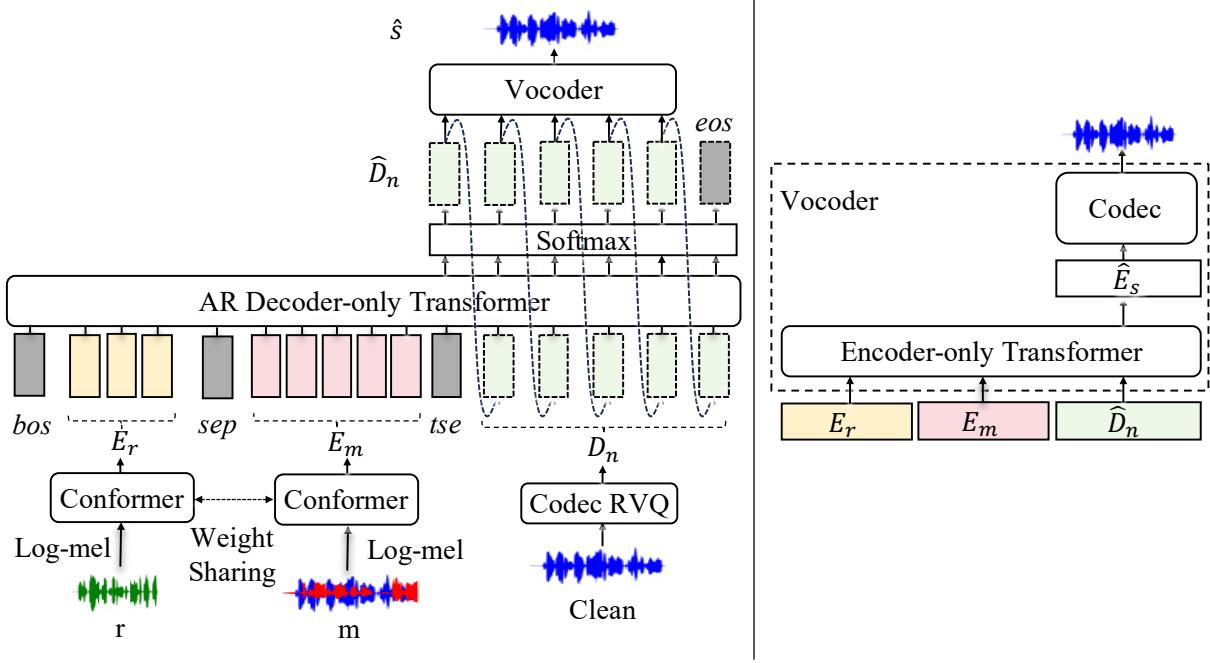


Fig. 2. The diagram of LauraTSE network. ‘m’ and ‘r’ denote the mixed speech and reference speech, respectively. We use two weight sharing conformer to process the mixed and reference speech separately.

exploits information from the mixed and enrollment speech to directly predict the summed embeddings of all codec layers, thereby compensating for the limitations of auto-regressive modeling in modeling long-range temporal dependencies and mitigating error accumulation. In the following, we provide a detailed description of LauraTSE’s architecture and design.

1) *Encoder*: The first stage of LauraTSE is the encoding stage. Following LauraGPT’s processing strategy for speech enhancement, we first compute log-mel spectrogram features for both the enrollment speech and the mixed speech, denoted as \mathbf{M}_m and \mathbf{M}_r . These two feature streams are then fed into a parameter-sharing Conformer [78] encoder, producing continuous representations for the reference speech and the mixed speech:

$$\mathbf{E}_m = C(\mathbf{M}_m) \quad (1)$$

$$\mathbf{E}_r = C(\mathbf{M}_r) \quad (2)$$

where $\mathbf{E}_m \in \mathbb{R}^{N \times L_m}$ and $\mathbf{E}_r \in \mathbb{R}^{N \times L_r}$ represent the encoded outputs of the \mathbf{M}_m and \mathbf{M}_r , respectively. $C(\cdot)$ denotes the conformer block. N is the feature dimension. L_m and L_r are the number of time steps.

This encoding stage serves as a feature adapter within the overall framework. Rather than directly performing target speaker extraction, its primary objective is to map raw acoustic features into a continuous representation space that is more suitable for subsequent modeling by the auto-regressive decoder-only language model, thereby providing high-quality and structured inputs for generative modeling. It is worth noting that, unlike SpeechX [54], which uses discrete representations produced by a neural audio codec as inputs to the AR model, this work, like LauraGPT [77], preserves task-driven continuous feature representations. This design choice

avoids potential information loss introduced by discretization, particularly for fine-grained speaker-related acoustic characteristics.

2) *Auto-Regressive Decoder-Only Language Model*: The auto-regressive decoder-only language model is designed to learn and predict the joint probability distribution of the coarse-grained discrete representations of the target speech, conditioned on the enrollment speech and the mixed speech. Specifically, the model factorizes the joint distribution of the target speech representations according to the chain rule of probability as follows:

$$P_\theta(\hat{\mathbf{D}}_n | \mathbf{E}_m, \mathbf{E}_r) = \prod_{i \leq T} P_\theta(\hat{\mathbf{D}}_n^{(i)} | \hat{\mathbf{D}}_n^{(1:i-1)}, \mathbf{E}_m, \mathbf{E}_r) \quad (3)$$

where T denotes the length of the output signal, and θ denotes the model parameters, and $\hat{\mathbf{D}}_n$ denotes the generated discrete representation of the target speech.

During training, the input sequence to the AR decoder-only language model is organized as $[\text{bos}, \mathbf{E}_r, \text{sep}, \mathbf{E}_m, \text{tse}, \mathbf{D}_n]$, where bos is a learnable beginning-of-sequence token, sep separates the enrollment and mixed speech embeddings, tse marks the boundary between conditional inputs and target outputs, and \mathbf{D}_n denotes the sum of embeddings from the first n residual vector quantization (RVQ) layers of the target speech. The AR model is trained to predict the discrete representations of the first n RVQ layers. After generating hidden states, n parallel linear layers estimate token distributions for each RVQ layer, and a cross-entropy loss is applied between the predicted and ground-truth token distributions. The predicted tokens are then mapped to continuous embeddings using the codec decoder’s embedding tables and summed across layers

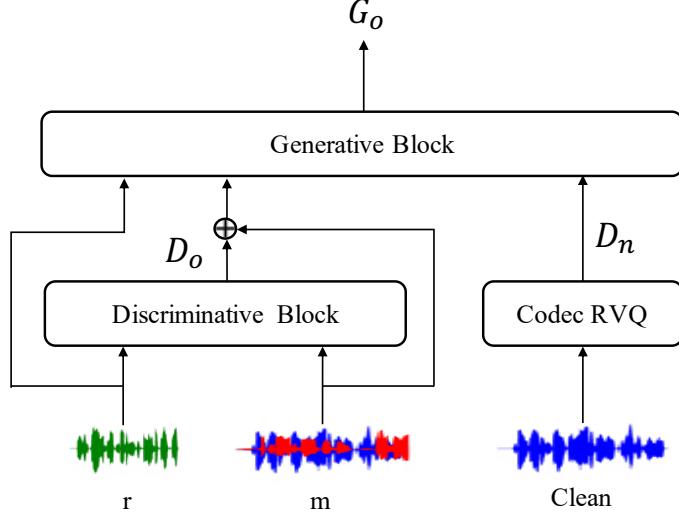


Fig. 3. The diagram of discriminative-generative target speaker extraction framework. ‘m’ and ‘r’ denote the mixed speech and reference speech, respectively.

to form a coarse-grained representation $\hat{\mathbf{D}}_n$. During inference, the model generates $\hat{\mathbf{D}}_n$ auto-regressively, frame by frame.

3) *Vocoder*: The objective of the vocoder module is to reconstruct a high-fidelity time-domain waveform of the target speaker by fully leveraging information from the mixed speech and the enrollment speech, based on the coarse-grained representations generated by the auto-regressive model. To this end, we design a vocoder consisting of an encoder-only language model and a frozen, pre-trained neural audio codec decoder. The encoder-only language model is built upon self-attention mechanisms, enabling effective modeling of long-range temporal dependencies and capturing fine-grained acoustic structures and speech details. Unlike SpeechX [54], which predicts RVQ codes layer-wise, our design adopts a one-step encoder-only language model that directly predicts the summed embeddings across all RVQ layers of the target speech. This formulation substantially simplifies the modeling process while improving both training and inference efficiency.

Specifically, the input to the encoder-only language model is the concatenated feature sequence $[\mathbf{E}_r, \mathbf{E}_m, \hat{\mathbf{D}}_n]$:

$$[\cdot, \cdot, \hat{\mathbf{E}}_s] = \text{EL}([\mathbf{E}_r, \mathbf{E}_m, \hat{\mathbf{D}}_n]) \quad (4)$$

where \mathbf{E}_r and \mathbf{E}_m denote the continuous embeddings of the enrollment speech and the mixed speech, respectively, and $\hat{\mathbf{D}}_n$ represents the embedding corresponding to the coarse-grained target speech representation generated by the first-stage AR decoder-only language model. $\text{EL}(\cdot)$ denotes the encoder-only language model. The encoder-only model processes this sequence and outputs $[\cdot, \cdot, \hat{\mathbf{E}}_s]$, where $\hat{\mathbf{E}}_s$ denotes the predicted fine-grained acoustic embedding of the target speaker. During training, the predicted embedding $\hat{\mathbf{E}}_s$ is supervised against the ground-truth target speech embedding \mathbf{E}_s , obtained from the neural audio codec as the sum of embeddings across all RVQ layers. Both L1 and L2 losses are jointly employed to optimize reconstruction accuracy and training stability. Finally, the frozen codec decoder converts the predicted embedding $\hat{\mathbf{E}}_s$ into the time-domain waveform of the target speaker’s speech.

It is worth emphasizing that the AR decoder-only language model and the encoder-only language model are jointly trained end-to-end.

B. Architecture

To leverage the advantages of both discriminative and generative approaches simultaneously, this work proposes a two-stage discriminative–generative framework for target speaker extraction. As illustrated in Fig. 3, the framework consists of two collaborative modules: a discriminative module and a generative module. The discriminative module explicitly extracts target-speaker-related speech components or intermediate acoustic representations from the mixed speech, providing high-quality and low-interference conditional inputs for the subsequent generative module. The generative module then performs generative reconstruction based on the outputs of the discriminative module, further enhancing the perceptual quality of the target speech.

In the discriminative module, the discriminative block takes the reference speech r and the mixed speech m as inputs, and extracts target-related information by suppressing interference from non-target speakers. This module outputs a coarse target representation \mathbf{D}_o , which can be interpreted as an estimated target speech signal or an intermediate acoustic representation:

$$\mathbf{D}_o = \mathcal{D}(\mathbf{m}, \mathbf{r}) \quad (5)$$

where $\mathcal{D}(\cdot)$ denotes the discriminative extraction function.

In parallel, the ground-truth clean target speech is encoded by a neural audio codec with RVQ, producing a coarse discrete representation \mathbf{D}_n :

$$\mathbf{D}_n = \mathcal{Q}(\mathbf{s}) \quad (6)$$

where \mathbf{s} denotes the clean target speech and $\mathcal{Q}(\cdot)$ represents the codec encoder.

In the generative module, the generative block takes the discriminative output \mathbf{D}_o as a conditional input and leverages

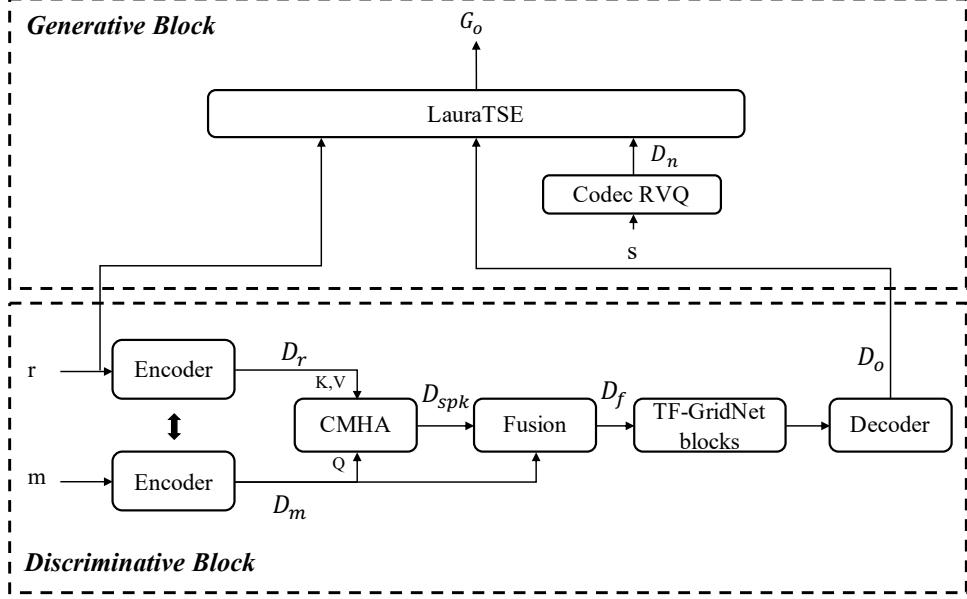


Fig. 4. The diagram of USEF-Laura-TSE. ‘m’ and ‘r’ denote the mixed speech and reference speech, respectively.

generative modeling to reconstruct a refined target speech representation. During training, the generative block is supervised using the codec representation \mathbf{D}_o , and the final output \mathbf{G}_o is generated as:

$$\mathbf{G}_o = \mathcal{G}(\mathbf{D}_o, \mathbf{D}_n) \quad (7)$$

where $\mathcal{G}(\cdot)$ denotes the generative reconstruction function. At inference time, only \mathbf{D}_o is required, and the generative block produces the enhanced target speech output \mathbf{G}_o .

Through this two-stage design, the discriminative block provides a low-interference and well-aligned target representation. In contrast, the generative block further refines speech details and improves perceptual quality via distribution-level modeling.

C. USEF-Laura-TSE

To validate the effectiveness of the proposed framework, we construct a two-stage discriminative–generative target speaker extraction network, termed USEF-Laura-TSE, that employs USEF-TFGridNet [45] as the discriminative front-end and LauraTSE as the generative back-end. The overall architecture of LauraTSE is illustrated in Fig 4.

1) *Discriminative Block (USEF-TFGridNet)*: Given the reference speech \mathbf{r} and the mixed speech \mathbf{m} , both signals are first transformed into the time–frequency (T–F) domain using the short-time Fourier transform (STFT), followed by 2-D convolutional encoders:

$$\mathbf{D}_m = \text{Enc}(\mathbf{m}) \quad (8)$$

$$\mathbf{D}_r = \text{Enc}(\mathbf{r}) \quad (9)$$

where $\text{Enc}(\cdot)$ denotes the shared encoder composed of STFT and 2-D convolution layers.

\mathbf{D}_m and \mathbf{D}_r are fed into the CMHA module, where a cross multi-head attention mechanism is applied to extract frame-level features of the target speaker:

$$\mathbf{D}_{\text{spk}} = \text{CMHA}(q = \mathbf{D}_m; k, v = \mathbf{D}_r) \quad (10)$$

where \mathbf{E}_m and \mathbf{E}_r represent the encoder outputs of the mixed speech and reference speech, respectively. The Cross Multi-Head Attention operation is denoted as $\text{CMHA}(\cdot)$, and \mathbf{E}_{spk} is the output of the CMHA module. The CMHA module in USEF-TFGridNet [45] uses mixed speech encoding as the query. This approach produces a frame-level feature with the same length as \mathbf{D}_m , allowing the mixed and reference speech lengths to differ in the USEF-TFGridNet [45].

The extracted speaker-aware representation \mathbf{D}_{spk} is then fused by direct concatenation with the mixed-speech features:

$$\mathbf{D}_f = \text{Concat}(\mathbf{D}_m, \mathbf{D}_{\text{spk}}) \quad (11)$$

The fused features are processed by a stack of TF-GridNet blocks to model global T–F dependencies. Finally, a decoder composed of 2-D transposed convolutions and inverse STFT (iSTFT) reconstructs the discriminative output \mathbf{D}_o .

2) *Generative Block (LauraTSE)*: The output of the discriminative block \mathbf{D}_o is fed into the generative block as a conditional input. During training, the clean target speech s is encoded by a neural audio codec with RVQ to obtain a coarse discrete representation \mathbf{D}_n . LauraTSE learns to model the conditional distribution of the target speech and generates the final output \mathbf{G}_o . The detailed procedure of LauraTSE is described in Section III-A

IV. EXPERIMENTAL SETUP

A. Datasets

The main experiments in this work are conducted using the 460-hour clean speech subset of the LibriSpeech [79] corpus, referred to as LibriSpeech-460h. The training data

are generated using an online mixing strategy, where speech samples are randomly selected and mixed during training. The relative signal-to-noise ratio (SNR) is randomly sampled from 0 to 5 dB to simulate realistic target-speaker extraction scenarios.

For validation, the clean development set of Libri2Mix [80] is used. During both training and evaluation, the enrollment speech is randomly cropped to 5 seconds to improve robustness to variations in enrollment duration. In the test phase, the clean test set of Libri2Mix is used for evaluation, where an enrollment utterance is randomly selected for each target speaker, thereby more closely reflecting practical target-speaker extraction conditions.

It should be noted that LauraTSE is first pre-trained on LibriSpeech-460h to learn robust speech and speaker representations from large-scale clean speech data. The model is then fine-tuned on the Libri2Mix clean training set to better adapt to the mixed-speech conditions of the target speaker extraction task. For the LauraTSE ablation studies, to ensure fair and controlled comparisons, the model is trained exclusively on the Libri2Mix clean training set.

B. Network Configuration

1) *LauraTSE*: For LauraTSE, we adopt LauraGPT [77] as the backbone of the AR decoder-only language model and employ FunCodec [76] as the neural audio codec. In LauraTSE, the AR model predicts $n = 2$ codec output layers, i.e., only the first two RVQ layers are modeled, resulting in a coarse-grained representation of the target speech. In the feature encoding stage, both the enrollment speech and the mixed speech are analyzed using a window length of 512 samples and a frame shift of 256 samples. The extracted features are then processed by a shared Conformer encoder consisting of six Conformer layers, each with eight attention heads and a hidden dimension of 512, to obtain continuous acoustic representations with rich contextual information. In the generative module, the AR decoder-only Transformer comprises 10 Transformer blocks, each with eight attention heads and a hidden dimension of 512. Conditioned on the encoded representations of the enrollment and mixed speech, the AR model predicts the discrete codec representations corresponding to the first n RVQ layers of the target speech in an auto-regressive, frame-by-frame manner. For waveform reconstruction, an encoder-only Transformer is employed as a refinement module to recover fine-grained acoustic details from the coarse-grained AR outputs. This network consists of six Transformer layers, with eight attention heads and a hidden dimension of 512. Through self-attention mechanisms, the encoder-only Transformer jointly fuses information from the mixed speech, the enrollment speech, and the AR-predicted codec representations, and directly estimates the complete RVQ representation of the target speech, enabling high-fidelity waveform reconstruction.

2) *USEF-LauraTSE*: USEF-LauraTSE employs USEF-TFGridNet [45] as the discriminative front-end. In the encoder, the STFT is computed with a 20 ms window length and a 10 ms frame shift, using a 128-point FFT, producing 161-dimensional complex-valued STFT features per frame. These

features are processed by 2-D convolutional layers with a kernel size of 3×3 and a stride of 1, with two input and 128 output channels. The cross-head attention module uses a single-layer cross-attention structure with four parallel attention heads and a feed-forward network (FFN) with a hidden dimension of 512. In both the full-band and sub-band modules, bidirectional long short-term memory (BLSTM) networks with 256 hidden units are employed as sequence modeling components to capture contextual dependencies along the temporal and frequency dimensions, respectively. Subsequently, a cross-frame self-attention module with a single attention layer, four attention heads, and a 512-dimensional FFN is applied to model global correlations across time-frequency units. The number of TF-GridNet blocks in the separator is set to 2 and 6 for the USEF-TFGridNet-S and USEF-TFGridNet-L configurations, respectively. In the decoder, 2-D transposed convolutional layers are used to reduce the feature channel dimension from 256 to 2, with kernel sizes and strides mirroring those of the encoder convolutional layers. The decoder outputs a complex time-frequency spectrum estimate of the target speaker’s speech, which is subsequently transformed back into the time domain.

C. Training Details

In the proposed discriminative-generative two-stage target speaker extraction framework, a stage-wise training strategy is adopted. The first stage corresponds to the discriminative front-end based on the USEF-TFGridNet [45]. This stage is first pre-trained independently on the Libri2Mix dataset to obtain stable and discriminative target speaker representations, as well as preliminary speech reconstruction capability. After pre-training, several training strategies are explored when jointly training with the generative module:(1) Freezing the discriminative module parameters and training only the generative module, in order to preserve the stability of the discriminative front-end. (2) Unfreezing the discriminative module parameters and performing end-to-end joint training further to enhance the collaborative modeling between the two modules. (3) Introducing an additional Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [81] loss at the output of the discriminative module to regularize its reconstruction quality during joint training.The SI-SDR loss is defined as follows:

$$\begin{cases} \mathbf{s}_T = \frac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s}}{\|\mathbf{s}\|^2} \\ \mathbf{s}_E = \hat{\mathbf{s}} - \mathbf{s}_T \\ \text{SI-SDR} = -10 \lg \frac{\|\mathbf{s}_T\|^2}{\|\mathbf{s}_E\|^2} \end{cases} \quad (12)$$

where $\hat{\mathbf{s}} \in \mathbb{R}^{1 \times T}$ represents the estimated target speaker speech, while $\mathbf{s} \in \mathbb{R}^{1 \times T}$ represents the clean source speech. $\langle \mathbf{s}, \mathbf{s} \rangle$ denotes the power of the signal \mathbf{s} .

The back-end corresponds to the generative module, which uses LauraTSE, an auto-regressive decoder-only language model, as its backbone. Unlike the discriminative module, LauraTSE is trained from scratch without relying on any pre-trained weights. The total number of model parameters is approximately 77M, of which 36M are allocated to the decoder-only Transformer. In contrast, the remaining parameters primarily come from the continuous feature encoder and

the vocoder-related modules. During training, the Adam [82] optimizer is used with an initial learning rate of 1×10^{-3} . To stabilize the training of the large-scale generative model, a 10,000-step warm-up learning rate schedule is applied. When validation performance does not improve for three consecutive epochs, the learning rate is halved. The total number of training epochs is set to 100.

D. Evaluation Metrics

Because vocoder-based waveform generation may introduce deviations in temporal alignment and phase details relative to the original clean speech, traditional intrusive speech quality metrics that rely on a reference signal may fail to reflect perceived quality when evaluating generative models accurately. Therefore, in this work, intrusive metrics such as PESQ [83] and STOI [84] are not adopted. Instead, we primarily employ the following evaluation metrics that are more suitable for generative speech modeling, most of which are non-intrusive:

- **DNSMOS** [85]: A non-intrusive objective speech quality metric that outputs three scores ranging from 1 to 5, including SIG (speech signal quality), BAK (background noise suppression), and OVRL (overall perceptual quality). DNSMOS has been shown to correlate well with human subjective judgments.
- **NISQA** [86]: Another non-intrusive speech quality assessment metric that predicts an overall perceptual quality score (1–5). It exhibits high correlation with subjective listening tests across diverse real-world scenarios.
- **SpeechBERT** [87]: A semantic similarity metric inspired by BERTScore, operating in a self-supervised speech representation space. It measures semantic consistency between the generated speech and the target speech. In this work, speech features are extracted using the HuBERT-base [88] model.
- **Differential Word Error Rate (dWER)** [89]: An intelligibility-oriented metric that computes the difference in word error rate between the generated speech and the ground-truth speech using an automatic speech recognition system. It reflects both intelligibility and semantic fidelity. We employ the *base* version of Whisper [90] for evaluation.
- **Speaker Similarity**: This metric evaluates speaker identity preservation by computing the cosine similarity between the generated speech and the ground-truth target speech in a high-dimensional speaker embedding space. Two speaker verification models are used: WavLM-base¹ and the *ResNet_221LM* model from WeSpeaker [91].

V. RESULTS AND DISCUSSIONS

This section presents a systematic experimental evaluation of the proposed discriminative-generative two-stage TSE framework. Experiments are primarily conducted on the LibriMix dataset, with performance assessed from multiple perspectives, including speech quality, semantic consistency, and speaker similarity. To better understand the respective roles

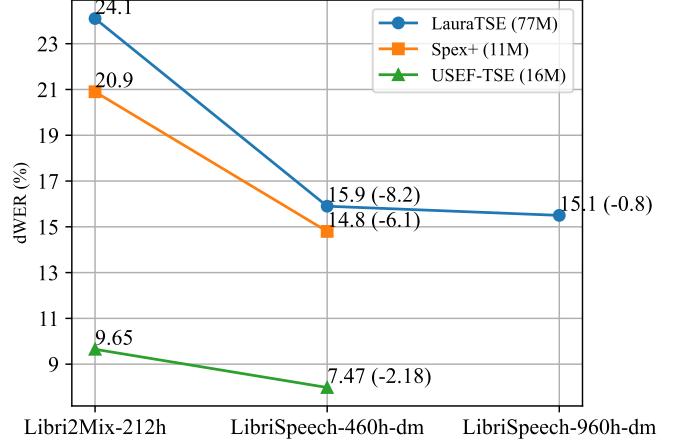


Fig. 5. dWER versus training data scale across models. Annotations “(−X)” denote relative dWER reduction (percentage points) compared to the preceding smaller dataset.

and advantages of different modeling paradigms in TSE, the experimental analysis is organized into two parts. First, we evaluate the generative model LauraTSE on the LibriMix dataset, with a particular focus on examining the modeling capacity, strengths, and limitations of the auto-regressive decoder-only generative paradigm for the TSE task. Subsequently, we assess the proposed discriminative–generative two-stage system, USEF-LauraTSE, and investigate how the introduction of a discriminative front-end improves the stability, robustness, and overall performance of the generative model.

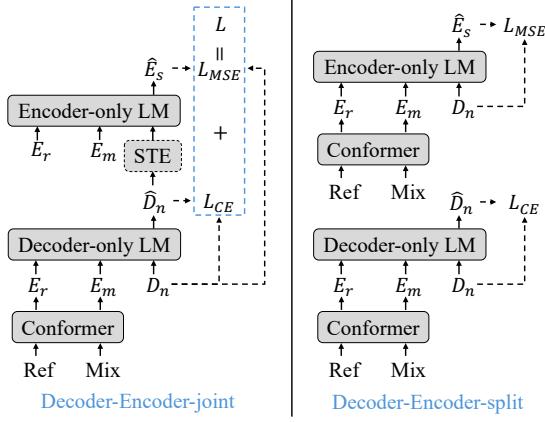
A. Ablation Results of LauraTSE

1) *Data Scalability of LauraTSE*: To evaluate the scalability of LauraTSE with respect to training data size, we compare it with two representative discriminative TSE models, namely SpEx+ [36] and USEF-TSE (USEF-TFGridNet-L) [45]. The parameter sizes of SpEx+ and USEF-TFGridNet-L are approximately 11 M and 16 M, respectively, whereas LauraTSE contains approximately 77 M parameters. In this experiment, LauraTSE is trained using three progressively larger training datasets: (1) *Libri2Mix-212h*, derived from the Libri2Mix-clean subset and consisting of approximately 212 hours of clean mixed speech; (2) *LibriSpeech-460h-dm*, constructed from the LibriSpeech-460h subset using a dynamic mixing strategy; and (3) *LibriSpeech-960h-dm*, generated from the full 960-hour LibriSpeech dataset, also using dynamic mixing.

The experimental results are presented in Fig. 5. As shown, when the training data are expanded from Libri2Mix to LibriSpeech, the performance of the discriminative model USEF-TSE [45] remains relatively stable, indicating limited sensitivity to training data scale. In contrast, both SpEx+ [36] and LauraTSE exhibit pronounced performance improvements as the amount of training data increases. In particular, LauraTSE achieves an absolute reduction of 8.2% in dWER, which is notably larger than the 6.1% improvement observed for SpEx+ [36]. These results indicate that LauraTSE benefits more substantially from large-scale training data, highlighting its superior scalability and data efficiency. This observation

¹<https://huggingface.co/microsoft/wavlm-base-plus-sv>

Fig. 6. Decoder-Encoder Joint vs. Split. *Decoder-Encoder-join* denotes the proposed LauraTSE model where Decoder and Encoder are trained together by Cross-Entropy Loss and MSE Loss. In *Decoder-Encoder-split*, the Decoder and Encoder is trained separately.



further suggests that auto-regressive generative models can better exploit increased data diversity when sufficient training data are available.

Notably, when the training data are increased from 460 to 960 hours, LauraTSE’s performance gains become relatively marginal. This phenomenon may be attributed to two factors. First, data quality may play an important role: the LibriSpeech-960h corpus contains a large proportion of utterances from the *train-other-500* subset, which involves more challenging acoustic conditions and higher noise levels, potentially offsetting the benefits of increased data quantity. Second, model capacity may constitute a limiting factor, as the current scale of LauraTSE may be insufficient to leverage the additional information provided by substantially larger datasets fully.

Moreover, even with increased training data, LauraTSE does not consistently outperform USEF-TSE [45] on semantic consistency metrics. This observation raises an important open question: can generative models comprehensively surpass well-established discriminative approaches in terms of semantic consistency for the target speaker extraction task? Addressing this question represents a promising direction for future research.

2) Impact Analysis of the Encoder-Only Language Model on LauraTSE: Ideally, a generative TSE system would rely solely on a decoder-only language model without introducing an additional encoder-only module. However, due to the limited capacity of a single-layer neural audio codec to simultaneously preserve fine-grained speech details and multi-layer acoustic information, the coarse discrete representations generated by a decoder-only language model alone are often insufficient for high-quality speech reconstruction. To address this limitation, the LauraTSE framework incorporates an encoder-only language model to refine the coarse representations produced by the decoder-only language model into high-resolution continuous acoustic embeddings. This subsection systematically investigates the role of the encoder-only language model in LauraTSE.

We first examine whether joint optimization of the decoder-only and encoder-only language models is necessary. Al-

TABLE I
EVALUATION RESULTS FOR DIFFERENT DECODER-ENCODER CONFIGURATIONS. *Decoder-Encoder-join* AND *Decoder-Encoder-split* REFER TO THE TWO INTEGRATION STRATEGIES ILLUSTRATED IN FIGURE 6. *Target-n* DENOTES THE RECONSTRUCTED TARGET CLEAN AUDIO USING ONLY THE FIRST n LAYERS OF THE CODEC. *No-Encoder* USES SUMMATION OF ONLY THE FIRST n LAYERS OF THE DECODER-ONLY LM OUTPUT TO GENERATE SPEECH WITHOUT THE ENCODER.

Model	NISQA \uparrow	dWER \downarrow	WeSpeaker Sim \uparrow
Decoder-Encoder-join	4.241	0.241	0.847
Decoder-Encoder-split	4.253	0.232	0.858
Target- n ($n = 2$)	3.644	0.301	0.740
No-Encoder	3.807	0.579	0.709

TABLE II
INPUT COMPOSITION RESULTS FOR THE ENCODER-ONLY LM.

Model	Input		NISQA \uparrow	dWER \downarrow	WeSpeaker Sim \uparrow	
Model	E_r	E_m	D_n	NISQA \uparrow	dWER \downarrow	WeSpeaker Sim \uparrow
Encoder-All	✓	✓	✓	4.241	0.241	0.847
Encoder-Mix	✗	✓	✓	4.173	0.239	0.842
Encoder-Ref	✓	✗	✓	4.187	0.480	0.763

though the two modules are architecturally coupled, their learning objectives differ substantially: the decoder-only language model focuses on auto-regressive sequence modeling, whereas the encoder-only language model emphasizes reconstruction and refinement of continuous acoustic representations. Accordingly, we consider two training strategies, as illustrated in Fig. 6: (1) *Decoder-Encoder-join*, which follows the original LauraTSE training scheme, where both modules are jointly optimized end-to-end using the straight-through estimator (STE) [92] to enable gradient propagation through the softmax-based discretization; and (2) *Decoder-Encoder-split*, where the decoder-only language model and the encoder-only language model are trained separately, with the latter optimized using fixed outputs from the former. As reported in Table I, the split training strategy yields slightly better overall performance than joint training. This result suggests that, under the current configuration, strict end-to-end joint optimization between the decoder-only and encoder-only language models is not essential.

We further analyze the contribution of different input sources to the encoder-only language model. In the original LauraTSE framework, the encoder-only language model takes feature representations from both the mixed speech and the enrollment speech as inputs. To isolate the effect of each input, we construct three variants: (1) *Encoder-All*, which uses both mixed speech and enrollment speech representations (original setting); (2) *Encoder-Mix*, which uses only the mixed speech representation; and (3) *Encoder-Ref*, which uses only the enrollment speech representation. The corresponding results are summarized in Table II. It can be observed that *Encoder-Mix* achieves performance comparable to *Encoder-All*, whereas *Encoder-Ref* suffers from a pronounced performance degradation. These results indicate that the mixed speech representation provides indispensable information for the encoder-only language model, while the enrollment speech plays a relatively

TABLE III

ABLATION STUDIES OF LAURATSE. n - DENOTES THE OUTPUT LAYER NUMBER OF THE DECODER-ONLY LM. THE *Ref output* FORMATS THE OUTPUT OF THE DECODER-ONLY LM TO CONTAIN BOTH THE CLEAN AND REFERENCE SPEECH. *Discrete IO* USES DISCRETE CODEC EMBEDDINGS RATHER THAN CONTINUOUS FEATURES AS THE INPUT FEATURES. FOR *WavLM input*, THE WAVLM [53] EMBEDDINGS ARE UTILIZED AS THE INPUT FEATURES.

Model	DNSMOS \uparrow			NISQA \uparrow	SpeechBERT \uparrow	dWER \downarrow	WavLM Sim \uparrow	Wespeaker Sim \uparrow
	SIG	BAK	OVL					
Base (n -2)	3.626	4.102	3.360	4.241	0.880	0.241	0.965	0.847
n -1	3.604	4.100	3.339	4.201	0.861	0.266	0.958	0.830
n -3	3.618	4.095	3.350	4.270	0.880	0.235	0.967	0.853
Ref output	3.588	4.071	3.318	4.182	0.859	0.237	0.962	0.851
Discrete IO	3.562	4.035	3.268	3.940	0.810	0.421	0.952	0.835
WavLM input	3.507	3.951	3.137	3.220	0.792	0.447	0.860	0.633

limited role at this stage. This finding further suggests that the encoder-only language model does not merely serve as a post-processing vocoder, but also continues to participate in task-related modeling for target-speaker extraction.

3) *Comparison Results of the Decoder-Only Language Model:* Table III summarizes the ablation results under different input and output configurations of the decoder-only language model. *Base* denotes the proposed LauraTSE baseline, while n - indicates the number of RVQ layers predicted by the auto-regressive decoder-only language model. Varying n from 1 to 3 yields only marginal performance differences, suggesting that predicting a small number of coarse-grained RVQ layers is sufficient for the target-speaker extraction task.

To investigate whether strict length alignment between the conditional input and the generated output is necessary, we reformulate the decoder-only input sequence as $[bos, E_r, E_m, tse]$ and require the model to generate an output sequence containing both the enrollment and the enhanced speech, referred to as *Ref output*. During inference, only the segment corresponding to the mixed speech is retained. This variant achieves performance comparable to the original setting, indicating that strict input–output length alignment is not required and that the decoder-only language model can focus solely on generating the clean target speech.

Inspired by SpeechX [54], we further evaluate a discrete-input variant (*Discrete IO*), in which continuous log-mel features are replaced with discrete RVQ codebook indices. This configuration consistently performs worse than the continuous-feature baseline, likely due to information loss from discretization and a mismatch between codec representations trained on clean speech and the mixed-speech inputs encountered during target speaker extraction.

Finally, we examine the use of WavLM [53] features as model inputs (*WavLM input*). Consistent with previous observations, this variant exhibits degraded speaker similarity performance, which may be attributed to the reduced preservation of speaker-related acoustic characteristics in the discretized representations.

B. Ablation Results of USEF-LauraTSE

This subsection presents a systematic analysis of the proposed discriminative–generative two-stage target speaker extraction model, USEF-LauraTSE, evaluated on the Libri2Mix dataset. The analysis focuses on three aspects: (1) the impact

of the discriminative front-end on overall system performance; (2) the behavior of both the discriminative front-end and the generative back-end when an additional SI-SDR loss is imposed on the front-end outputs; and (3) the performance differences between auto-regressive and non-auto-regressive inference strategies under identical training conditions.

1) *Impact Analysis of the Discriminative Front-End:* To investigate the effect of introducing a discriminative front-end prior to the generative model, we consider two training strategies: (i) a *frozen* setting, where the pre-trained discriminative front-end is kept fixed and used solely as a feature extractor; and (ii) an *unfrozen* setting, where the discriminative front-end is jointly optimized together with the generative back-end. Table IV reports the test results on Libri2Mix for the purely discriminative model (USEF-TFGridNet-S), the purely generative model (LauraTSE), and the proposed two-stage model (USEF-LauraTSE-S).

A comparison between the purely generative model LauraTSE and the two-stage model USEF-LauraTSE-S shows that introducing a discriminative front-end consistently improves performance across multiple evaluation metrics. Under the Libri2Mix training configuration, USEF-LauraTSE-S achieves comparable or improved speech quality (DNSMOS-OVRL: 3.336 \rightarrow 3.341), semantic consistency (SpeechBERT: 0.908 \rightarrow 0.910), and speaker similarity (WavLM similarity: 0.974 \rightarrow 0.973), while further reducing the dWER (0.159 \rightarrow 0.153). These results indicate that the discriminative front-end provides more structured and interference-suppressed intermediate representations, thereby easing the semantic modeling burden and speaker preservation requirements of the generative back-end.

Further comparison between frozen and unfrozen training strategies reveals that freezing the discriminative front-end is suboptimal. Although the frozen configuration still outperforms the purely generative model, it exhibits noticeable degradation in semantic-related metrics, such as SpeechBERT (0.869) and dWER (0.266), compared with the unfrozen setting (SpeechBERT: 0.910, dWER: 0.153). This observation suggests that joint optimization enables effective feedback from the generative back-end to the discriminative front-end, guiding it toward intermediate representations that are more amenable to generative modeling.

When joint training is conducted on a larger-scale dataset (Training Data = 2) with the discriminative front-end unfrozen,

TABLE IV

RESULTS OF THE DISCRIMINATIVE-GENERATIVE MODELS ON THE LIBRIMIX CLEAN TEST SET. IN THE “CATEGORY” COLUMN, “D” DENOTES A DISCRIMINATIVE MODEL, “G” DENOTES A GENERATIVE MODEL, AND “D-G” DENOTES A DISCRIMINATIVE-GENERATIVE MODEL. IN THE “TRAINING DATA” COLUMN, “1” INDICATES TRAINING ON LIBRIMIX, WHILE “2” DENOTES TRAINING WITH ONLINE MIXING ON LIBRISPEECH FOLLOWED BY FINE-TUNING ON LIBRIMIX. USEF-TFGRIDNET-S REFERS TO THE USEF-TFGRIDNET MODEL WITH TWO TF-GRIDNET BLOCKS, AND USEF-LAURA-TSE-S DENOTES THE DISCRIMINATIVE-GENERATIVE MODEL THAT EMPLOYS USEF-TFGRIDNET-S AS THE DISCRIMINATIVE FRONT-END AND LAURATSE AS THE GENERATIVE BACK-END. “SBERT” DENOTES SPEECHBERT SCORE.

Model	Category	Frozen	Training Data	DNSMOS ↑			NISQA ↑	SBERT ↑	dWER ↓	WavLM ↑	WeSpeaker ↑
				SIG	BAK	OVRL					
USEF-TFGridNet-S	D	-	1	3.308	3.745	2.926	3.349	0.807	0.228	0.961	0.912
LauraTSE	G	-	1	3.629	4.102	3.360	4.241	0.879	0.241	0.965	0.847
			2	3.609	4.084	3.336	4.333	0.908	0.159	0.974	0.876
USEF-Laura-TSE-S	D-G	✓	1	3.606	4.100	3.344	4.304	0.869	0.266	0.963	0.851
		✗	1	3.609	4.086	3.341	4.350	0.910	0.153	0.973	0.879
		✗	2	3.592	4.061	3.313	4.453	0.925	0.120	0.978	0.895

USEF-LauraTSE-S achieves the best overall performance across nearly all evaluation metrics. Specifically, the dWER is further reduced to 0.120, SpeechBERT improves to 0.925, and speaker similarity reaches its highest levels (WavLM: 0.978, WeSpeaker: 0.895). In contrast, while the purely discriminative model USEF-TFGridNet-S remains competitive in terms of dWER (0.228) and speaker similarity (0.961 / 0.912), it lags behind both generative and discriminative-generative models in perceptual speech quality metrics, such as DNSMOS-OVRL (2.926) and NISQA (3.349).

This performance gap can be attributed to two factors. First, the discriminative front-end employs only two TF-GridNet blocks, which constrains its modeling capacity. Second, discriminative approaches are more prone to over-suppression and signal distortion, whereas generative models offer inherent advantages for reconstructing natural, perceptually pleasing speech. By integrating these complementary strengths, the proposed discriminative-generative framework effectively balances robust target localization and interference suppression with high-quality speech reconstruction.

2) *Bidirectional Interaction Between the Discriminative and Generative Modules:* Previous experiments primarily evaluate the discriminative-generative framework from the perspective of its generative outputs. In this subsection, we shift the focus to the discriminative module’s outputs to investigate the bidirectional interaction between the discriminative front-end and the generative back-end. The corresponding experimental results are summarized in Table V.

We first compare the standalone discriminative and generative models under the Training Data = 2 setting. As shown in the first three rows of Table V, the discriminative model USEF-TFGridNet-L achieves strong overall performance, with a DNSMOS-OVRL score of 3.272, a NISQA score of 4.319, a low dWER of 0.075, and speaker similarity scores exceeding 0.98 for both WavLM and WeSpeaker. In contrast, the purely generative model LauraTSE slightly outperforms USEF-TFGridNet-L in perceptual quality metrics (DNSMOS-OVRL: 3.336; NISQA: 4.333), but exhibits inferior semantic consistency and speaker preservation, as reflected by a higher dWER (0.159) and a lower WeSpeaker score (0.876). A similar trend is observed for USEF-TFGridNet-S, where the discriminative model maintains advantages in dWER and speaker similarity, while lagging behind the generative model

in perceptual quality.

For USEF-LauraTSE-S without an additional SI-SDR constraint (“SI-SDR = No, O = G”), the discriminative front-end degrades noticeably during joint training. This behavior indicates that the front-end no longer prioritizes the perceptual quality of its own outputs. Instead, it adapts to produce intermediate representations that are more favorable for the generative back-end. Despite the degraded discriminative output, the generative output of USEF-LauraTSE-S consistently outperforms the standalone LauraTSE across multiple metrics, including NISQA (4.453 vs. 4.333), SpeechBERT (0.925 vs. 0.908), dWER (0.120 vs. 0.159), and speaker similarity (WavLM: 0.978 vs. 0.974; WeSpeaker: 0.895 vs. 0.876). These results suggest that even a weakened discriminative front-end can still provide sufficient target alignment and coarse separation, enabling the generative back-end to exploit its strong reconstruction capability.

To prevent excessive degradation of the discriminative front-end, an SI-SDR loss is introduced to constrain its outputs explicitly. With this constraint (“SI-SDR = Yes”), the discriminative output of USEF-LauraTSE-S (“O = D”) shows substantial improvements in semantic-related metrics compared with the pre-trained USEF-TFGridNet-S, with dWER reduced from 0.228 to 0.113, WavLM increased from 0.961 to 0.977, and WeSpeaker improved from 0.912 to 0.937. DNSMOS-OVRL increases slightly, while NISQA decreases to some extent, indicating that the discriminative front-end sacrifices a small amount of perceptual quality in exchange for more structured and semantically reliable representations. Correspondingly, the generative output exhibits a modest degradation in NISQA (4.416 vs. 4.453), an increase in dWER (0.120 → 0.154), and a slight decrease in speaker similarity, suggesting that a stronger discriminative constraint limits the flexibility of the generative model in adjusting fine-grained waveform details.

A similar pattern is observed for USEF-LauraTSE-L. Compared with the standalone USEF-TFGridNet-L, the discriminative output after joint training shows a marginal increase in dWER (0.075 → 0.076), accompanied by decreases in DNSMOS-OVRL and NISQA. It indicates that, under the discriminative-generative objective, a stronger discriminative front-end does not over-optimize its own perceptual quality, but instead produces intermediate representations that are easier for the generative back-end to reconstruct. In con-

TABLE V

RESULTS ON THE LIBRI2MIX CLEAN TEST SET FOR THE DISCRIMINATIVE–GENERATIVE MODELS WITH AN ADDITIONAL SI-SDR LOSS. IN THE “O” COLUMN, “D” DENOTES THE OUTPUT OF THE DISCRIMINATIVE MODEL, AND “G” DENOTES THE OUTPUT OF THE GENERATIVE MODEL. IN THE “MODEL” COLUMN, USEF-LAURA-TSE-S REFERS TO THE DISCRIMINATIVE–GENERATIVE SYSTEM THAT USES USEF-TFGRIDNET-S AS THE DISCRIMINATIVE FRONT-END AND LAURATSE AS THE GENERATIVE BACK-END, WHILE USEF-LAURA-TSE-L EMPLOYS USEF-TFGRIDNET-L (WITH SIX TF-GRIDNET BLOCKS) AS THE DISCRIMINATIVE FRONT-END AND LAURATSE AS THE GENERATIVE BACK-END. FOR ALL USEF-LAURA-TSE VARIANTS, THE DISCRIMINATIVE FRONT-END IS PRE-TRAINED ONLY ON LIBRI2MIX. IN THE “TRAINING DATA” COLUMN, “1” INDICATES TRAINING ON LIBRIMIX, AND “2” DENOTES TRAINING WITH ONLINE MIXING ON LIBRISPEECH FOLLOWED BY FINE-TUNING ON LIBRI2MIX. “SBERT” DENOTES THE SPEECHBERT SCORE.

Model	Training Data	SI-SDR Loss?	O	DNSMOS		NISQA	SBERT	dWER	WavLM	Wespeaker
			SIG	BAK	OVRL					
USEF-TFGridNet-L	1	-	D	3.514	4.041	3.249	0.909	0.104	0.982	0.953
USEF-TFGridNet-L	2	-	D	3.555	4.051	3.272	0.935	0.075	0.988	0.968
LauraTSE	2	-	G	3.609	4.066	3.336	0.908	0.159	0.974	0.876
	2	No	D	1.187	1.144	1.100	1.014	0.451	0.693	0.672
USEF-Laura-TSE-S	2	No	G	3.592	4.061	3.313	4.453	0.925	0.120	0.978
	2	Yes	D	3.422	3.661	2.979	3.172	0.884	0.113	0.977
	2	Yes	G	3.603	4.080	3.329	4.416	0.915	0.154	0.975
USEF-Laura-TSE-L	2	Yes	D	3.528	3.955	3.202	3.648	0.933	0.076	0.987
	2	Yes	G	3.592	4.075	3.319	4.450	0.934	0.117	0.982
										0.902

trast, the generative output of USEF-LauraTSE-L significantly outperforms the standalone LauraTSE, with improvements in NISQA (4.450 vs. 4.333), SpeechBERT (0.934 vs. 0.908), dWER (0.117 vs. 0.159), and speaker similarity metrics. Compared with the smaller front-end configuration, the larger discriminative front-end further reduces dWER and improves speaker consistency, demonstrating that increased discriminative capacity provides higher-quality structural information for generative reconstruction.

Overall, these results reveal a fundamental trade-off in discriminative–generative TSE frameworks. Without an SI-SDR constraint, the generative model can fully exploit its distribution modeling capability and exhibits strong robustness to imperfections in the discriminative front-end. Introducing SI-SDR loss substantially strengthens the discriminative module but partially constrains the generative model’s flexibility, resulting in reduced gains in perceptual quality. These findings highlight the importance of carefully balancing front-end controllability and back-end generative freedom when designing training strategies for discriminative–generative target speaker extraction systems.

3) *Auto-Regressive and Non-Auto-Regressive Inference Strategies:* Under identical training conditions, the proposed discriminative–generative framework supports two inference modes at test time: auto-regressive (AR) and non-auto-regressive (NAR). In the standard setting, the decoder-only language model generates discrete target speech representations in an auto-regressive, frame-by-frame manner. Within the discriminative–generative architecture, an alternative NAR inference strategy can be adopted by leveraging the outputs of the discriminative front-end as pseudo labels for the generative model. Specifically, in the NAR inference mode, the training procedure remains unchanged. At inference time, the target speech representations produced by the discriminative front-end are used as pseudo-labels and injected into the decoder-only language model’s decoding sequence, along with the mixed and enrollment speech representations. By controlling the pseudo-label injection ratio R , the inference process can be flexibly adjusted between fully generative decoding ($R = 0$)

and firm reliance on the discriminative front-end estimates ($R = 1$). A smaller R preserves more generative flexibility, whereas a larger R emphasizes the robustness and controllability of the discriminative front-end. Comparative experiments are conducted on four configurations: USEF-LauraTSE-S with a frozen front-end, USEF-LauraTSE-S with an additional SI-SDR loss, a decoupled-training discriminative–generative model, and USEF-LauraTSE-S with SI-SDR loss under decoupled training. The results are reported in Table VI.

As shown in Table VI, the proposed two-stage framework exhibits consistent trends under AR and NAR inference. Taking USEF-LauraTSE-S as an example, when an SI-SDR loss is applied, the AR generative output ($O = G$) achieves the best perceptual quality for this configuration, with DNSMOS-OVRL of approximately 3.33 and NISQA of 4.416, while maintaining high speaker similarity scores. However, the dWER remains at 0.154, indicating that although AR inference maximizes perceptual quality, it is still susceptible to semantic errors and content drift.

When switching to NAR inference and gradually increasing the injection ratio R , a clear trade-off between perceptual quality and intelligibility is observed. As R increases from 0 to 1, DNSMOS-OVRL and NISQA decrease monotonically from approximately 3.33/4.416 to 3.23/4.060, while dWER consistently improves from 0.154 to 0.133. These results indicate that using discriminative front-end outputs as pseudo-labels effectively suppresses auto-regressive drift and hallucination, at the cost of moderate perceptual quality degradation, in exchange for improved semantic stability and intelligibility.

A highly consistent trend is observed for the larger front-end configuration (USEF-LauraTSE-L). Under AR inference, the model again achieves the highest perceptual quality. In NAR mode, increasing R leads to a slight reduction in perceptual metrics (e.g., NISQA decreases from 4.450 to 4.302), while further reducing dWER from 0.112 to 0.099. This consistency across different front-end scales suggests that AR inference prioritizes perceptual quality, whereas NAR inference enhances intelligibility by constraining the generative process.

It is worth noting that the cascaded baseline USEF-

TABLE VI

RESULTS ON THE LIBRI2MIX CLEAN TEST SET FOR USEF-LAURA-TSE UNDER AUTO-REGRESSIVE (AR) AND NON-AUTO-REGRESSIVE (NAR) INFERENCE. IN THE “MODEL” COLUMN, “USEF-TFGRIDNET-L + LAURATSE (SPLIT)” DENOTES THAT USEF-TFGRIDNET-L AND LAURATSE ARE TRAINED SEPARATELY, AND DURING NAR INFERENCE THE OUTPUTS OF USEF-TFGRIDNET-L ARE USED AS PSEUDO LABELS FOR THE GENERATIVE MODEL. IN THE “INFERENCE MODE” COLUMN, “AR” INDICATES AUTOREGRESSIVE INFERENCE AND “NAR” INDICATES NON-AUTOREGRESSIVE INFERENCE. “R” DENOTES THE INJECTION RATIO OF THE DISCRIMINATIVE OUTPUTS. “SBERT” DENOTES THE SPEECHBERT SCORE.

Model	Inference Mode	SI-SDR Loss?	O	R	DNSMOS			NISQA	SBERT	dWER	WavLM	WeSpeaker
			SIG	BAK	OVRL	O						
USEF-Laura-TSE-S	AR	No	D	-	1.187	1.144	1.100	1.014	0.451	0.693	0.672	0.642
		No	G	-	3.592	4.061	3.313	4.453	0.925	0.120	0.978	0.895
	NAR	No	G	0.0	2.647	2.061	1.905	1.864	0.473	1.024	0.796	0.661
		No	G	0.5	2.452	1.768	1.724	1.635	0.454	1.069	0.782	0.647
		No	G	1.0	1.844	1.353	1.404	1.362	0.424	1.105	0.752	0.637
	AR	Yes	D	-	3.422	3.661	2.979	3.172	0.884	0.113	0.977	0.934
		Yes	G	-	3.603	4.080	3.329	4.416	0.915	0.154	0.975	0.880
		Yes	G	0.0	3.590	4.027	3.291	4.217	0.910	0.149	0.975	0.881
USEF-Laura-TSE-S	NAR	Yes	G	0.5	3.578	3.991	3.263	4.099	0.907	0.148	0.975	0.882
		Yes	G	1.0	3.568	3.944	3.232	3.960	0.906	0.133	0.975	0.883
		-	D	-	3.555	4.051	3.272	4.319	0.935	0.075	0.988	0.968
	+ LauraTSE (split)	-	G	-	3.609	4.084	3.336	4.333	0.908	0.159	0.974	0.876
		-	G	0.0	3.587	4.089	3.322	4.512	0.881	0.216	0.969	0.866
		-	G	0.5	3.604	4.101	3.343	4.553	0.898	0.166	0.972	0.872
	-	G	1.0	3.619	4.114	3.363	4.583	0.913	0.120	0.974	0.878	
USEF-Laura-TSE-L	AR	Yes	D	-	3.528	3.955	3.202	3.648	0.933	0.076	0.987	0.950
		Yes	G	-	3.592	4.075	3.319	4.450	0.934	0.117	0.982	0.902
	NAR	Yes	G	0.0	3.580	4.048	3.294	4.346	0.927	0.115	0.981	0.901
		Yes	G	0.5	3.574	4.035	3.283	4.316	0.927	0.112	0.981	0.902
		Yes	G	1.0	3.570	4.022	3.272	4.302	0.929	0.099	0.982	0.903

TFGridNet-L + LauraTSE (decoupled training) achieves the highest DNSMOS-OVRL and NISQA scores across all configurations. This behavior can be attributed to the independently trained discriminative front-end, which provides near-ideal magnitude spectra, and a fully decoupled generative back-end that focuses exclusively on waveform reconstruction without being constrained by SI-SDR loss during joint training. However, this configuration exhibits inferior dWER and speaker similarity compared with the jointly trained USEF-LauraTSE-L, revealing a clear trade-off between perceptual quality and content fidelity. In contrast, the jointly trained USEF-LauraTSE-L achieves a more balanced trade-off among perceptual quality, intelligibility, and speaker consistency. These results suggest that moderate SI-SDR constraints combined with end-to-end optimization effectively prevent excessive generative freedom, thereby improving the reliability and stability of the overall system.

Overall, the USEF-Laura-TSE enables a controllable trade-off between speech quality and intelligibility under a unified training paradigm by flexibly switching between AR and NAR inference modes and adjusting the injection ratio R at inference time. When perceptual quality is the primary objective, pure auto-regressive inference is preferred; when higher intelligibility, ASR robustness, or semantic consistency is required, NAR inference with a larger R provides a more suitable solution by explicitly guiding the generative model with discriminative front-end outputs.

C. Comparison With Previous Models

Table VII summarizes the overall experimental results on the Libri2Mix dataset. The proposed discriminative–generative model USEF-Laura-TSE-L achieves a more balanced performance across perceptual quality, semantic fidelity, and speaker

consistency. Compared with the purely generative LauraTSE, USEF-Laura-TSE-L significantly improves semantic consistency and speaker similarity, with dWER reduced from 0.159 to 0.117, and speaker similarity increased from 0.974/0.876 to 0.982/0.902 (WavLM/WeSpeaker), while maintaining comparable DNSMOS-OVRL and achieving the best NISQA score (4.450) among all systems. This indicates that a stronger discriminative front-end (USEF-TFGridNet-L) provides more reliable and structured intermediate representations, which effectively guide the generative back-end toward improved content stability without sacrificing perceptual quality.

Compared with the strong discriminative baseline USEF-TFGridNet-L [45], which attains the best dWER and speaker similarity, USEF-Laura-TSE-L substantially improves perceptual quality (DNSMOS-OVRL and NISQA), demonstrating that the discriminative–generative framework effectively bridges the gap between discriminative robustness and generative naturalness. Moreover, despite being trained on only 460 hours of data, USEF-Laura-TSE-L achieves performance comparable to or better than large-scale generative systems such as AnyEnhance [49], highlighting the data efficiency and effectiveness of task-oriented discriminative–generative modeling for target speaker extraction.

Overall, these results confirm that combining a strong discriminative front-end with a generative AR decoder-only back-end yields a robust and well-balanced solution, validating the effectiveness of the proposed discriminative–generative two-stage paradigm.

VI. CONCLUSION

This paper first proposes LauraTSE, a generative target speaker extraction (TSE) method based on an auto-regressive decoder-only language model. By leveraging con-

TABLE VII

RESULTS ON LIBRI2MIX CLEAN. IN THE "CATEGORY" COLUMN, "G" REFERS TO GENERATIVE MODELS, WHILE "D" REFERS TO DISCRIMINATIVE MODELS.

Model	Category	DNSMOS \uparrow			NISQA \uparrow	SBERT \uparrow	dWER \downarrow	WavLM \uparrow	Wespeaker \uparrow
		SIG	BAK	OVL					
Mixture	-	3.383	3.098	2.653	2.453	0.572	0.792	0.847	0.759
Spex+ [36]	D	3.472	4.027	3.186	3.349	0.878	0.148	0.973	0.935
WeSep [93]	D	3.486	3.838	3.118	3.892	0.895	0.123	0.980	0.945
USEF-TFGridNet-L [45]	D	3.555	4.051	3.272	4.319	0.935	0.0747	0.988	0.968
TSELM-L [52]	G	3.489	4.041	3.212	3.961	0.793	0.297	0.887	0.627
AnyEnhance [49]	G	3.638	4.066	3.353	4.277	0.735	-	0.914	-
LauraTSE	G	3.609	4.084	3.336	4.333	0.908	0.159	0.974	0.876
USEF-Laura-TSE-L	D-G	3.592	4.075	3.319	4.450	0.934	0.117	0.982	0.902

tinuous acoustic features and a neural audio codec, LauraTSE enables end-to-end generative TSE without explicit speaker embeddings. Experimental results demonstrate its competitive performance in speech quality, speaker similarity, and semantic consistency, and data-scaling experiments further show stronger scalability than conventional discriminative models. Analysis reveals that coarse auto-regressive generation alone is insufficient for fine-grained reconstruction, motivating the introduction of an encoder-only LM to refine acoustic details. Building on this, the chapter presents a discriminative–generative framework in which a USEF-TFGridNet-based discriminative front-end provides structured target representations to guide generative reconstruction. Experiments show that the discriminative–generative design significantly improves speaker consistency and intelligibility, validating the complementary roles of discriminative and generative modeling. Further exploration of SI-SDR-constrained training and non-autoregressive inference highlights the framework’s ability to achieve a controllable trade-off between perceptual quality and semantic robustness.

REFERENCES

- [1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America*, vol. 25, pp. 975–979, 1953.
- [2] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta acustica united with acustica*, vol. 86, no. 1, pp. 117–128, 2000.
- [3] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. of Interspeech*, vol. 2. Citeseer, 2006, pp. 2–5.
- [4] A. Cichocki, R. Zdunek, and S.-i. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 5, 2006, pp. V–V.
- [5] R. Lyon, "A computational model of binaural localization and separation," in *Proc. of ICASSP*, vol. 8, 1983, pp. 1148–1151.
- [6] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.
- [7] G. Hu and D. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 15, no. 2, pp. 396–405, 2007.
- [8] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. of ICASSP*, 2016, pp. 31–35.
- [9] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.
- [10] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. of ICASSP*, 2018, pp. 1–5.
- [11] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. of ICASSP*, 2017, pp. 246–250.
- [12] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 26, no. 10, pp. 1702–1726, 2018.
- [13] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.
- [14] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. of ICASSP*, 2017, pp. 241–245.
- [15] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.
- [16] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *Proc. of ICASSP*, 2018, pp. 696–700.
- [17] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. of ICASSP*, 2020, pp. 46–50.
- [18] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm-rf: Efficient networks for universal audio source separation," in *Proc. of MLSP*, 2020, pp. 1–6.
- [19] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [20] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2840–2849, 2021.
- [21] C. Li, Y. Luo, C. Han, J. Li, T. Yoshioka, T. Zhou, M. Delcroix, K. Kinoshita, C. Boeddeker, Y. Qian et al., "Dual-path rnn for long recording speech separation," in *Proc. of SLT*, 2021, pp. 865–872.
- [22] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.
- [23] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. of ICASSP*, 2021, pp. 21–25.
- [24] K. Li, R. Yang, and X. Hu, "An efficient encoder-decoder architecture with top-down attention for speech separation," *arXiv preprint arXiv:2209.15200*, 2022.
- [25] J. Rixen and M. Renz, "Qdpn-quasi-dual-path network for single-channel speech separation," in *Proc. of Interspeech*, 2022, pp. 5353–5357.
- [26] J. Q. Yip, S. Zhao, Y. Ma, C. Ni, C. Zhang, H. Wang, T. H. Nguyen, K. Zhou, D. Ng, E. S. Chng et al., "Spqm: Prioritizing local features for enhanced speech separation performance," in *Proc. of ICASSP*, 2024, pp. 326–330.
- [27] S. Zhao and B. Ma, "Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions," in *Proc. of ICASSP*, 2023, pp. 1–5.

- [28] S. Zhao, Y. Ma, C. Ni, C. Zhang, H. Wang, T. H. Nguyen, K. Zhou, J. Q. Yip, D. Ng, and B. Ma, "Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation," in *Proc. of ICASSP*, 2024, pp. 10356–10360.
- [29] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Proc. of Interspeech*, 2019, pp. 2728–2732.
- [30] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černocký, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.
- [31] M. Delcroix, T. Ochiai, K. Žmolíková, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *Proc. of ICASSP*, 2020, pp. 691–695.
- [32] Y. Hao, J. Xu, J. Shi, P. Zhang, L. Qin, and B. Xu, "A unified framework for low-latency speaker extraction in cocktail party environments," in *Proc. of Interspeech*, 2020, pp. 1431–1435.
- [33] T. Li, Q. Lin, Y. Bao, and M. Li, "Atss-Net: Target Speaker Separation via Attention-Based Neural Network," in *Proc. of Interspeech*, 2020, pp. 1411–1415.
- [34] Z. Zhang, B. He, and Z. Zhang, "X-tasnet: Robust and accurate time-domain speaker extraction network," in *Proc. of Interspeech*, 2020.
- [35] C. Xu, W. Rao, E. S. Chng, and H. Li, "Spex: Multi-scale time domain speaker extraction network," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1370–1384, 2020.
- [36] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network," in *Proc. of Interspeech*, 2020, pp. 1406–1410.
- [37] W. Wang, C. Xu, M. Ge, and H. Li, "Neural speaker extraction with speaker-speech cross-attention network," in *Proc. of Interspeech*, 2021, pp. 3535–3539.
- [38] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Multi-stage speaker extraction with utterance and frame-level reference signals," in *Proc. of ICASSP*, 2021, pp. 6109–6113.
- [39] K. Liu, Z. Du, X. Wan, and H. Zhou, "X-seperformer: End-to-end speaker extraction network with explicit optimization on speaker confusion," in *Proc. of ICASSP*, 2023, pp. 1–5.
- [40] F. Hao, X. Li, and C. Zheng, "X-tf-gridnet: A time–frequency domain target speaker extraction network with adaptive speaker embedding fusion," *Information Fusion*, vol. 112, p. 102550, 2024.
- [41] M. Elminshawi, W. Mack, S. R. Chetupalli, S. Chakrabarty, and E. A. Habets, "New insights on target speaker extraction," *arXiv preprint arXiv:2202.00733*, 2022.
- [42] B. Zeng, H. Suo, Y. Wan, and M. Li, "Sef-net: Speaker embedding free target speaker extraction network," in *Proc. of Interspeech*, 2023, pp. 3452–3456.
- [43] Y. Hu, H. Xu, Z. Guo, H. Huang, and L. He, "Smma-net: An audio clue-based target speaker extraction network with spectrogram matching and mutual attention," in *Proc. of ICASSP*, 2024, pp. 1496–1500.
- [44] X. Yang, C. Bao, J. Zhou, and X. Chen, "Target speaker extraction by directly exploiting contextual information in the time-frequency domain," in *Proc. of ICASSP*, 2024, pp. 10476–10480.
- [45] B. Zeng and M. Li, "Usef-tse: Universal speaker embedding free target speaker extraction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 2110–2124, 2025.
- [46] P. Wang, K. Tan *et al.*, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 39–48, 2019.
- [47] N. Kamo, M. Delcroix, and T. Nakatani, "Target speech extraction with conditional diffusion model," in *Proceeding of Interspeech*, 2023, pp. 176–180.
- [48] H. Erdogan, S. Wisdom, X. Chang, Z. Borsos, M. Tagliasacchi, N. Zeghidour, and J. R. Hershey, "Tokensplit: Using discrete speech representations for direct, refined, and transcript-conditioned speech separation and recognition," in *Proceeding of Interspeech*, 2023, pp. 3462–3466.
- [49] J. Zhang, J. Yang, Z. Fang, Y. Wang, Z. Zhang, Z. Wang, F. Fan, and Z. Wu, "Anyenhance: A unified generative model with prompt-guidance and self-critic for voice enhancement," *arXiv preprint arXiv:2501.15417*, 2025.
- [50] B. Kang, X. Zhu, Z. Zhang, Z. Ye, M. Liu, Z. Wang, Y. Zhu, G. Ma, J. Chen, L. Xiao *et al.*, "Llase-g1: Incentivizing generalization capability for llama-based speech enhancement," *arXiv preprint arXiv:2503.00493*, 2025.
- [51] R. Wang, L. Li, and T. Toda, "Dual-channel target speaker extraction based on conditional variational autoencoder and directional information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1968–1979, 2024.
- [52] B. Tang, B. Zeng, and M. Li, "Tselm: Target speaker extraction using discrete tokens and language models," *arXiv preprint arXiv:2409.07841*, 2024.
- [53] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [54] X. Wang, M. Thakker, Z. Chen, N. Kanda, S. E. Eskimez, S. Chen, M. Tang, S. Liu, J. Li, and T. Yoshioka, "Speechx: Neural codec language model as a versatile speech transformer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [55] B. Tang, B. Zeng, and M. Li, "Lauratse: Target speaker extraction using auto-regressive decoder-only language models," *arXiv preprint arXiv:2504.07402*, 2025.
- [56] Q. Wang, I. L. Moreno, M. Saglam, K. Wilson, A. Chiao, R. Liu, Y. He, W. Li, J. Pelecanos, M. Nika *et al.*, "Voicefilter-lite: Streaming targeted voice separation for on-device speech recognition," *arXiv preprint arXiv:2009.04323*, 2020.
- [57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of ICPR*, 2016, pp. 770–778.
- [58] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Proc. of Interspeech*, 2020.
- [59] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. of ICASSP*, 2018, pp. 4879–4883.
- [60] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. of ICPR*, 2019, pp. 4690–4699.
- [61] X. Xiao, Z. Chen, T. Yoshioka, H. Erdogan, C. Liu, D. Dimitriadis, J. Droppo, and Y. Gong, "Single-channel speech extraction using speaker inventory and attention network," in *Proc. of ICASSP*, 2019, pp. 86–90.
- [62] L. Yang, W. Liu, L. Tan, J. Yang, and H.-G. Moon, "Target speaker extraction with ultra-short reference speech by ve-ve framework," in *Proc. of ICASSP*, 2023, pp. 1–5.
- [63] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. of ICASSP*, 2018, pp. 5329–5333.
- [64] A. Li, G. Yu, Z. Xu, C. Fan, X. Li, and C. Zheng, "Tabe: Decoupling spatial and spectral processing with taylor's unfolding method in the beamspace domain for multi-channel speech enhancement," *Information Fusion*, vol. 101, p. 101976, 2024.
- [65] R. Scheibler, Y. Ji, S.-W. Chung, J. Byun, S. Choe, and M.-S. Choi, "Diffusion-based generative speech source separation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [66] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2351–2364, 2023.
- [67] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Ieee, 2022, pp. 7402–7406.
- [68] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2724–2737, 2023.
- [69] N. Kamo, M. Delcroix, and T. Nakatani, "Target speech extraction with conditional diffusion model," in *Interspeech 2023*, 2023, pp. 176–180.
- [70] L. Zhang, Y. Qian, L. Yu, H. Wang, H. Yang, S. Liu, L. Zhou, and Y. Qian, "Ddts: Discriminative diffusion model for target speech extraction," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 294–301.
- [71] R. Wang, L. Li, and T. Toda, "Target speaker extraction based on conditional variational autoencoder and directional information in underdetermined condition," *IEICE Technical Report; IEICE Tech. Rep.*, vol. 121, no. 383, pp. 76–81, 2022.
- [72] ———, "Dual-channel target speaker extraction based on conditional variational autoencoder and directional information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1968–1979, 2024.

- [73] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.
- [74] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [75] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.
- [76] Z. Du, S. Zhang, K. Hu, and S. Zheng, “Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec,” in *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 591–595.
- [77] Z. Du, J. Wang, Q. Chen, Y. Chu, Z. Gao, Z. Li, K. Hu, X. Zhou, J. Xu, Z. Ma *et al.*, “Laurapt: Listen, attend, understand, and regenerate audio with gpt,” *arXiv preprint arXiv:2310.04673*, 2023.
- [78] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented transformer for speech recognition,” in *Proceeding of Interspeech*, 2020, pp. 5036–5040.
- [79] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [80] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, “Librimix: An open-source dataset for generalizable speech separation,” *arXiv preprint arXiv:2005.11262*, 2020.
- [81] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “Sdr-half-baked or well done?” in *Proc. of ICASSP*, 2019, pp. 626–630.
- [82] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. of ICLR*, San Diego, CA, USA, 2015.
- [83] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 2001, pp. 749–752.
- [84] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4214–4217.
- [85] C. K. Reddy, V. Gopal, and R. Cutler, “Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” in *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 886–890.
- [86] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, “Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets,” in *Proceeding of Interspeech*, 2021, pp. 2127–2131.
- [87] T. Saeki, S. Maiti, S. Takamichi, S. Watanabe, and H. Saruwatari, “Speechbertscore: Reference-aware automatic evaluation of speech generation leveraging nlp evaluation metrics,” in *Proceeding of Interspeech*, 2024, pp. 4943–4947.
- [88] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.
- [89] Z.-Q. Wang, H. Erdogan, S. Wisdom, K. Wilson, D. Raj, S. Watanabe, Z. Chen, and J. R. Hershey, “Sequential multi-frame neural beamforming for speech separation and enhancement,” in *Proceeding of IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 905–911.
- [90] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *Proceeding of International conference on machine learning*, 2023, pp. 28492–28518.
- [91] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, “Wespeaker: A research and production oriented speaker embedding learning toolkit,” in *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [92] Y. Bengio, N. Léonard, and A. C. Courville, “Estimating or propagating gradients through stochastic neurons for conditional computation,” *CoRR*, vol. abs/1308.3432, 2013.
- [93] S. Wang, K. Zhang, S. Lin, J. Li, X. Wang, M. Ge, J. Yu, Y. Qian, and H. Li, “Wesep: A scalable and flexible toolkit towards generalizable target speaker extraction,” in *Proceeding of Interspeech*, 2024, pp. 4273–4277.