SepALM: Audio Language Models Are Error Correctors for Robust Speech Separation

Zhaoxi Mu , **Xinyu Yang*** and **Gang Wang**
Xi'an Jiaotong University
{wsmzxxh, wanggang911}@stu.xjtu.edu.cn, yxyphd@mail.xjtu.edu.cn

Abstract

While contemporary speech separation technologies adeptly process lengthy mixed audio waveforms, they are frequently challenged by the intricacies of real-world environments, including noisy and reverberant settings, which can result in artifacts or distortions in the separated speech. To overcome these limitations, we introduce SepALM, a pioneering approach that employs audio language models (ALMs) to rectify and re-synthesize speech within the text domain following preliminary separation. SepALM comprises four core components: a separator, a corrector, a synthesizer, and an aligner. By integrating an ALM-based end-to-end error correction mechanism, we mitigate the risk of error accumulation and circumvent the optimization hurdles typically encountered in conventional methods that amalgamate automatic speech recognition (ASR) with large language models (LLMs). Additionally, we have developed Chain-of-Thought (CoT) prompting and knowledge distillation techniques to facilitate the reasoning and training processes of the ALM. Our experiments substantiate that SepALM not only elevates the precision of speech separation but also markedly bolsters adaptability in novel acoustic environments.

1 Introduction

Speech separation, also known as the cocktail party problem, involves isolating individual speech sources from a mixture of audio signals. Prevailing state-of-the-art techniques in speech separation are predicated on a time-domain dual-path methodology [Luo *et al.*, 2020; Subakan *et al.*, 2021; Mu *et al.*, 2023b], characterized by an encoder-dual-path separation network-decoder framework that adeptly manages lengthy mixed audio waveforms. Despite these advancements, existing methods falter when tasked with separating speech from recordings captured in complex real-world acoustics, such as those rife with noise and reverberation, where residual artifacts or distortions persist in the output. Recent endeavours have sought to surmount this challenge through multi-stage processing [Li *et al.*, 2021; Mu *et al.*, 2023a; Neri and Braun, 2023] and the use of generative models [Chen *et al.*, 2020; Wang *et al.*, 2024a]. Yet, the efficacy of these approaches in practical scenarios is often constrained, predominantly owing to the heterogeneity of real-world interference that transcends the scope of typical training datasets.

To tackle this issue, we propose a novel approach that involves correcting and re-synthesizing preliminary separated speech, which may contain artifacts and distortions, within the text domain, as depicted in Figure 1. The underlying insight behind our decision to perform error correction in the text domain rather than the conventional speech domain is that speech is *high-resolution* data, while text is *low-resolution* data, thus facilitating error correction in a simpler form. Additionally, this textual correction is less susceptible to various types of noise, thereby enhancing the model's adaptability to novel acoustic disturbances.

To obtain transcriptions of the preliminary separated speech and to correct these transcriptions, we propose leveraging the capabilities of large language models (LLMs). These models have recently showcased exceptional proficiency in logical reasoning and language generation, leading to significant achievements and swift progress across various natural language processing tasks [OpenAI, 2023; Touvron *et al.*, 2023]. Trained on extensive textual corpora, LLMs exhibit a robust grasp of world knowledge and nuanced contextual understanding. This has prompted exploration into the application of LLMs for automatic speech recognition (ASR) [Lakomkin *et al.*, 2024; Fathullah *et al.*, 2024] and the subsequent refinement of ASR outputs [Radhakrishnan *et al.*, 2023; Chen *et al.*, 2023]. However, we have identified several challenges associated with employing a cascade of ASR models and LLMs for error correction. Initially, LLMs can only assess potential errors based on the context provided by the N-best hypotheses decoded by ASR models, lacking the capacity to perceive and utilize the original speech information. This limitation can lead to grammatically coherent yet contextually inaccurate outcomes [Hu *et al.*, 2024]. Furthermore, the cascade of ASR models and LLMs escalates computational expenses and the risk of error accumulation, while incompatibilities between the models can complicate optimization and diminish expressive capability.

In light of these insights, we have turned to audio language

*Corresponding author

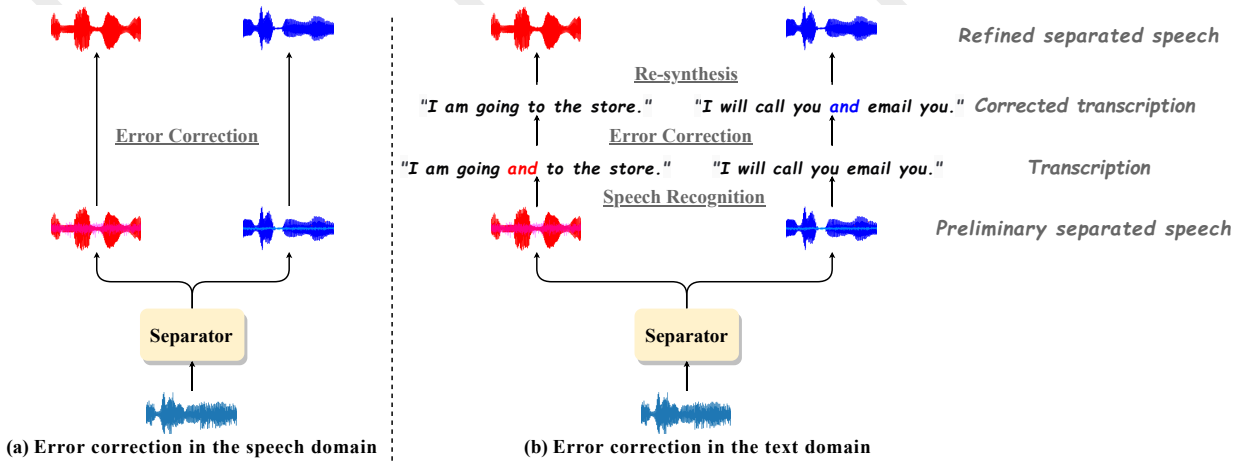

Figure 1: The illustration contrasts (a) the conventional method of error correction for preliminary separated speech in the audio domain, and (b) our proposed method of error correction for preliminary separated speech in the text domain.

models (ALMs) [Latif *et al.*, 2023], which are derived from LLMs and exhibit remarkable capacity in processing and generating both textual and auditory information. We advocate for a streamlined, end-to-end error correction strategy utilizing a single ALM. This approach presents two key benefits. First, it capitalizes on the ALM's cross-modal competencies to incorporate original speech data, thereby enhancing the correction process. Second, by integrating a solitary ALM, we alleviate the complexity associated with optimization and inference. To bolster the precision of our single ALM and mitigate potential hallucinations, we have adopted Chain-of-Thought (CoT) prompting [Wei *et al.*, 2022]. This technique bifurcates the error correction into two phases, enhancing the ALM's reasoning capabilities. Given the paucity of annotated data, we have crafted a knowledge distillation technique that harnesses a pre-trained ASR model as a teacher to direct the ALM's training. Our empirical findings suggest that this CoT strategy rivals the efficacy of more intricate cascaded systems.

Upon securing the corrected transcription, we utilize it alongside the preliminary separated speech to re-synthesize the refined speech, diverging from the conventional use of transcription as a conditional input for speech separation. This decision stems from the recognition that straightforward feature fusion techniques that incorporate transcription information as a condition may precipitate modality imbalance issues [Mu and Yang, 2024], where the textual modality is prone to ineffectiveness. We experimented with two speech synthesis methods based on the neural codec language model[1]: autoregressive (AR) generation [Borsos *et al.*, 2023a; Wang *et al.*, 2023] and non-autoregressive (NAR) masked generation [Kharitonov *et al.*, 2023; Borsos *et al.*, 2023b] to resynthesize the refined speech. In our approach, the transcription and the preliminary separated speech are treated equally, being encoded into tokens that serve as inputs to the language model, thereby mitigating the modality imbalance issue. Additionally, we capitalize on the generative model's ability to effectively learn the distribution of clear speech as a refined prior, enhancing the model's generalization capacity when confronted with novel acoustic disturbances.

Due to the lack of phase information during the re-synthesis of refined speech, the synthesized signal may undergo phase shifts over time, which can diminish the precision of metrics evaluated on a time-sample basis, such as the scale-invariant signal-to-noise ratio (SI-SNR) [Roux *et al.*, 2019]. To mitigate this issue, we implement a realignment process for the refined speech against the preliminary separated speech within the time-frequency domain.

Our contributions are summarized as follows:

- We propose a novel speech separation paradigm that harnesses the prowess of the audio language model to rectify and re-synthesize preliminary separated speech within the text domain, thereby enhancing the system's ability to separate noisy mixed audio.
- We advocate for a streamlined, single ALM-based end-to-end error correction mechanism, circumventing the error accumulation and optimization challenges inherent in the traditional sequential integration of ASR models with LLMs. To further enhance the ALM's reasoning and training processes, we have developed techniques of CoT prompting and knowledge distillation.
- We utilize a speech synthesis method based on the neural codec language model to re-synthesize the refined speech, effectively neutralizing modality imbalance issues and bolstering the model's generalization capability. Furthermore, we incorporate time-frequency domain alignment techniques to resolve phase shift issues, thereby markedly elevating objective evaluation metrics.

2 Related Work

Recent advancements have led to significant improvements in speech separation performance [Luo *et al.*, 2020; Subakan *et*

[1]This can also be referred to as the audio language model. However, for the sake of differentiation from the previously referenced audio language model, we term this the speech synthesis model based on the neural codec language model.

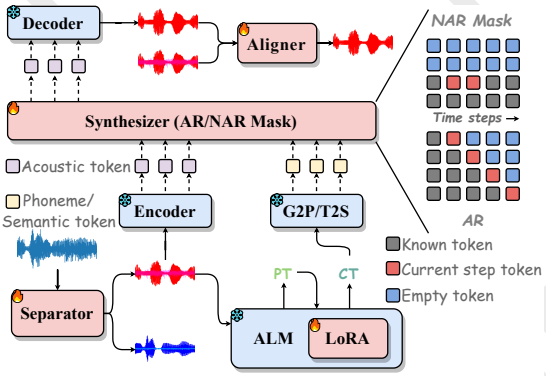

Figure 2: The structural framework of our proposed SepALM. For clarity, the diagram illustrates the processing flow for a single separated speech signal. In the diagram, *'PT'* signifies Preliminary Transcription, and *'CT'* denotes Corrected Transcription.

al., 2021]. Despite these strides, existing models frequently struggle to maintain separation quality under more demanding acoustic scenarios, such as the presence of more intrusive noise types or diminished signal-to-noise ratios. To surmount this challenge, some studies have put forth generative correction techniques to refine the separated speech further [Lutati *et al.*, 2024; Erdogan *et al.*, 2023; Hirano *et al.*, 2023; Wang *et al.*, 2024a]. The underpinning logic of these techniques is that generative models aim to learn the prior distribution of data, can approximate intricate data distributions, and typically exhibit superior generalization capabilities, resulting in the production of more natural and higher-quality samples. Notable approaches similar to ours include Separate and Diffuse [Lutati *et al.*, 2024] and TokenSplit [Erdogan *et al.*, 2023]. The method proposed in Separate and Diffuse [Lutati *et al.*, 2024] amalgamates the strengths of deterministic models with those of stochastic generative models to augment the efficacy of speech separation and perform linear combinations in the time-frequency domain. In TokenSplit [Erdogan *et al.*, 2023], the authors propose a scheme for predicting enhanced audio tokens derived from the speech separated by conventional speech separation models and their transcriptions, aiming to eliminate distortions and artifacts present in the separation estimates.

3 Methodology

3.1 Overall Framework

In this section, we introduce SepALM, a cutting-edge system designed to separate mixed speech and re-synthesize target speech in complex acoustic environments. As depicted in Figure 2, SepALM consists of four principal components: 1) a separator for the preliminary separation of mixed speech; 2) a corrector for error correction of the preliminary separated speech within the low-resolution text domain; 3) a synthesizer that refines the preliminary separated speech by re-synthesizing it based on the corrected text transcription; and 4) an aligner that carries out phase compensation and alignment on the refined speech. Each component will be elaborated upon in the subsequent sections.

3.2 Separator

Given a noisy mixed speech signal $x \in \mathbb{R}^T$, our goal is to estimate C individual speech sources, denoted as $s^{(i)} \in \mathbb{R}^T$ for $i = 1, 2, \ldots, C$. The mixed signal x can be expressed as:

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

Here, $n \in \mathbb{R}^T$ signifies the background noise component, T represents the total number of data points in the signal, and C corresponds to the count of individual speech sources. For the sake of simplicity and without generality being compromised, we assume a scenario with $C = 2$ sources.

We initially engage the widely used time-domain dual-path speech separation network, SepFormer [Subakan *et al.*, 2021], as the separator. This network executes preliminary separation on the noisy mixed speech, yielding preliminary estimates $\hat{s}^{(i)} = f_{sep}(x)$ for $i \in 1, 2$. The separator f_{sep} is trained to directly minimize the discrepancy between the estimated signals $\hat{s}$ and the true signals s. We utilize the SI-SNR loss to optimize the separator's parameters. However, this conventional speech separation approach frequently encounters issues such as over-suppression or under-separation, particularly in noisy environments. These issues can introduce distortions and artifacts in $\hat{s}$. Consequently, additional corrective measures and refinement processes are warranted to enhance the quality of the preliminary separated speech.

3.3 Corrector

In lieu of the prevalent two-stage GER paradigm that incorporates ASR models and LLMs, we posit that employing a singular ALM can yield comparable GER efficacy while simultaneously curtailing the model's computational inference expenses and potential for error accumulation. This hypothesis is substantiated in the experimental section of our study. We harness the pre-trained and fine-tuned ALM, SpeechGPT[2], to rectify the preliminary estimates $\hat{s}$ within the low-resolution text domain. SpeechGPT undergoes training by segmenting speech into 1,000 HuBERT units and integrating them into the LLaMA-7b[3] [Touvron *et al.*, 2023] tokenizer, succeeded by fine-tuning LLaMA-7b to learn cross-modal mappings. Equipped with this multi-modal capability, we can adapt the correction process to the source speech.

Our corrector, denoted as $f_{\text{cor}}(\hat{s}) = \tilde{t}$, can be formulated as a probabilistic model:

$$\tilde{t} = \arg\max_{t} \mathcal{P}_{\text{cor}}(t \mid \hat{s}) \tag{2}$$

Here, $\tilde{t}$ signifies the corrected transcription. To regulate the response quality and mitigate the risk of potential hallucinations within the ALM, we adopt a CoT approach, bifurcating this task into two sub-steps:

(I) The ALM first acts as an ASR model to recognize the speech $\hat{s}$, yielding a preliminary transcription $\hat{t}$. This step is mathematically formulated as:

$$\hat{t} = \arg\max_{t} \mathcal{P}_{\text{asr}}(t \mid \hat{s}) \tag{3}$$

[2]https://huggingface.co/fnlp/SpeechGPT-7B-cm
[3]https://huggingface.co/yahma/llama-7b-hf

(II) Subsequently, the ALM functions as a GER model to refine the preliminary transcription $\hat{t}$ by leveraging the original speech $\hat{s}$, resulting in the corrected transcription $\tilde{t}$. This process can be represented as:

$$\tilde{t} = \arg\max_{t} \mathcal{P}_{\text{ger}}(t \mid \hat{t}, \hat{s}) \tag{4}$$

This methodical approach ensures a structured refinement of the transcription, enhancing accuracy and reliability. To implement this procedure, we have crafted an instructive prompt template as follows:

[Human]: *Please transcribe the following speech input, and then use the speech input to correct any errors in the transcribed text. You can do it step by step. This is the input:* {*speech unit*} <*eoh*>.

[SpeechGPT]: *Preliminary transcription:* {*preliminary transcription*}. *Corrected transcription:* {*corrected transcription*} <*eoa*>.

Notably, unlike prior studies that rely on N-best hypotheses generated by ASR systems, we focus on the top-1 hypothesis, which carries the highest probability and is generally deemed to be of the highest quality. This greedy decoding strategy avoids heavy beam search decoding and rescoring procedures, thereby rendering the decoding process more expeditious and efficient. Experimental results presented in Table 3 corroborate that the ALM with greedy decoding performs comparably to, or even surpasses, the amalgamations of ASR models and LLMs that are deployed with more intricate decoding techniques.

Given the unavailability of ground-truth transcriptions for these two steps, we have devised a knowledge distillation strategy that employs a pre-trained ASR model as a teacher to direct the training of SpeechGPT. Specifically, we engage Whisper [Radford *et al.*, 2023] to execute greedy decoding on both the preliminary separated speech $\hat{s}$ and the true clean speech s, yielding the target transcriptions $\hat{t}^*$ and $\tilde{t}^*$, respectively. Overall, the optimization objectives are encapsulated by the following equations:

$$\mathcal{L}_{\text{asr}} = \sum_{n=1}^{N_1} -\log \mathcal{P}(\hat{t}^*_n \mid \hat{t}^*_{n-1}, \ldots, \hat{t}^*_1, \hat{s}) \tag{5}$$

$$\mathcal{L}_{\text{ger}} = \sum_{n=1}^{N_2} -\log \mathcal{P}(\tilde{t}^*_n \mid \tilde{t}^*_{n-1}, \ldots, \tilde{t}^*_1, \hat{t}, \hat{s}) \tag{6}$$

Here, $\mathcal{L}_{\text{asr}}$ and $\mathcal{L}_{\text{ger}}$ signify the cross-entropy losses associated with the ASR and GER steps, respectively. $\hat{t}^*_n$ and $\tilde{t}^*_n$ correspond to the n-th tokens of the transcriptions $\hat{t}^*$ and $\tilde{t}^*$, respectively. N_1 and N_2 denote the total count of tokens within the transcriptions. Given the substantial scale of the ALM, we employ the parameter-efficient low-rank adaptation (LoRA) technique [Hu *et al.*, 2022] to fine-tune SpeechGPT, thereby reducing computational and memory overhead.

3.4 Synthesizer

Upon acquiring the corrected transcription, our objective is to refine the preliminary separated speech utilizing the corrected transcription. We eschew the conventional approach of re-separating the speech conditioned on the transcription [Rahimi *et al.*, 2022], instead opting for a strategy that involves re-generating the refined speech. This methodology offers dual benefits: First, by affording equal significance to both text and speech as inputs, we circumvent the potential issue of modality imbalance, which might otherwise render the text modality less effective. Second, by employing a generative model to learn the distribution of clear speech as a prior for refinement, rather than training the model to learn a direct mapping from preliminary separated speech to refined speech, we enhance the model's capacity to generalize to novel acoustic environments.

Inspired by recent advances in neural codec language models for speech synthesis [Borsos *et al.*, 2023a], we conceptualize speech synthesis as a conditional language modelling task utilizing neural codec codes, also known as acoustic tokens. Specifically, we utilize a pre-trained neural codec, DAC [Kumar *et al.*, 2023], to tokenize each audio sample into discrete acoustic tokens. We subsequently train a decoder-only language model to generate the acoustic tokens S^* of the refined speech s^* using either an AR or NAR masked generation manner, conditioned on the corrected transcription $\tilde{t}$ and the preliminary separated speech $\hat{s}$, represented by the acoustic tokens $\hat{S}$. DAC is a convolutional residual vector quantization audio codec that features a Q-level quantizer comprising B entries. The preliminary separated speech $\hat{s}$ is discretized and encoded into $\hat{S} \in \mathbb{R}^{2\times T'\times Q}$, where T' denotes the length of the downsampled acoustic tokens.

For the AR generation approach, aligning with methods from prior studies [Borsos *et al.*, 2023a; Wang *et al.*, 2023], we adopt a two-stage modelling strategy. In the first stage, an AR language model generates the acoustic tokens for the first quantizer in an AR fashion. In the second stage, a NAR language model generates the acoustic tokens for the remaining $Q-1$ quantizers in parallel. This amalgamation of AR and NAR generation strategies effectively balances the fidelity of the synthesized speech with the inference speed. The training process can be represented as:

$$\begin{aligned}\mathcal{P}(S \mid \hat{S}, p) = &\prod_{t=0}^{T''} \mathcal{P}_{\text{AR}}(S_{t,1} \mid S_{<t,1}, \hat{S}_{:,1}, p) \\ &\times \prod_{i=2}^{Q} \mathcal{P}_{\text{NAR}}(S_{:,i} \mid S_{:,<i}, \hat{S}, p)\end{aligned} \tag{7}$$

where $S \in \mathbb{R}^{2\times T''\times Q}$ denotes the acoustic tokens of the true target speech s, with T'' indicating the length of S. The notation $S_{:,<i}$ represents the acoustic token layers in S where the layer indices are less than i. The terms $\mathcal{P}_{\text{AR}}$ and $\mathcal{P}_{\text{NAR}}$ represent the AR and NAR generation processes, respectively. We utilize a grapheme-to-phoneme (G2P) tool to convert the corrected transcription $\tilde{t}$ into a phoneme sequence p. p is then concatenated with the acoustic tokens $\hat{S}$ of the preliminary separated speech to form the prefix tokens.

For the NAR masked generation approach, akin to methods employed in previous studies [Kharitonov *et al.*, 2023; Wang *et al.*, 2024b], during training, at each time step t, we randomly select a subset of tokens from the i-th quantizer $S_{:,i}$ of the true target speech s's acoustic tokens S for masking, resulting in $S_{m,i}$. The synthesizer is then trained to generate the

complete target acoustic tokens $S_{:,i}$ in a NAR manner, which can be expressed as:

$$\mathcal{P}_{\text{Mask}}(S_{:,i} \mid S_{m,i}, S_{:,<i}, \hat{S}, P) \tag{8}$$

We utilize a pre-trained text-to-semantic (T2S) model based on w2v-BERT [Chung *et al.*, 2021], sourced from SPEAR-TTS[4], to transform the corrected transcription $\tilde{t}$ into semantic tokens P. Subsequently, the embeddings of the semantic tokens P are added to the embeddings of the acoustic tokens $\hat{S}$ and $S_{:,<i}$ to serve as the condition.

The optimization of the synthesizer is achieved by minimizing the negative log-likelihood objective, which equates to the cross-entropy loss between the generated acoustic tokens and the true acoustic tokens. Ultimately, the refined speech s^* is synthesized utilizing the DAC decoder from the generated acoustic tokens S^*.

3.5 Aligner

The potential for phase shifts in the output generated by the generation method necessitates a subsequent alignment process to ensure the refined speech s^* is accurately aligned with the target. To address this, we integrate the preliminary separated speech $\hat{s}$ and apply phase compensation and alignment to the refined speech s^*. Adhering to the approach outlined in [Lutati *et al.*, 2024], we align the two estimated speech signals through a linear combination in the time-frequency domain, yielding the final aligned refined speech $\tilde{s}$, which can be expressed as:

$$\tilde{s} = \text{iSTFT}(\alpha_1 \odot \text{STFT}(\hat{s}) + \alpha_2 \odot \text{STFT}(s^*)) \tag{9}$$

Here, STFT and iSTFT denote the short-time Fourier transform and its inverse, respectively, while $\odot$ represents the Hadamard product. The weighting coefficients α_1 and α_2 are determined by the aligner A:

$$[\alpha_1, \alpha_2] = A(\text{STFT}(\hat{s}), \text{STFT}(s^*)) \tag{10}$$

A detailed exposition of the alignment procedure is provided in the technical appendix. In congruence with the preliminary separated speech $\hat{s}$, the aligned refined speech $\tilde{s}$ is also evaluated using the SI-SNR as the objective function. This guides the training of the entire model by minimizing the divergence between $\tilde{s}$ and the ground-truth speech s.

4 Experiments

4.1 Datasets

To bolster the generalization capability of our model, we train it using a unified dataset comprising noisy mixed datasets WHAM! [Wichern *et al.*, 2019], WHAMR! [Maciejewski *et al.*, 2020], and Libri2Mix [Cosentino *et al.*, 2020]. WHAM! is a noisy version of WSJ0-2mix [Hershey *et al.*, 2016], incorporating noise samples recorded in environments such as cafes, restaurants, and bars. Building upon WHAM!, WHAMR! introduces reverberation effects to the speech sources, supplementing the pre-existing noise components. Libri2Mix is constructed by simulating noise data from WHAM! and speech segments from Librispeech [Panayotov *et al.*, 2015], and it encompasses two training subsets: train-360 and train-100. All datasets are sampled at a rate of 16 kHz. The model was trained on the complete audio segments from these datasets to produce semantically coherent transcriptions. During training, the audio segments were padded to ensure uniform length. Upon completion of training, we conduct individual evaluations on each dataset's respective test set.

[4]https://github.com/lucidrains/spear-tts-pytorch

4.2 Setup

For the separator, we utilized SepFormer, which comprises 26M parameters. For the corrector, during fine-tuning, we configured the rank of LoRA to 8, integrated LoRA weights into the query, key, value, and output layers of each Transformer block, and trained the newly added LoRA parameters, resulting in a total of 8M trainable parameters. The optimization process was facilitated by the AdamW optimizer with a peak learning rate of $2e^{-4}$. Training proceeded for 5 epochs with a batch size of 128. During inference, we set the maximum sequence length to 1024 and implemented both Top-k and Top-p sampling strategies, with k set to 40 and p to 0.9. The temperature parameter was set to 0.1, and the beam search was configured with a size of 1. We utilized Whisper-Tiny as the teacher model. For the synthesizer, we implemented a Transformer-based model consisting of 12 layers, 16 attention heads, attention dimensions of 1024, and feed-forward network dimensions of 4096, totaling 202M parameters. The AdamW optimizer was also applied here, with a peak learning rate of $5e^{-4}$. For the NAR masked generation method, we set the number of inference steps to 25. During the training of the synthesizer, we employed classifier-free guidance [Ho and Salimans, 2022], randomly discarding prompts with a probability of 0.1. For the neural codec DAC, we configured B and Q to 1024 and 12, respectively. For the aligner A, we employed a two-layer convolutional neural network with residual connections, which has 0.13M parameters. For the entire model, the training strategy commenced with the fine-tuning of SpeechGPT, after which its parameters were frozen while the remaining model components were trained. We employed a permutation-invariant loss function throughout the training process.

4.3 Evaluation Metrics

For objective evaluation, we utilized reference-based perceptual evaluation metrics, encompassing scale-invariant signal-to-noise ratio improvement (SI-SNRi), signal-to-distortion ratio improvement (SDRi), perceptual evaluation of speech quality improvement (PESQi), and extended short-time objective intelligibility improvement (ESTOIi). To quantify speech intelligibility, we measured the word error rate (WER) of the generated audio when transcribed by ASR. We utilized Whisper-Tiny to perform ASR on the separated speech to assess the transcription accuracy. To establish a benchmark, we also applied Whisper-Tiny to the original clean speech, treating the resulting transcription as the true reference. To measure the retention of the target speaker's voice by our re-synthesis-based approach, we utilized a speaker verification system based on WavLM [Chen *et al.*, 2022] to calculate the

Method	Libri2Mix				WHAM!				WHAMR!			
	SI-SNRi	SDRi	NMOS	SMOS	SI-SNRi	SDRi	NMOS	SMOS	SI-SNRi	SDRi	NMOS	SMOS
Conv-TasNet [2019]	12.1	12.5	$3.23_{\pm 0.13}$	$3.01_{\pm 0.21}$	12.7	13.2	$3.32_{\pm 0.30}$	$3.25_{\pm 0.34}$	8.3	7.8	$2.82_{\pm 0.24}$	$3.17_{\pm 0.32}$
DPRNN [2020]	11.3	11.6	$3.21_{\pm 0.28}$	$3.04_{\pm 0.19}$	13.7	14.1	$3.48_{\pm 0.35}$	$3.31_{\pm 0.40}$	10.3	9.7	$3.06_{\pm 0.37}$	$3.29_{\pm 0.16}$
Wavesplit [2021]	15.1	15.8	$3.67_{\pm 0.27}$	$3.28_{\pm 0.42}$	16.0	16.6	$3.50_{\pm 0.40}$	$3.34_{\pm 0.38}$	13.2	12.2	$3.17_{\pm 0.12}$	$3.38_{\pm 0.44}$
SepFormer [2021]	12.9	13.5	$3.52_{\pm 0.37}$	$3.35_{\pm 0.11}$	16.4	16.7	$3.63_{\pm 0.19}$	$3.37_{\pm 0.23}$	14.0	13.0	$3.19_{\pm 0.49}$	$\mathbf{3.45}_{\pm 0.37}$
MossFormer2 [2024]	16.0	16.6	$3.78_{\pm 0.34}$	$\mathbf{3.48}_{\pm 0.29}$	18.1	18.5	$3.88_{\pm 0.13}$	$3.52_{\pm 0.31}$	17.0	15.9	$3.20_{\pm 0.31}$	$3.43_{\pm 0.47}$
MossFormer2* [2024]	16.0	16.5	$3.81_{\pm 0.27}$	$3.45_{\pm 0.42}$	18.2	18.6	$3.92_{\pm 0.21}$	$\mathbf{3.55}_{\pm 0.52}$	17.2	16.0	$3.22_{\pm 0.13}$	$3.37_{\pm 0.25}$
DiffSep† [2023]	8.9	9.5	$3.60_{\pm 0.13}$	$3.08_{\pm 0.26}$	12.4	12.9	$3.67_{\pm 0.34}$	$3.15_{\pm 0.42}$	9.5	8.6	$3.16_{\pm 0.49}$	$3.13_{\pm 0.19}$
SepALM$^{\ddagger *}$ (*AR*)	17.4	17.9	$\mathbf{3.91}_{\pm 0.33}$	$3.29_{\pm 0.28}$	**18.8**	**19.3**	$\mathbf{4.01}_{\pm 0.14}$	$3.43_{\pm 0.33}$	18.1	17.2	$\mathbf{3.44}_{\pm 0.54}$	$3.36_{\pm 0.29}$
SepALM$^{\ddagger *}$ (*Mask*)	**17.6**	**18.2**	$3.86_{\pm 0.08}$	$3.36_{\pm 0.22}$	18.7	19.2	$3.98_{\pm 0.13}$	$3.40_{\pm 0.26}$	**18.2**	**17.4**	$3.37_{\pm 0.45}$	$3.38_{\pm 0.17}$

Table 1: Performance comparison of SepALM with other state-of-the-art speech separation models on the Libri2Mix, WHAM!, and WHAMR! benchmark datasets. The symbol † denotes methods based on generative models, while ‡ signifies methods that integrate both discriminative and generative models. *AR* represents AR generation, and *Mask* signifies NAR masked generation. The superscript $*$ signifies training conducted with the combined dataset of three sources.

Method	MUSAN				DEMAND			
	SI-SNRi ↑	PESQi ↑	ESTOIi ↑	SIM ↑	SI-SNRi ↑	PESQi ↑	ESTOIi ↑	SIM ↑
SepFormer [2021]	9.1	0.50	0.20	**0.65**	9.9	0.67	0.21	0.66
DiffSep† [2023]	8.4	0.41	0.19	0.48	9.5	0.62	0.20	0.55
Refiner‡ [2023]	8.8	0.51	0.20	0.53	9.6	0.60	0.20	0.56
Fast-GeCo‡ [2024a]	12.3	0.75	0.30	0.58	13.3	0.92	0.29	0.61
SepALM‡ (*AR*)	13.7	0.82	0.38	0.63	**14.5**	1.15	0.46	**0.67**
SepALM‡ (*Mask*)	**13.9**	**0.86**	**0.39**	0.62	14.4	**1.19**	**0.47**	0.65

Table 2: Performance comparison of SepALM with other state-of-the-art speech separation models on the MUSAN and DEMAND out-of-domain noise datasets. The symbol † signifies methods based on generative models, while ‡ marks generative correction methods that integrate both discriminative and generative models.

cosine similarity (SIM) between the speaker embeddings of the generated samples and the real audio. Additionally, We report the real-time factor (RTF) for each method when processing 5 seconds of speech on an A100 GPU to compare inference efficiency. For subjective evaluation, we utilized two metrics: the naturalness mean opinion score (NMOS) and the similarity mean opinion score (SMOS), to assess the naturalness and speaker similarity of the separated speech, respectively. A panel of fifteen human evaluators rated 30 randomly selected speech segments on a scale from 1 to 5, where 1 indicated the lowest quality and 5 the highest.

4.4 Comparison with State-of-the-Art

We conducted a comprehensive comparison of our proposed SepALM method against various state-of-the-art techniques, as detailed in Table 1. For previously published results, we reference the original data; otherwise, we present our reproduced outcomes. All baseline methods were trained and evaluated on individual datasets by default. The comparative analysis reveals that SepALM, whether based on AR or NAR masked generation, consistently outperforms baseline methods in most evaluation metrics. This highlights the efficacy of our integrated approach of separation, correction, synthesis, and alignment in processing noisy mixed audio.

Furthermore, it is observed that speech separation methods based on generative models, such as DiffSep [Scheibler *et al.*, 2023], tend to underperform on objective metrics, including simulated exact reconstruction and signal-to-noise ratio (SNR), as well as on the subjective metric SMOS, which measures speaker similarity. However, these methods excel in the subjective metric NMOS, which assesses naturalness. This discrepancy arises because, under conditions of high distortion, multiple distinct clean samples may yield identical observed samples post-distortion, compounded by the inherent sampling randomness of generative models. Our method combines the strengths of discriminative and generative models, refining the generative output through alignment to mitigate the non-positive definite issue, thus achieving robust performance across both objective and subjective metrics. In addition, the results presented in rows 5 and 6 of Table 1 demonstrate that a simple increase in the quantity of training data did not lead to a noticeable improvement in the performance of MossFormer. This finding suggests that the performance gains observed in our method are predominantly attributed to the error correction and re-synthesis processes we employ.

Evaluation on out-of-domain data. To assess the generalization capability of our method when faced with unseen noise types, similar to the WHAM! dataset, we utilized the test set of WSJ0-2mix as the speech source and noise audio from the MUSAN [Snyder *et al.*, 2015] and DEMAND [Hadad *et al.*, 2014] datasets as the noise source to simulate out-of-domain noisy mixed audio. Noise is introduced by randomly sampling SNR values from a uniform distribution that spans from -6 to +3 dB. All evalu-

Exp.	Method	SI-SNRi ↑	SDRi ↑	NMOS ↑	SMOS ↑	WER (%)↓	RTF ↓
(a)	Re-separation model	14.5	15.1	$3.64_{\pm 0.52}$	$3.31_{\pm 0.14}$	3.76/4.79	1.83
(b)	SepALM (*AR*)	17.4	17.9	$3.91_{\pm 0.33}$	$3.29_{\pm 0.28}$	3.76/4.10	2.58
(c)	SepALM (*Mask*)	**17.6**	**18.2**	$3.86_{\pm 0.08}$	$\mathbf{3.36}_{\pm 0.22}$	3.76/4.03	1.91
(d)	- Separator only	13.2	13.8	$3.62_{\pm 0.27}$	$3.35_{\pm 0.12}$	−/5.68	**0.52**
(e)	- Cascaded method	17.4	17.9	$3.93_{\pm 0.29}$	$3.32_{\pm 0.30}$	3.85/4.32	2.95
(f)	- *w/o* Fine-tuning	16.3	16.8	$3.64_{\pm 0.45}$	$3.28_{\pm 0.45}$	4.56/4.96	2.95
(g)	- *w/o* Aligner	15.7	16.2	$\mathbf{3.98}_{\pm 0.19}$	$3.34_{\pm 0.42}$	3.76/**3.99**	1.89
(h)	- *w/o* Fine-tuning	17.2	17.8	$3.79_{\pm 0.24}$	$3.30_{\pm 0.19}$	3.93/4.25	1.91

Table 3: Ablation study on the Libri2Mix dataset. The first and second WER values correspond to the word error rates of the corrected transcriptions and the model's output speech, respectively.

ated methods were trained on the Libri2Mix dataset. The results, as depicted in Table 2, indicate that compared to other speech separation methods that apply generative correction directly in the speech domain [Hirano *et al.*, 2023; Wang *et al.*, 2024a], our text-domain error correction approach exhibits superior generalization performance in out-of-domain noise scenarios.

4.5 Ablation Study

In this section, we validate the efficacy of each key design within our method through ablation studies. All models were trained on a unified training set comprising the training sets from WHAM!, WHAMR!, and Libri2Mix and evaluated on the Libri2Mix test set.

Ablation study on the corrector. Initially, we aimed to substantiate the merit of our proposed text-domain error correction by excluding all elements bar the separator. The outcomes are delineated in Exp. (d) of Table 3, revealing that the correction process significantly improves the quality of the separated speech and reduces the WER. To further validate the feasibility of employing a standalone ALM for GER, we devised a two-stage cascaded model for comparison. Specifically, following Whispering-LLaMA [Radhakrishnan *et al.*, 2023], we deployed Whisper-Tiny to decode the preliminary separated speech into 3-best hypotheses and then employed LLaMA-7B to perform GER on these hypotheses. The corrected transcriptions were subsequently employed to re-synthesize the separated speech. The training regimen mirrored that of our method. As depicted in Exp. (e) of Table 3, our CoT reasoning-based ALM approach rivals the cascaded paradigm using more intricate decoding techniques and even evinces minor advantages in certain metrics, with a lower RTF. This underscores the superior efficiency of our method compared to the more intricate cascaded paradigm. Furthermore, our method remains effective even without fine-tuning the ALM, as evidenced in Exp. (h) of Table 3. This effectiveness is primarily attributed to the CoT prompts we employ, which enhance the zero-shot reasoning capabilities of the ALM. This feature is absent in traditional two-stage cascaded systems, as illustrated in Exp. (f) of Table 3.

Ablation study on the synthesizer. To verify the efficacy of the synthesis process within our method, we substituted the synthesizer with a re-separation module analogous to the separator, employing the corrected transcriptions as a condition. This is akin to target speech extraction (TSE) methods [Zmolíková *et al.*, 2023; Mu *et al.*, 2024] under given conditions. Specifically, we employed a G2P tool to convert the text into phoneme sequences, which were subsequently mapped onto a series of learnable embedding vectors. We utilized the audio embeddings from the intermediate layers of the SepFormer as the value and key, while the phoneme embeddings served as the query for feature fusion based on cross-attention. The outcomes are presented in Exp. (a) of Table 3, revealing that our re-synthesis method surpasses the re-separation method. We attribute this superiority to two principal factors. Initially, for source separation tasks, the performance upper bound of generative models is higher than that of deterministic models [Lutati *et al.*, 2024], thereby bolstering the model's generalization capabilities. Second, by harnessing a neural codec language model for speech synthesis, we positioned the corrected transcriptions and the preliminary separated speech on an equal footing, effectively neutralizing the modality imbalance issue. Furthermore, comparing the results of Exp. (b) and (c) in Table 3 reveals that the NAR masked generation method outperforms the AR generation method in terms of generation quality and inference speed. This superiority is largely due to its bidirectional attention to the full context and its parallel generation approach.

Ablation study on the aligner. To assess the contribution of the alignment phase, we omitted the aligner from our process and assessed its impact. The findings, as depicted in Exp. (g) of Table 3, indicate that including the aligner markedly improves objective metrics calculated on time samples, such as SI-SNRi. However, it has a minor detrimental effect on the naturalness (NMOS) of the separated speech. Nonetheless, this trade-off is deemed justifiable given the overall enhancement in the quality of the separated speech.

5 Conclusion

This paper introduces SepALM, a novel approach that capitalizes on the prowess of audio language models to correct errors in the preliminary separated speech within the text domain and subsequently re-synthesize it. Our method has been shown to significantly enhance the separation of noisy mixed audio. Ablation studies further substantiated the efficacy of each key component of SepALM. Collectively, our findings highlight the potential of the speech separation-correction-synthesis-alignment paradigm within the text domain, offering new avenues for future research on speech separation under more intricate acoustic conditions.

References

[Borsos *et al.*, 2023a] Zalán Borsos, Raphaël Marinier, Damien Vincent, et al. Audiolm: A language modeling approach to audio generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2523–2533, 2023.

[Borsos *et al.*, 2023b] Zalán Borsos, Matthew Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. Soundstorm: Efficient parallel audio generation. *CoRR*, abs/2305.09636, 2023.

[Chen *et al.*, 2020] Jingjing Chen, Qirong Mao, and Dong Liu. On synthesis for supervised monaural speech separation in time domain. In *INTERSPEECH*, pages 2627–2631. ISCA, 2020.

[Chen *et al.*, 2022] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518, 2022.

[Chen *et al.*, 2023] Chen Chen, Yuchen Hu, Chao-Han Huck Yang, et al. Hyporadise: An open baseline for generative speech recognition with large language models. In *NeurIPS*, 2023.

[Chung *et al.*, 2021] Yu-An Chung, Yu Zhang, Wei Han, et al. w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *ASRU*, pages 244–250. IEEE, 2021.

[Cosentino *et al.*, 2020] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. Librimix: An open-source dataset for generalizable speech separation. *CoRR*, abs/2005.11262, 2020.

[Erdogan *et al.*, 2023] Hakan Erdogan, Scott Wisdom, Xuankai Chang, et al. Tokensplit: Using discrete speech representations for direct, refined, and transcript-conditioned speech separation and recognition. In *INTERSPEECH*, pages 3462–3466. ISCA, 2023.

[Fathullah *et al.*, 2024] Yassir Fathullah, Chunyang Wu, Egor Lakomkin, et al. Prompting large language models with speech recognition abilities. In *ICASSP*, pages 13351–13355. IEEE, 2024.

[Hadad *et al.*, 2014] Elior Hadad, Florian Heese, Peter Vary, and Sharon Gannot. Multichannel audio database in various acoustic environments. In *IWAENC*, pages 313–317. IEEE, 2014.

[Hershey *et al.*, 2016] John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *ICASSP*, pages 31–35. IEEE, 2016.

[Hirano *et al.*, 2023] Masato Hirano, Kazuki Shimada, Yuichiro Koyama, Shusuke Takahashi, and Yuki Mitsufuji. Diffusion-based signal refiner for speech separation. *CoRR*, abs/2305.05857, 2023.

[Ho and Salimans, 2022] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022.

[Hu *et al.*, 2022] Edward J. Hu, Yelong Shen, Phillip Wallis, et al. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net, 2022.

[Hu *et al.*, 2024] Yuchen Hu, Chen Chen, Chengwei Qin, et al. Listen again and choose the right answer: A new paradigm for automatic speech recognition with large language models. In *ACL (Findings)*, pages 666–679. Association for Computational Linguistics, 2024.

[Kharitonov *et al.*, 2023] Eugene Kharitonov, Damien Vincent, Zalán Borsos, et al. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Trans. Assoc. Comput. Linguistics*, 11:1703–1718, 2023.

[Kumar *et al.*, 2023] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, et al. High-fidelity audio compression with improved RVQGAN. In *NeurIPS*, 2023.

[Lakomkin *et al.*, 2024] Egor Lakomkin, Chunyang Wu, Yassir Fathullah, et al. End-to-end speech recognition contextualization with large language models. In *ICASSP*, pages 12406–12410. IEEE, 2024.

[Latif *et al.*, 2023] Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Heriberto Cuayáhuitl, and Björn W. Schuller. Sparks of large audio models: A survey and outlook. *CoRR*, abs/2308.12792, 2023.

[Li *et al.*, 2021] Andong Li, Wenzhe Liu, Xiaoxue Luo, Guochen Yu, Chengshi Zheng, and Xiaodong Li. A simultaneous denoising and dereverberation framework with target decoupling. In *Interspeech*, pages 2801–2805. ISCA, 2021.

[Luo and Mesgarani, 2019] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 27(8):1256–1266, 2019.

[Luo *et al.*, 2020] Yi Luo, Zhuo Chen, and Takuya Yoshioka. Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation. In *ICASSP*, pages 46–50. IEEE, 2020.

[Lutati *et al.*, 2024] Shahar Lutati, Eliya Nachmani, and Lior Wolf. Separate and diffuse: Using a pretrained diffusion model for better source separation. In *ICLR*. OpenReview.net, 2024.

[Maciejewski *et al.*, 2020] Matthew Maciejewski, Gordon Wichern, Emmett McQuinn, and Jonathan Le Roux. Whamr!: Noisy and reverberant single-channel speech separation. In *ICASSP*, pages 696–700. IEEE, 2020.

[Mu and Yang, 2024] Zhaoxi Mu and Xinyu Yang. Separate in the speech chain: Cross-modal conditional audio-visual target speech extraction. In *IJCAI*, pages 6415–6423. ijcai.org, 2024.

[Mu *et al.*, 2023a] Zhaoxi Mu, Xinyu Yang, Xiangyuan Yang, and Wenjing Zhu. A multi-stage triple-path method for speech separation in noisy and reverberant environments. In *ICASSP*, pages 1–5. IEEE, 2023.

[Mu *et al.*, 2023b] Zhaoxi Mu, Xinyu Yang, and Wenjing Zhu. Multi-dimensional and multi-scale modeling for

speech separation optimized by discriminative learning. In *ICASSP*, pages 1–5. IEEE, 2023.

[Mu *et al.*, 2024] Zhaoxi Mu, Xinyu Yang, Sining Sun, and Qing Yang. Self-supervised disentangled representation learning for robust target speech extraction. In *AAAI*, pages 18815–18823. AAAI Press, 2024.

[Neri and Braun, 2023] Julian Neri and Sebastian Braun. Towards real-time single-channel speech separation in noisy and reverberant environments. In *ICASSP*, pages 1–5. IEEE, 2023.

[OpenAI, 2023] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[Panayotov *et al.*, 2015] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *ICASSP*, pages 5206–5210. IEEE, 2015.

[Radford *et al.*, 2023] Alec Radford, Jong Wook Kim, Tao Xu, et al. Robust speech recognition via large-scale weak supervision. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 2023.

[Radhakrishnan *et al.*, 2023] Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, et al. Whispering llama: A cross-modal generative error correction framework for speech recognition. In *EMNLP*, pages 10007–10016. Association for Computational Linguistics, 2023.

[Rahimi *et al.*, 2022] Akam Rahimi, Triantafyllos Afouras, and Andrew Zisserman. Reading to listen at the cocktail party: Multi-modal speech separation. In *CVPR*, pages 10483–10492. IEEE, 2022.

[Roux *et al.*, 2019] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. SDR - half-baked or well done? In *ICASSP*, pages 626–630. IEEE, 2019.

[Scheibler *et al.*, 2023] Robin Scheibler, Youna Ji, Soo-Whan Chung, Jaeuk Byun, Soyeon Choe, and Min-Seok Choi. Diffusion-based generative speech source separation. In *ICASSP*, pages 1–5. IEEE, 2023.

[Snyder *et al.*, 2015] David Snyder, Guoguo Chen, and Daniel Povey. MUSAN: A music, speech, and noise corpus. *CoRR*, abs/1510.08484, 2015.

[Subakan *et al.*, 2021] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *ICASSP*, pages 21–25. IEEE, 2021.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.

[Wang *et al.*, 2023] Chengyi Wang, Sanyuan Chen, Yu Wu, et al. Neural codec language models are zero-shot text to speech synthesizers. *CoRR*, abs/2301.02111, 2023.

[Wang *et al.*, 2024a] Helin Wang, Jesus Villalba, Laureano Moro-Velazquez, Jiarui Hai, Thomas Thebaud, and Najim Dehak. Noise-robust speech separation with fast generative correction. *CoRR*, abs/2406.07461, 2024.

[Wang *et al.*, 2024b] Yuancheng Wang, Haoyue Zhan, Liwei Liu, et al. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *CoRR*, abs/2409.00750, 2024.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

[Wichern *et al.*, 2019] Gordon Wichern, Joe Antognini, Michael Flynn, et al. Wham!: Extending speech separation to noisy environments. In *INTERSPEECH*, pages 1368–1372. ISCA, 2019.

[Zeghidour and Grangier, 2021] Neil Zeghidour and David Grangier. Wavesplit: End-to-end speech separation by speaker clustering. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:2840–2849, 2021.

[Zhao *et al.*, 2024] Shengkui Zhao, Yukun Ma, Chongjia Ni, et al. Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation. In *ICASSP*, pages 10356–10360. IEEE, 2024.

[Zmolíková *et al.*, 2023] Katerina Zmolíková, Marc Delcroix, Tsubasa Ochiai, Keisuke Kinoshita, Jan Cernocký, and Dong Yu. Neural target speech extraction: An overview. *IEEE Signal Process. Mag.*, 40(3):8–29, 2023.