ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446418

GENERATION-BASED TARGET SPEECH EXTRACTION WITH SPEECH DISCRETIZATION AND VOCODER

Linfeng Yu, Wangyou Zhang, Chenpeng Du, Leying Zhang, Zheng Liang, Yanmin Qian[†]

Auditory Cognition and Computational Acoustics Lab
MoE Key Lab of Artificial Intelligence, AI Institute
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Target speech extraction (TSE) is a task aiming at isolating the speech of a specific target speaker from an audio mixture, with the help of an auxiliary recording of that target speaker. Most existing TSE methods employ discrimination-based models to estimate the target speaker's proportion in the mixture, but they often fail to compensate for the missing or highly corrupted frequency components in the speech signal. In contrast, the generation-based methods can naturally handle such scenarios via speech resynthesis. In this paper, we propose a novel discrete token based TSE approach by combining state-of-the-art speech discretization and vocoder techniques. By predicting a sequence of discrete tokens with the auxiliary audio and employing a vocoder that takes discrete tokens as input, the target speech can be effectively re-synthesized while eliminating interference. Our experiments conducted on the WSJ0-2mix and Libri2mix datasets demonstrate that our proposed method yields high-quality target speech without interference.

Index Terms— Target speech extraction, speech discretization, speech synthesis, vocoder

1. INTRODUCTION

With the recent development of speech-related intelligent devices and related applications, front-end speech enhancement has become a popular research topic. Speech enhancement aims to obtain clean speech from noisy recordings, which consists of many sub-tasks such as noise reduction, dereverberation, speech separation, and target speech extraction (TSE). In this paper, we focus on the TSE task, which aims to extract the target speaker's speech from multi-talker mixture recordings with the help of the target speaker's voice print or a short recording. This area has made rapid progress with the advent of deep learning. [1, 2, 3, 4].

The TSE models use an additional speaker encoder to help model the information of the target speaker, and a fusion module to embed speaker information into the model. Some methods use a speaker verification model to get the target speaker's embedding, which is then used as the condition of TSE model [5, 6, 7]. Other methods are to train the neural embedding extractor jointly with a speaker encoder [2, 8, 9], the resulting speaker representations are thus directly optimized for TSE tasks. However, most existing TSE methods use the discrimination neural network which is trained to directly map noisy speech to clean speech by optimizing a signal level metric between the enhanced signal and a clean speech reference. This may lead to the inability to compensate for the missing or highly corrupted frequency components.

The vocoder is an important part of the text-to-speech system, which synthesizes raw waveform from the acoustic features. Many vocoders are generative adversarial networks (GAN) based vocoders [10]. GAN-based vocoders consist of two components: a generator to synthesize audio from acoustic features, and a discriminator to evaluate the generated audio for its authenticity compared to real audio data. The training criteria for the vocoder have ensured its consistent capability to output clean and high-fidelity monaural speech.

Recently, researchers have been attempting to apply generation-based neural networks in speech enhancement, speech separation, and TSE from different aspects and have made some progress. Some works [11, 12, 13, 14] use diffusion technique to re-synthesize speech from noisy recordings. Some works [15, 16] use a vocoder to generate clean speech without interference. In [17], Shi *et al.* propose a generation-based approach in speech separation and speech enhancement, based on the recognition of discrete symbols, and converting the discrete symbols to clean monaural speech. Yet this method requires the front-end model to predict both the speaker's discrete symbols and embedding, which will degrade the overall performance of the re-synthesized speech.

In this paper, we proposed discrete token based TSE, which is the first system to apply a generation-based neural network in the area of audio-only TSE, which consists of two components: discrete token prediction and discrete vocoder. For each training sample, we will use the pre-trained self-supervised training model to extract representations of speech, and then transform these representations into a sequence of discrete tokens via a clustering algorithm. After that, we train a prediction model to predict each target speech's discrete token sequence from noisy recordings. At last, a discrete vocoder will take the discrete tokens as input to synthesize the target speaker's speech. We evaluated the performance of the discrimination-based method, the mel-spectrogram based re-synthesis method, and our proposed discrete token based method. The experiments show that the proposed discrete token based model can obtain higher-quality target speech without interference than existing methods. The sound demo for this paper is available at `https://earthmanylf.github.io/DiscreteTSE/`.

2. METHODOLOGY

Our proposed discrete token based method consists of two parts: discrete token prediction, and discrete vocoder. Before training, we will discretize all the speech data, and then train the discrete token prediction model and discrete vocoder separately based on the discrete sequence of each speech. During inference, the discrete token prediction model will take mixed audio and the target speaker's enrollment as input to predict the target speaker's discrete sequence. Finally, the

[†] Yanmin Qian is the corresponding author.

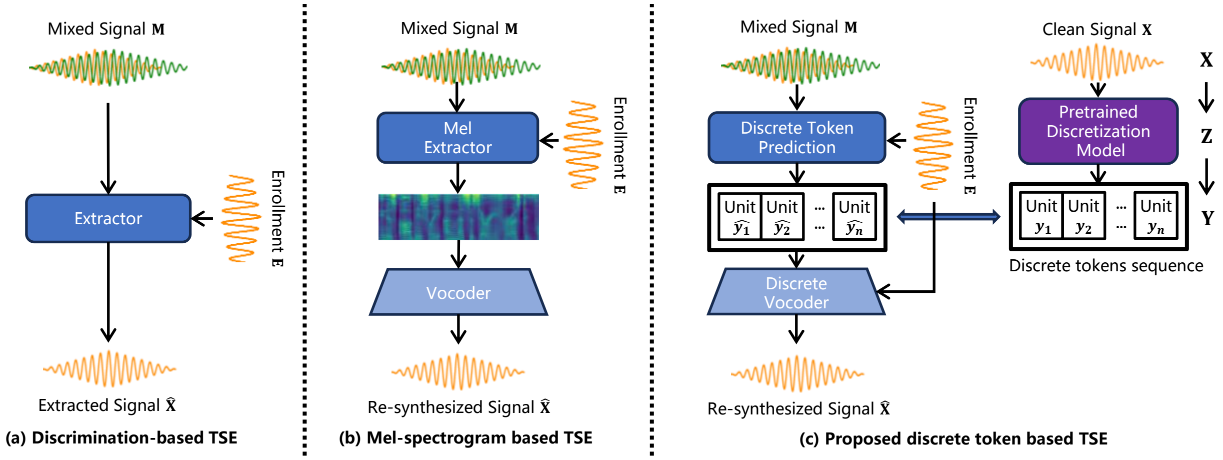

Fig. 1: Illustration of two existing TSE methods and our proposed discrete token based TSE.

discrete vocoder can convert the discrete sequence to a clean target speech with the help of the target speaker's enrollment. Fig. 1 shows two existing TSE methods (a) (b) and the proposed TSE method (c). With the help of discrete vocoder, we can synthesize target speaker's speech with high speech quality and without any interference, which is difficult to avoid in traditional discrimination-based methods. In the following discussion, we assume that the mixed signal is $\mathbf{M}$, the target clean speech is $\mathbf{X}$, and the enrollment of the target speaker is $\mathbf{E}$.

2.1. Speech Discretization

Speech discretization aims to encode the audio input into a discrete sequence, which is similar to the low-bitrate speech codecs [18]. Ideally, we want to encode as much information as possible so that the original speech can be restored without quality loss. In this paper, we use three methods to get the discrete units from the raw speech: HuBERT[19], vq-wav2vec[20] and Encodec [18]. The first two provide semantic speech representations, while the last extracts relatively low-level representations.

HuBERT is a self-supervised learning based model trained with the masked speech prediction objective. It has shown superior performance across multiple speech-related tasks. We follow the acoustic unit discovery process as in [19] to convert the continuous HuBERT representations into discrete tokens, where the k-means algorithm is adopted.

Vq-wav2vec is a self-supervised learning based model that integrates vector quantization into the wav2vec [21] framework. Vector quantization helps in discretizing the continuous audio signal into a sequence of discrete symbols.For the vq-wav2vec model, we use the official codebook provided by fairseq[1].

Encodec is an audio codec neural network, which consists of a streaming encoder-decoder architecture with quantized latent space trained in an end-to-end fashion. The quantized codecs compress speech and preserve low-level representations. For the Encodec model, we use the official implementation provided by Meta[2].

In our method, all the discretization tokens of speech are extracteded before training. A self-supervised learning based model will be used to discretize the raw wave $\mathbf{X}$ into a sequence of the discrete tokens $\mathbf{Y} = \{y_1, y_2, \cdots, y_n\}$, such as codebooks or HuBERT clustering. This process can be formulated as follows:

$$\mathbf{Z} = \{z_1, z_2, \cdots, z_n\} = F(\mathbf{X}) \tag{1}$$

$$y_i = Q(z_i) = \arg\min_j ||z_i - c_j|| \tag{2}$$

where F is the feature encoder (HuBERT, vq-wav2vec or Encodec encoder), Q is the discretization module, c_j is the j-th centroid in codebook or clustering, and n is the total length of the discrete sequence $\mathbf{Y} = \{y_1, y_2, \cdots, y_n\}$.

2.2. Discrete Token Prediction

With the discrete sequence $\mathbf{Y}$ prepared for the target speech, we can apply a discrete token prediction module to the mixed signal $\mathbf{M}$ with the help of the target speaker's enrollment $\mathbf{E}$. Instead of directly predicting the mask of target speech or mapping the spectrogram, we consider this process as a classification task, where we will predict the discrete tokens frame-by-frame. In the discrete token prediction module, we will compute the posterior probability of the discrete token sequence with the mixed signal $\mathbf{M}$ and the target speaker's enrollment $\mathbf{E}$ as inputs. This process can be formulated as follows:

$$p(\hat{\mathbf{Y}}|\mathbf{M}, \mathbf{E}) = \prod_{i=1}^{n} p(\hat{y}_i|\mathbf{M}, \mathbf{E}) \tag{3}$$

We adapt the architecture of the traditional discrimination-based TSE model, and replace the decoder with a softmax function as the classifier to predict the index of discrete tokens rather than reconstruct the clean signal itself. Based on this, we convert traditional discrimination-based TSE task to a classification task.

2.3. Discrete Vocoder

Unlike traditional neural vocoders from text-to-speech models, which usually use a mel-spectrogram to reconstruct the phase and to form the clean speech, the discrete vocoder uses discrete tokens as the input to generate higher-quality speech.

After the prediction of discrete token sequences $\hat{\mathbf{Y}}$, we will apply discrete vocoder on the discrete sequence to re-synthesize target speaker's speech $\hat{\mathbf{X}}$. Since the representations of self-supervised learning based models usually focus on the semantic information, we will use the target speaker's enrollment $\mathbf{E}$ as the condition to the discrete vocoder to help restore the information of the speaker's identity to the generated speech. The process of restoration can be formulated as follows:

$$\hat{\mathbf{X}} = \text{Vocoder}(\hat{\mathbf{Y}}, \mathbf{E}) \tag{4}$$

We train our discrete vocoder on the multi-speaker dataset so that the discrete vocoder is capable of modeling different speakers as required by the TSE task.

[1] https://github.com/facebookresearch/fairseq
[2] https://github.com/facebookresearch/encodec

Table 1: The overall evaluation of TSE performance with our proposed methods and some baseline models in WSJ0-2mix (clean) and Libri2Mix (noisy). The analysis of the poor performance of intrusive metrics such as PESQ, STOI, and SI-SDR is shown in Sec 3.3.2 in detail. *We adopt the setting and result of discrete speech separation from [17], which was evaluated on 8kHz WSJ0-2mix corpus.

Dataset	Model	Vocoder	PESQ	STOI	SI-SDR	OVRL	SIG	BAK
Clean	Mixture speech	-	2.05	77.80	2.50	2.81	3.42	3.27
	DPCCN-stft	-	**3.42**	**95.10**	**16.24**	3.13	3.42	4.07
	DPCCN-mel	HiFi-GAN	3.04	87.72	-28.35	3.29	3.52	4.13
	Conv-DPRNN*	Discrete HiFi-GAN	-	67.00	-	-	-	-
	SkiM	UniCATS(HuBERT-512)	1.32	71.65	-38.89	3.28	3.58	4.01
		UniCATS(HuBERT-4096)	1.33	71.15	-38.89	3.27	3.57	3.99
		UniCATS(vq-wav2vec)	1.64	73.54	-37.68	**3.37**	**3.62**	**4.10**
		Encodec	2.12	76.28	-1.65	2.13	2.48	3.31
Noisy	Mixture speech	-	1.51	64.73	-1.96	1.63	2.33	1.66
	DPCCN-stft	-	**2.61**	**85.22**	**9.36**	3.00	3.37	3.76
	DPCCN-mel	HiFi-GAN	2.30	79.78	-27.61	3.03	3.40	3.79
	SkiM	UniCATS(HuBERT-512)	1.08	65.85	-38.62	3.22	3.54	3.96
		UniCATS(HuBERT-4096)	1.06	66.42	-38.91	3.18	3.50	3.94
		UniCATS(vq-wav2vec)	1.29	68.52	-39.95	**3.27**	**3.56**	**4.02**
		Encodec	1.76	68.97	-2.35	1.94	2.20	3.35

3. EXPERIMENTS

3.1. Datasets

We perform experiments based on the public datasets: WSJ0-2mix[22] and Libri2Mix[23]. WSJ0-2mix is a dataset for speech separation without noise, and we follow the recipes[3] in [24] to form lists of enrollment for the TSE task. Libri2Mix is a dataset for speech separation with noise, we use train (train-100 + train-360) for training, and follow recipes[4] in [2] to form lists of enrollment for the TSE task. We use LibriTTS [25] to train all our vocoders. The sampling rate of all the audios is 16kHz, and we choose the min mode of WSJ0-2mix and Libri2Mix same as other TSE methods. All the experiments were conducted using the ESPnet [26] toolkit.

3.2. Implementation Details

We choose a recent model DPCCN [27] as our baseline for its good performance on TSE, standing for the traditional discrimination-based model. For the mel-spectrogram based method, we use DPCCN to predict the target speaker's mel-spectrogram, and then use HiFi-GAN [18] which is popular in text-to-speech synthesis to restore the clean speech from mel-spectrogram.

For our proposed method, to quickly predict the token, we apply SkiM [28] as the discrete token prediction module with an additional speaker encoder for the TSE task similar to SpeakerBeam [2], and UniCATS [29] as discrete vocoder under the cases of HuBERT and vq-wav2vec. UniCATS is a HiFiGAN-based vocoder that inputs discrete token sequence to generate high-quality speech. Specifically, we use the representations of the final layer from HuBERT-Base model trained on Librispeech [30] without finetuning. Then, we trained a k-means model to generate the discrete tokens for the corresponding data. And we use vq-wav2vec-kmeans model trained on Librispeech [30] without finetuning to extract discrete tokens. It should be noted that original UniCATS only uses discrete tokens from vq-wav2vec as the input, we train UniCATS with clustering discrete tokens from HuBERT additionally.

When we choose Encodec as our discrete tokens, we choose the 24kHz&6kbps setting of Encodec, which means we have 8 groups of tokens and each group has 1024 kinds of different tokens. Since Encodec is an audio codec model, its quantized code has modeled both semantic and speaker information, we will use the Encodec decoder to directly restore the target speech without enrollment rather than re-train a discrete vocoder. The output of Encodec decoder will be resampled to 16kHz under the case of Encodec. The settings of vocoders and corresponding discrete tokens are shown in Table 2.

Table 2: Training settings for the discrete vocoder. †: $A * B$ means that we have B groups discrete tokens where each group has A kinds of tokens

Vocoder Architecture	Discrete Tokens	Clusters	Token Dim
HiFi-GAN	mel-spectrogram	-	-
UniCATS	HuBERT	4096	768
	HuBERT	512	768
	vq-wav2vec	$320{*}2^\dagger$	512
Encodec decoder	Encodec	$1024{*}8^\dagger$	-

3.3. Results

3.3.1. *Evaluation Metrics*

As observed in previous works[17], it's difficult to compare the performance of generation-based models and discrimination-based models in the metrics of signal-level similarity, such as SI-SDR, PESQ, STOI, and so on. The reason is that these intrusive metrics are designed to measure the difference between the reconstructed signal and the original signal, which is sensitive to slight sample misalignment caused by phase errors. The existing generative methods intrinsically cannot handle the phase alignment issue in the generation process, thus leading to poor scores in terms of the aforementioned metrics.

Due to these observations, we apply non-intrusive metrics from DNSMOS [31] to evaluate the quality of the synthesized speech. DNSMOS is a non-intrusive speech quality metric developed by Microsoft. The scores reflect the overall quality of the audio clip, which are computed by a trained convolutional neural network (CNN) based model. This framework gives the standalone quality scores of speech and background noise in addition to the overall quality, including speech quality (SIG), background noise quality (BAK), and overall quality (OVRL).

3.3.2. *Analysis*

Table 1 shows the performance of our proposed methods and baseline models on both clean and noisy datasets. We compare the

[3] https://github.com/xuchenglin28/speaker_extraction

[4] https://github.com/BUTSpeechFIT/speakerbeam

discrimination-based model (DPCCN-stft), mel-spectrogram based model (DPCCN-mel), and our proposed discrete token based models using tokens of different kinds and different numbers of clusters.

On the one hand, the results on intrusive metrics show that all the generation-based models achieve worse performance than the discrimination-based model. The mel-spectrogram based model achieves better than discrete tokens based model, while Encodec based models achieve better than HuBERT and vq-wav2vec based models.

We think there are several reasons: 1) Vocoder synthesizes signal from mel-spectrogram and discrete token sequence, and the phase is regenerated, which is often misaligned with the original input signal; 2) Vocoder uses GAN-based loss criteria which does not force the model to reconstruct the signal perfectly; 3) Mel-spectrogram has more signal-level information, while the discrete tokens contain mainly semantic-level information; 4) Encodec is designed to compress audio for transmitting, leading to that the Encodec tokens allow for better storage of signal-level information and the decoder of Encodec has a greater ability to reconstruct speech signals.

On the other hand, the results on non-intrusive metrics are completely different from those on intrusive metrics. All the generation-based models outperform the discrimination-based model except for the Encodec-based discrete model, which demonstrated that the signals reconstructed by the vocoder are perceptually cleaner than those reconstructed by the discrimination-based model.

To be specific, HuBERT based model tends to have close performance to the mel-spectrogram based models, vq-wav2vec based model outperforms mel-spectrogram based models, while Encodec based model achieves the worst scores among generation-based models. This can be attributed to the fact that when a frame of speech is represented by multiple discrete tokens, it is difficult for the discrete token prediction module to accurately predict all the tokens of a frame at once, which leads to a degradation in the quality of the generated speech. This conjecture is also supported by the low accuracy.

Another interesting fact is that the gap between discrimination-based model and generation-based models of non-intrusive metrics on noisy dataset is greater than that on clean dataset. This shows that our proposed method is more advantageous in noisy environments as the re-synthesized speech completely removes any interference voice and noise.

3.4. Ablation Study

We based Libri2mix on our dataset and chose SkiM as the discrete token prediction module in our ablation studies. Table 3 compares the performance of discrete vocoder settings using ground truth tokens and predicted discrete token sequence, which shows the upper bound of our proposed method. We can see from the results that when the accuracy of the predicted tokens is greater than 30%, the non-intrusive metrics are essentially similar to when the ground truth tokens are used. This fact shows that as long as the prediction accuracy reaches a certain threshold, the quality of the generated speech will not degrade, which can prove that the discrete vocoder has some fault tolerance. Yet the accuracy of the predicted tokens can affect the intrusive metrics, a lower accuracy will degrade PESQ and STOI. We think the reason is that the wrong prediction on some specific combinations of tokens may mislead the vocoder to generate incorrect phonemes. Unsurprisingly, Encodec based model gets the worst prediction accuracy with the biggest gap between ground truth tokens and predicted tokens, which proves the difficulty of all the tokens of a frame at the same time when a frame of speech is represented by multiple tokens.

Table 4 shows the performance when we use re-synthesized speech of the target speaker as the condition for the audio mixture in the discrimination-based TSE model. The purpose is to show that our reconstructed speech already contains information about the target speaker. In this experiment, we use the same baseline model from Table 1 which is trained without any synthesized speech as the condition, and the same model trained with only synthesized speech as the condition. From the results, we can see that directly using synthesized speech as the condition for the discrimination-based model that has not seen synthesized speech in training will degrade the performance, while the model trained with synthesized speech as the condition can achieve the same performance as the traditional discrimination-based model. This phenomenon indicates that the synthesized speech from our proposed architecture contains information about the target speaker.

Table 3: The overall evaluation of TSE performance with our proposed methods between ground truth tokens and predicted tokens. GT denotes ground truth tokens. ACC denotes the accuracy of the predicted tokens.

Vocoder Setting	GT	PESQ	STOI (×100)	OVRL	SIG	BAK	ACC (%)
HuBERT-512	✓	1.17	70.44	3.23	3.54	3.99	100.00
	✗	1.08	65.85	3.22	3.54	3.96	48.14
HuBERT-4096	✓	1.14	70.87	3.18	3.50	3.94	100.00
	✗	1.06	66.42	3.18	3.51	3.93	41.29
vq-wav2vec	✓	1.45	73.90	3.19	3.52	3.93	100.00
	✗	1.29	68.52	3.27	3.56	4.02	30.44
Encodec	✓	3.17	93.52	2.91	3.29	3.74	100.00
	✗	1.76	68.97	1.94	2.20	3.35	15.73

Table 4: The performance of using synthesized speech as the condition for discrimination-based TSE model trained w/ or w/o synthesized speech as the condition. 'syn' denotes synthesized speech.

Training Setting	Condition Types	PESQ	STOI (×100)	SI-SDR (dB)	OVRL
w/o syn	Original	2.61	85.22	9.36	3.00
	HuBERT-512	2.62	84.76	8.99	2.99
w/ syn	HuBERT-512	2.66	85.52	9.41	2.96

4. CONCLUSIONS

In this paper, we proposed a new generation-based method for TSE task based on discrete token prediction and discrete vocoder. With the predicted discrete token sequence, we can restore the target speech via an advanced discrete vocoder. Experiments on both clean and noisy datasets in different settings show that our method can synthesize high-quality and human-hearing friendly target speech without any interference, which is hard to avoid in discrimination-based methods. Our method can achieve better scores on non-intrusive metrics such as DNSMOS, while worse scores on intrusive metrics such as PESQ and STOI, which is caused by the current vocoder. Future work includes further improving our proposed method and boosting the performance in terms of the intrusive metrics.

5. ACKNOWLEDGEMENT

This work was supported in part by China STI 2030-Major Projects under Grant No. 2021ZD0201500, in part by China NSFC projects under Grants 62122050 and 62071288, and in part by Shanghai Municipal Science and Technology Commission Project under Grant 2021SHZDZX0102, and in part by Shanghai Science and Technology Committee under Grant No. 21511101100.

6. REFERENCES

[1] 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. IEEE ICASSP*, 2023, pp. 1–5.

[2] 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.

[3] K. Zhang, M. Borsdorf, Z. Pan, H. Li, Y. Wei, and Y. Wang, "Speaker Extraction with Detection of Presence and Absence of Target Speakers," in *Proc. Interspeech*, 2023, pp. 3714–3718.

[4] B. Zeng, S. Hongbin, Y. Wan, and M. Li, "SEF-Net: Speaker Embedding Free Target Speaker Extraction Network," in *Proc. Interspeech*, 2023, pp. 3452–3456.

[5] 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. Interspeech*, 2019, pp. 2728–2732.

[6] K. Liu, Z. Du, X. Wan, and H. Zhou, "X-SEPFORMER: End-To-End Speaker Extraction Network with Explicit Optimization on Speaker Confusion," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[7] W. Liu and C. Xie, "Gated Convolutional Fusion for Time-Domain Target Speaker Extraction Network," in *Proc. Interspeech*, 2022, pp. 5368–5372.

[8] H. Sato, T. Ochiai, M. Delcroix, K. Kinoshita, T. Moriya, N. Makishima, M. Ihori, T. Tanaka, and R. Masumura, "Strategies to Improve Robustness of Target Speech Extraction to Enrollment Variations," in *Proc. Interspeech*, 2022, pp. 996–1000.

[9] M. Delcroix, K. Kinoshita, T. Ochiai, K. Zmolikova, H. Sato, and T. Nakatani, "Listen only to me! How well can target speech extraction handle false alarms?" in *Proc. Interspeech*, 2022, pp. 216–220.

[10] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Proc. NeurIPS*, vol. 33, 2020, pp. 17 022–17 033.

[11] N. Kamo, M. Delcroix, and T. Nakatani, "Target Speech Extraction with Conditional Diffusion Model," in *Proc. Interspeech*, 2023, pp. 176–180.

[12] H. Yen, F. G. Germain, G. Wichern, and J. L. Roux, "Cold Diffusion for Speech Enhancement," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[13] R. Scheibler, Y. Ji, S.-W. Chung, J. Byun, S. Choe, and M.-S. Choi, "Diffusion-Based Generative Speech Source Separation," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[14] B. Chen, C. Wu, and W. Zhao, "SEPDIFF: Speech Separation Based on Denoising Diffusion Model," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[15] B. Irvin, M. Stamenovic, M. Kegler, and L.-C. Yang, "Self-Supervised Learning for Speech Enhancement Through Synthesis," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[16] R. Mira, B. Xu, J. Donley, A. Kumar, S. Petridis, V. K. Ithapu, and M. Pantic, "LA-VOCE: LOW-SNR Audio-Visual Speech Enhancement Using Neural Vocoders," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[17] J. Shi, X. Chang, T. Hayashi, Y.-J. Lu, S. Watanabe, and B. Xu, "Discretization and re-synthesis: an alternative method to solve the cocktail party problem," *arXiv preprint arXiv:2112.09382*, 2021.

[18] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High Fidelity Neural Audio Compression," *arXiv preprint arXiv:2210.13438*, 2022.

[19] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Trans. ASLP.*, vol. 29, pp. 3451–3460, 2021.

[20] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations," in *Proc. ICLR*, 2020.

[21] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech*, 2019, pp. 3465–3469.

[22] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE ICASSP*, 2016, pp. 31–35.

[23] 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.

[24] C. Xu, W. Rao, E. S. Chng, and H. Li, "SpEx: Multi-Scale Time Domain Speaker Extraction Network," *IEEE Trans. ASLP.*, vol. 28, pp. 1370–1384, 2020.

[25] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Proc. Interspeech*, 2019, pp. 1526–1530.

[26] C. Li, J. Shi, W. Zhang, A. S. Subramanian, X. Chang, N. Kamo, M. Hira, T. Hayashi, C. Boeddeker, Z. Chen, and S. Watanabe, "ESPnet-SE: End-to-End Speech Enhancement and Separation Toolkit Designed for ASR Integration," in *Proc. IEEE SLT*, 2021, pp. 785–792.

[27] J. Han, Y. Long, L. Burget, and J. Černocký, "DPCCN: Densely-Connected Pyramid Complex Convolutional Network for Robust Speech Separation and Extraction," in *Proc. IEEE ICASSP*, 2022, pp. 7292–7296.

[28] C. Li, L. Yang, W. Wang, and Y. Qian, "SkiM: Skipping Memory Lstm for Low-Latency Real-Time Continuous Speech Separation," in *Proc. IEEE ICASSP*, 2022, pp. 681–685.

[29] C. Du, Y. Guo, F. Shen, Z. Liu, Z. Liang, X. Chen, S. Wang, H. Zhang, and K. Yu, "UniCATS: A Unified Context-Aware Text-to-Speech Framework with Contextual VQ-Diffusion and Vocoding," *arXiv preprint arXiv:2306.07547*, 2023.

[30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. IEEE ICASSP*, 2015, pp. 5206–5210.

[31] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE ICASSP*, 2022, pp. 886–890.