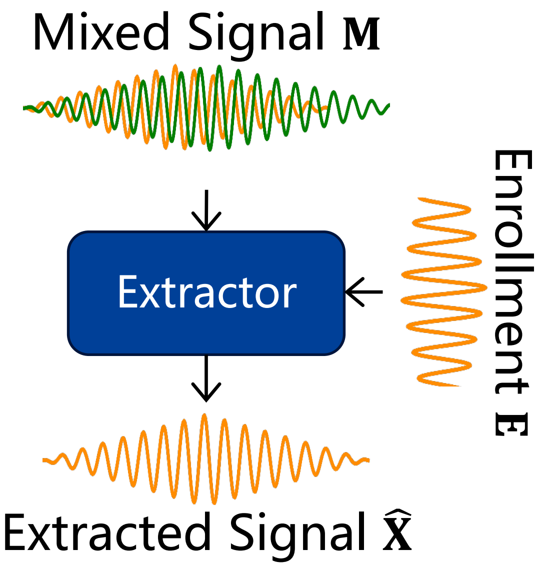
Generation-Based Target Speech Extraction with Speech Discretization and Vocoder

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

Shanghai Jiao Tong University, Shanghai, China

Target Speech Extraction(TSE)



Target speech extraction (TSE) aims at isolating the speech of a specific target speaker from an audio mixture, with the help of an auxiliary recording of target speaker.

Most existing TSE methods employ discriminative models to estimate the target speakers proportion in the mixture, but they often fail to compensate for the missing or highly corrupted frequency components in the speech signal.

Generation-Based Target Speech Extraction with Speech Discretization and Vocoder

Generative

Discriminative

Audio-only

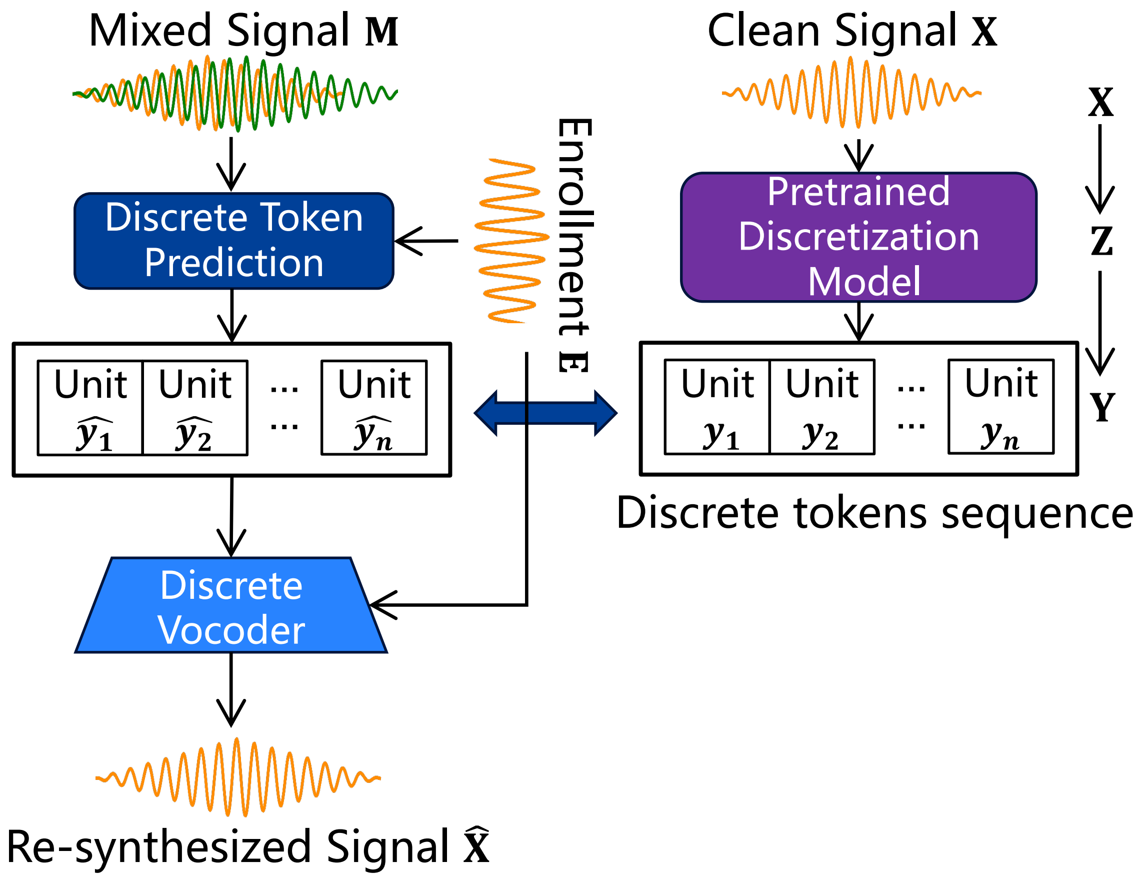
XXX

Multimodal

XXX

2018	2019	2020	2021	2022	2023 - Present	
	ICASSP'19 Speaker Inventory	ICASSP'20 TD-Speaker Beam	ICASSP'21 MuSE / ADL-MVDR	ICASSP'22 sDPCCN	SLT'22 LiMuSE	arXiv:2310.04791 DiffSpEx
ICASSP'18 Speaker Beam	ICASSP'19 TseNet	ICASSP'20 Speaker filter	Interspeech'21 SpEx_{pc} SpEx_{sc}	ICASSP'22 L-SpEx	ICASSP'23 LA-VOCE / AV-SepFormer	arXiv:2310.07284 LLM-TSE
ICASSP'18 MC-SpeakerBeam	Interspeech'19 Neural Spatial Filter	TASLP'20, Interspeech'20 SpEx SpEx+	Interspeech'21 AvaTr	MM'22 Concept Beam	ICASSP'23 X-SepFormer / VEVEN	arXiv:2401.15993 C-TSE
Interspeech'18 DENet	Interspeech'19 Voice Filter	TASLP'20 AV-CVAE	Interspeech'21 Personalized PercepNet	TASLP'22 USEV	Interspeech'23 SEF-Net / MC-SpEx	ICASSP'24 Active Extract
SLT'18 Zhuo *et al.*	Interspeech'19 Speaker Beam-AV	Interspeech'20 WaveFilter / X-TaSNet	CVPR'21 VisualVoice	Interspeech'22 SAMoM	Interspeech'23 Diff-TSE	IJCNN'24 AVHuMAR-TSE
etc.	etc.	etc.	etc.	etc.	etc.	etc.

Discrete Token based TSE



First system to apply a vocoder-based generative method in audio-only TSE.

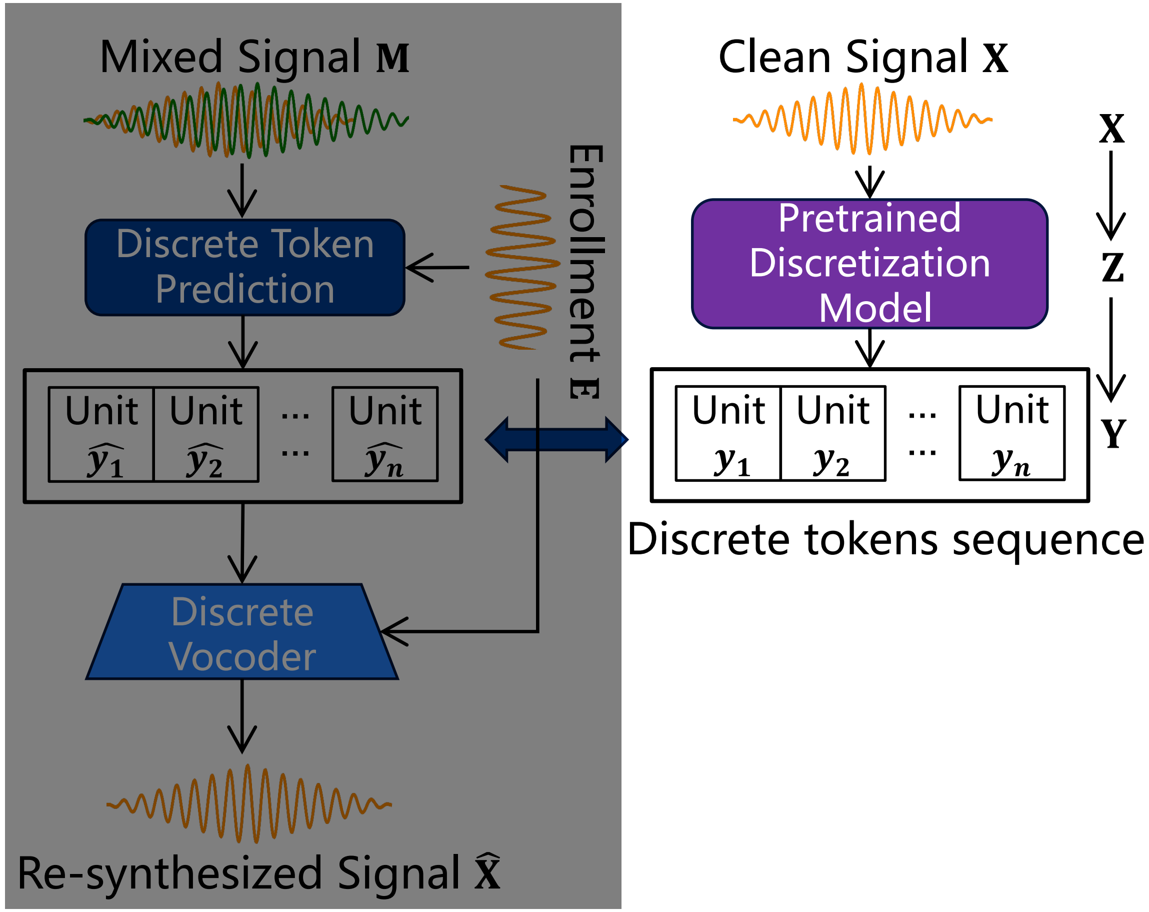
2 models:

1. Discrete Token Prediction
2. Discrete Vocoder

Discrete Token Prediction predicts the target speaker's discrete token sequence.

Discrete Vocoder converts the discrete sequence to a clean target speech.

Speech Discretization



Speech discretization aims to encode the audio input into a discrete sequence.

All the discretization tokens of speech are extracted before training.

$$\mathbf{Z} = \{z_1, z_2, \ldots, z_n\} = F(\mathbf{X})$$

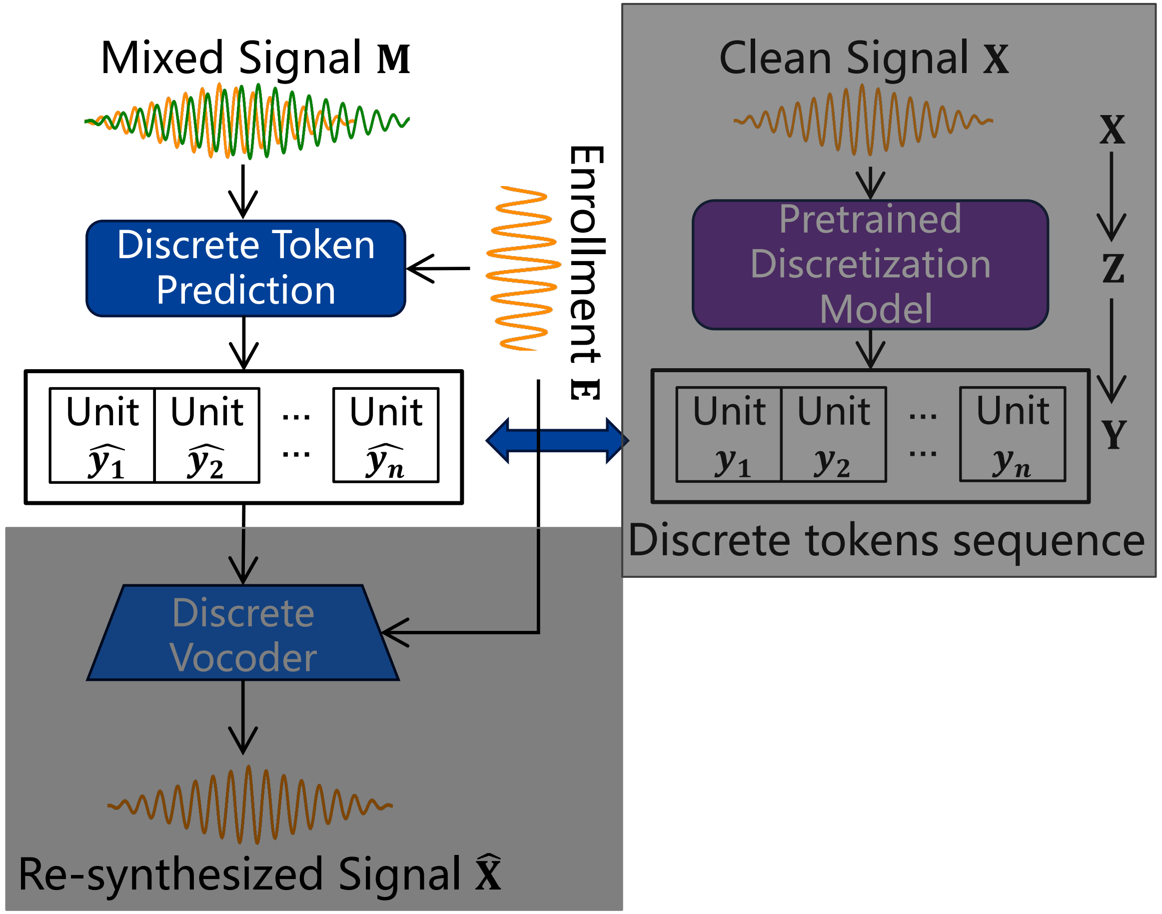
$$y_i = Q(z_i) = \arg\min_j ||\mathbf{z}_i - \mathbf{c}_j||$$

F is the feature encoder (HuBERT, vq-wav2vec or EnCodec encoder)

Q is the discretization module

$\mathbf{c}_j$ is the j-th centroid in codebook or clustering

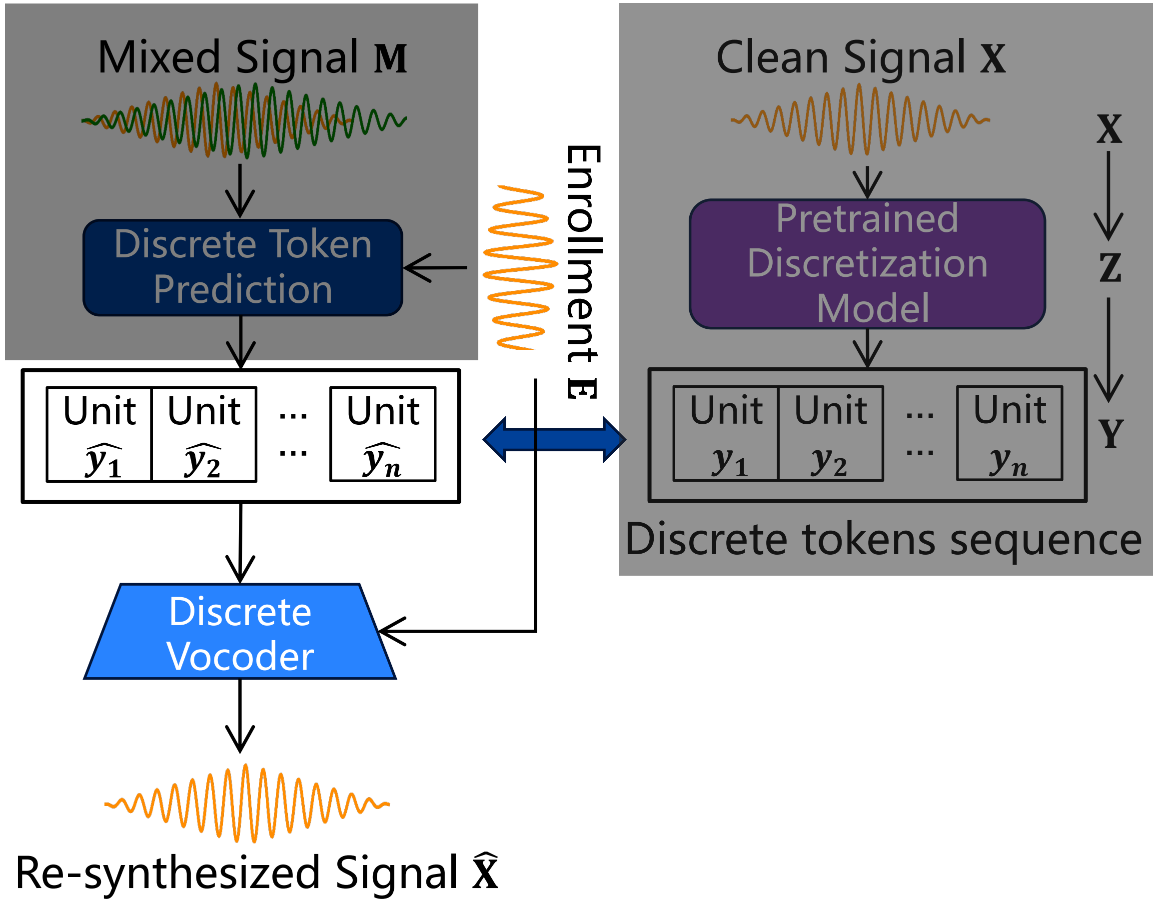
Discrete Token Prediction



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.

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

Discrete Vocoder



Discrete vocoder takes discrete tokens as input to generate higher-quality speech.

we use the enrollment $\mathbf{E}$ as a condition to the discrete vocoder to help restore the speaker characteristics in the re-synthesized speech.

$$\hat{\mathbf{X}} = \mathrm{Vocoder}(\hat{\mathbf{Y}}, \mathbf{E})$$

Experiments

Datasets

- WSJ0-2mix (clean)
- Libri2mix (noisy)

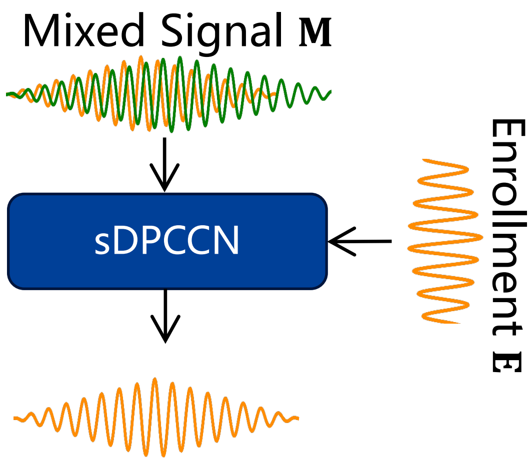
Models

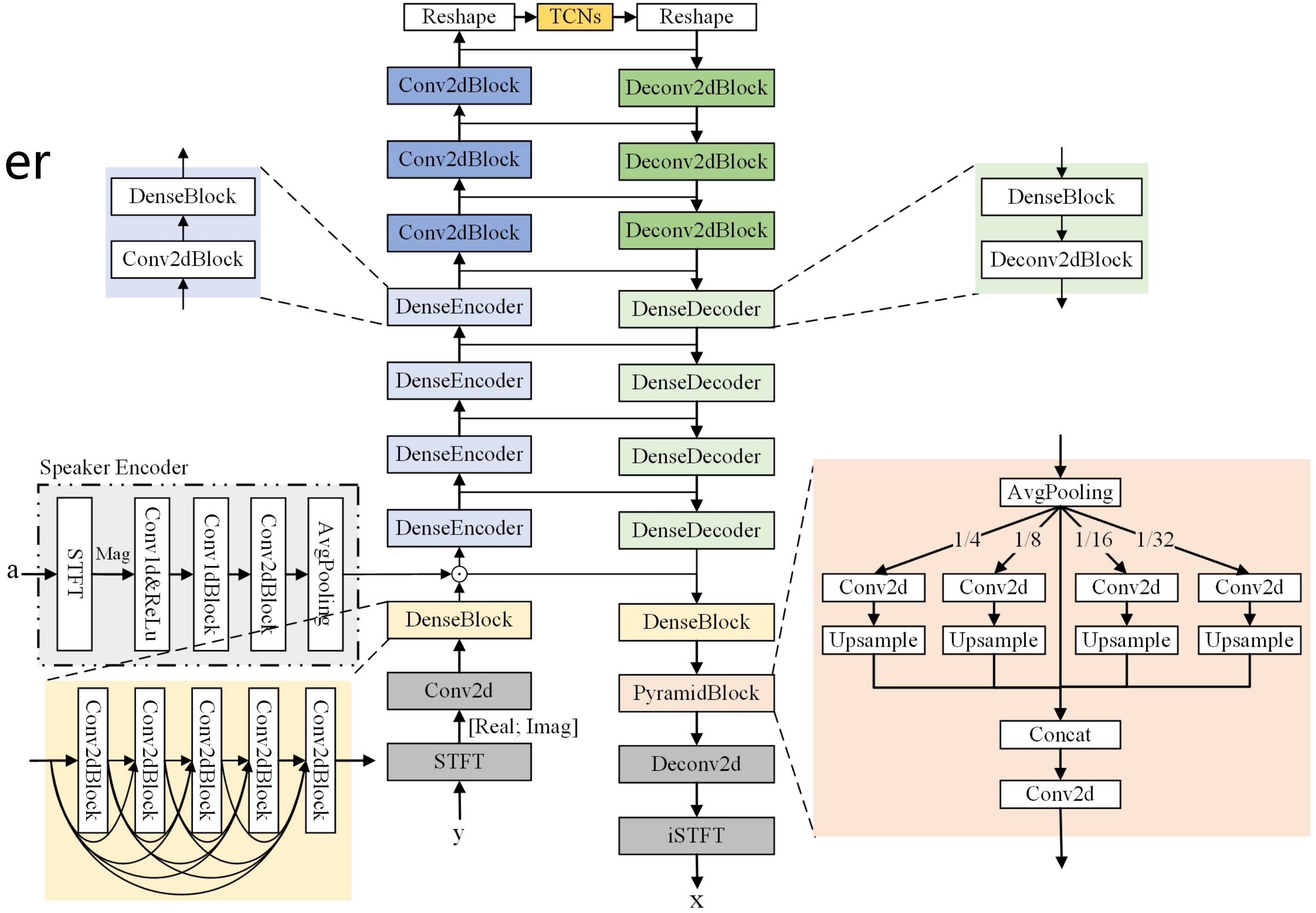
1) Baselines
   1. Discrimination-based: DPCCN (denoted as DPCCN-stft)
   2. Mel-spectrogram based: DPCCN (denoted as DPCCN-mel) and HiFi-GAN (denoted as vocoder)
2) (Proposed) Discrete Token based TSE
   - Discrete token prediction module: SkiM
   - Discrete Vocoder: UniCATS

Experiments

1) **Baselines** (discriminative)

1. DPCCN with speaker encoder (sDPCCN)

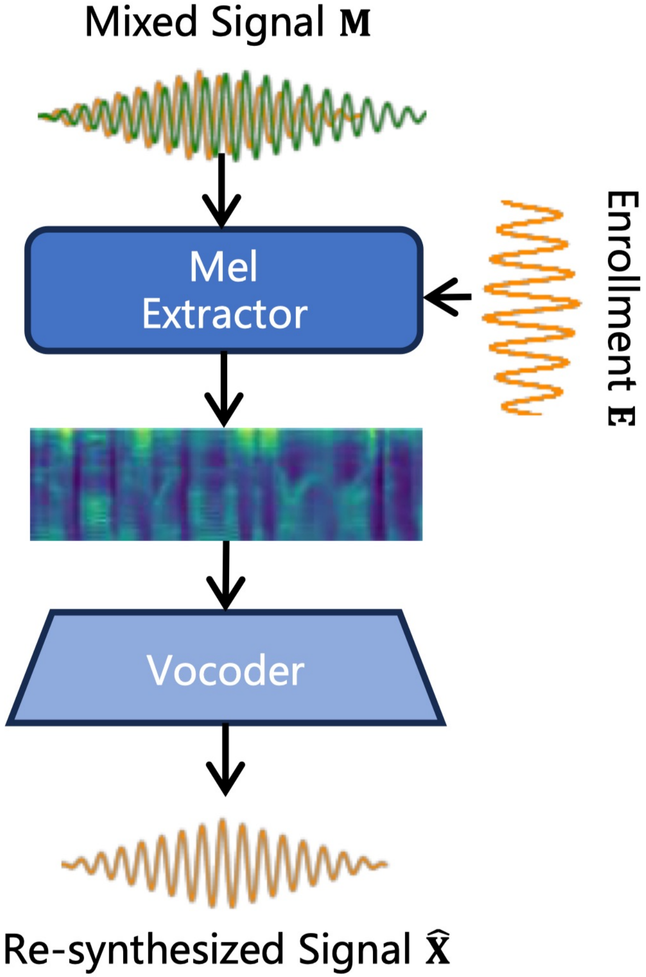




J. Han, Y. Long, L. Burget, and J. Černockỳ, "DPCCN: Densely-connected pyramid complex convolutional network for robust speech separation and extraction," in Proc. IEEE ICASSP, 2022, pp. 7292–7296.

Experiments

1) **Baselines** (generation-based)
   2. Mel-spectrogram based
      - ❑ Mel Extractor: sDPCCN
      - ❑ Vocoder: HiFi-GAN

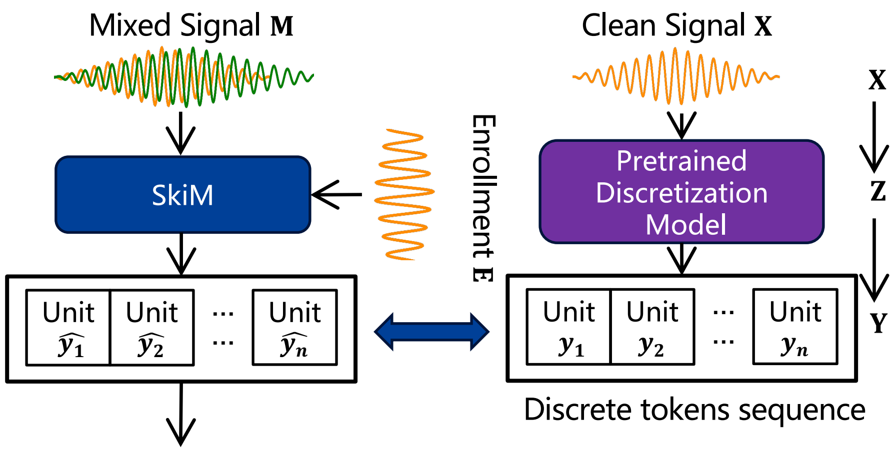


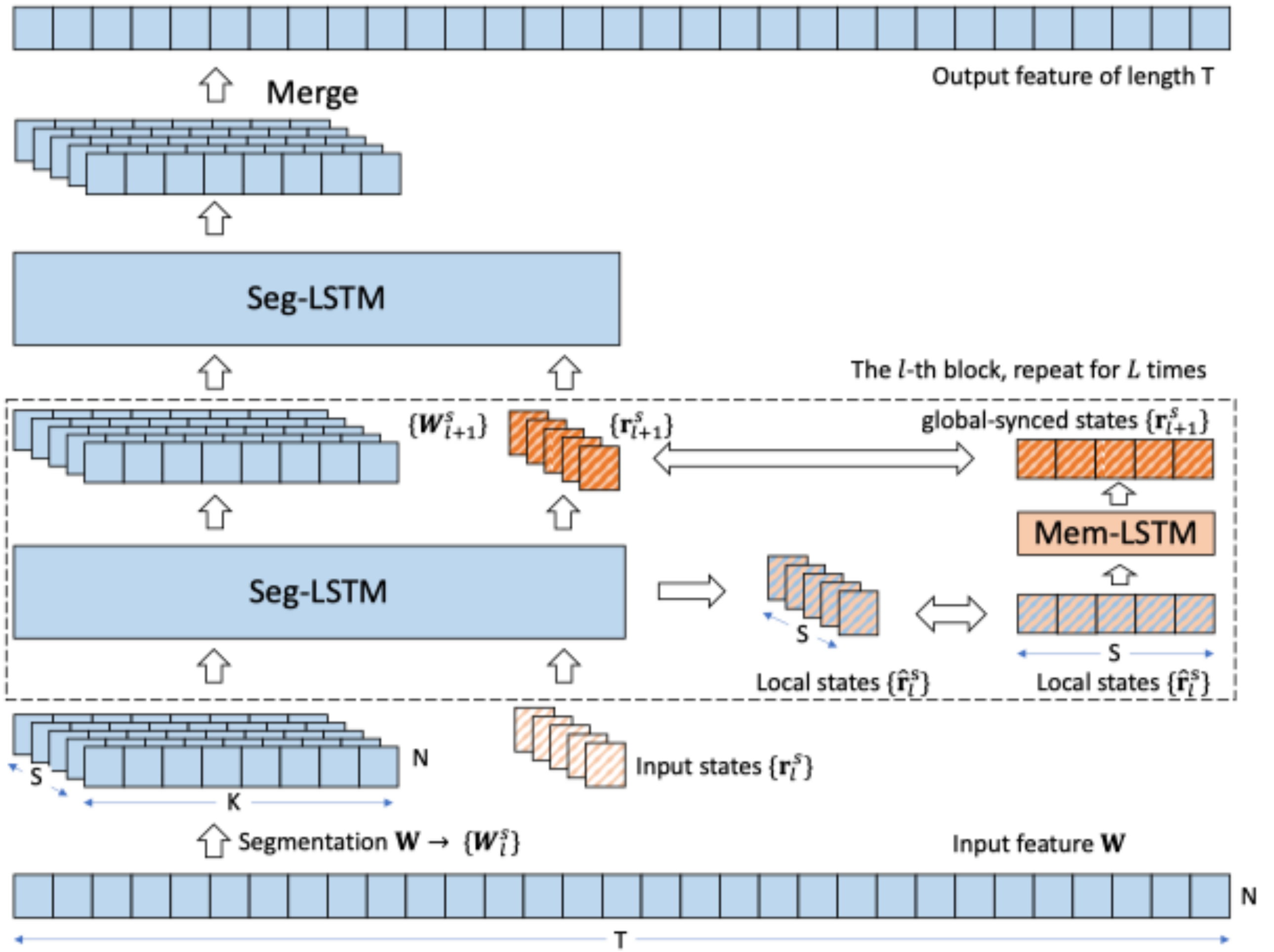
J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," Advances in Neural Information Processing Systems, vol. 33, pp. 17022–17033, 2020.

Experiments

2) (Proposed) Discrete Token based TSE

- **SkiM:** time-domain dual-path model



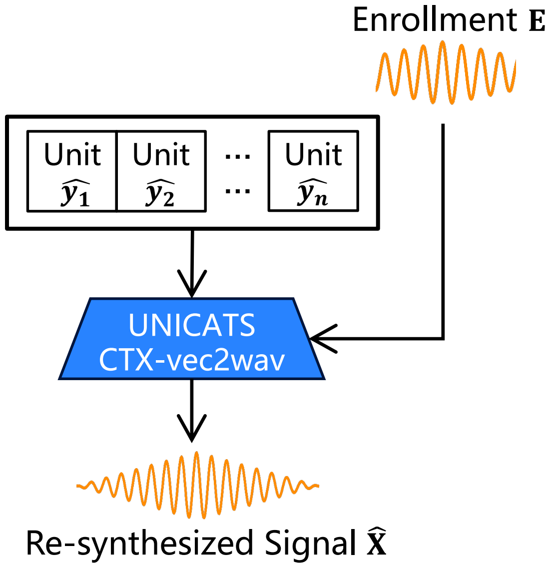


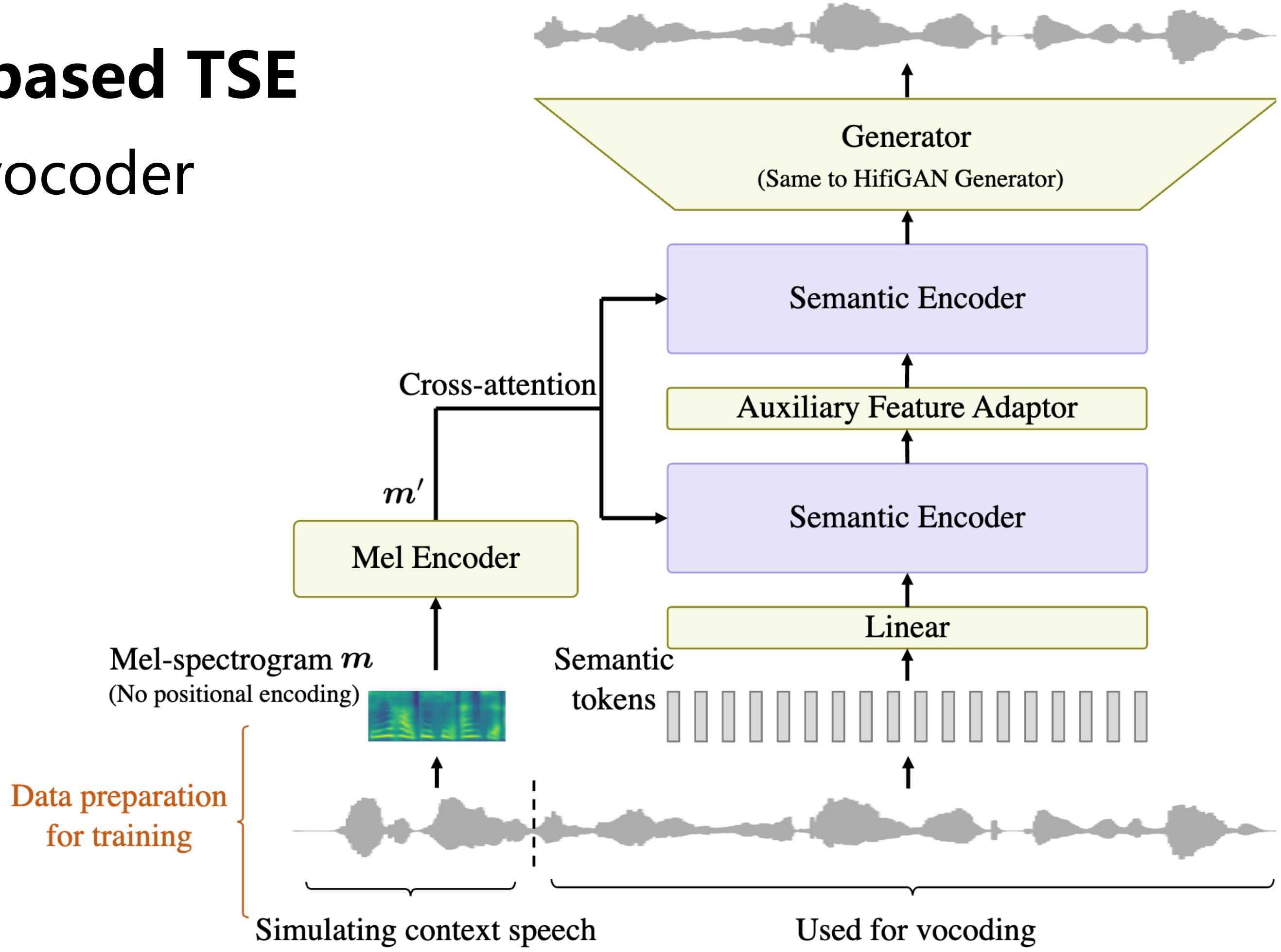
C. Li, L. Yang, W. Wang, and Y. Qian, "SkiM: Skipping memory LSTM for low-latency real-time continuous speech separation," in Proc. IEEE ICASSP, 2022, pp. 681–685.

Experiments

2) (Proposed) Discrete Token based TSE

- **UniCATS:** high-performance vocoder





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," in Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 17924–17932.

Experiments-Discrete Vocoder Settings

Vocoder Architecture	Discrete Token	Clusters	Token Dim
HiFi-GAN	mel-spectrogram	-	-
UniCATS	HuBERT*	4096	768
	HuBERT	512	768
	vq-wav2vec	320 × 2**	512
EnCodec-decoder	EnCodec	1024 × 8	-

*: We use HuBERT-base for the experiments.
**: A∗B means that we have B groups discrete tokens where each group has A kinds of tokens.

Experiments – Main Results

Dataset	Model	Vocoder	SI-SDR	OVRL	SIG	BAK
Clean WSJ0-2mix	Mixture	-	2.50	2.81	3.42	3.27
	DPCCN-stft	-	**16.24**	3.13	3.42	4.07
	DPCCN-mel	HiFi-GAN	-28.35	3.29	3.52	**4.13**
	SkiM	UniCATS(HuBERT-512)	-38.89	3.28	3.58	4.01
		UniCATS(HuBERT-4096)	-38.89	3.27	3.57	3.99
		UniCATS(vq-wav2vec)	-37.68	**3.37**	**3.62**	4.10
		Encodec	-1.65	2.13	2.48	3.31
Noisy Libri2Mix	Mixture	-	-1.96	1.63	2.33	1.66
	DPCCN-stft	-	**9.36**	3.00	3.37	3.76
	DPCCN-mel	HiFi-GAN	-27.61	3.03	3.40	3.79
	SkiM	UniCATS(HuBERT-512)	-38.62	3.22	3.54	3.96
		UniCATS(HuBERT-4096)	-38.91	3.18	3.50	3.94
		UniCATS(vq-wav2vec)	-39.95	**3.27**	**3.56**	**4.02**
		Encodec	-2.35	1.94	2.20	3.35

- Intrusive metric (SI-SDR)
  - All the generation-based models achieve worse performance than the discriminative models.
- Reasons:
  - Signals synthesized from vocoder have phase alignment issue.
  - GAN-based loss cannot force vocoder to reconstruct the signal perfectly.
  - Discrete tokens contain mainly semantic-level information.

Experiments – Main Results

Dataset	Model	Vocoder	SI-SDR	OVRL	SIG	BAK
Clean WSJ0-2mix	Mixture	-	2.50	2.81	3.42	3.27
	DPCCN-stft	-	16.24	3.13	3.42	4.07
	DPCCN-mel	HiFi-GAN	-28.35	3.29	3.52	**4.13**
	SkiM	UniCATS(HuBERT-512)	-38.89	3.28	3.58	4.01
		UniCATS(HuBERT-4096)	-38.89	3.27	3.57	3.99
		UniCATS(vq-wav2vec)	-37.68	**3.37**	**3.62**	4.10
		EnCodec	-1.65	2.13	2.48	3.31
Noisy Libri2Mix	Mixture	-	-1.96	1.63	2.33	1.66
	DPCCN-stft	-	9.36	3.00	3.37	3.76
	DPCCN-mel	HiFi-GAN	-27.61	3.03	3.40	3.79
	SkiM	UniCATS(HuBERT-512)	-38.62	3.22	3.54	3.96
		UniCATS(HuBERT-4096)	-38.91	3.18	3.50	3.94
		UniCATS(vq-wav2vec)	-39.95	**3.27**	**3.56**	**4.02**
		EnCodec	-2.35	1.94	2.20	3.35

- Non-intrusive metrics (OVRL, SIG, BAK)
  - All the generation-based models outperform the discriminative model except for the EnCodec-based discrete model.

Experiments – Ablation Study

We based Libri2mix as our dataset in ablation studies.

First, we compare the performance of discrete vocoder settings using ground truth tokens and predicted discrete token sequence

Token	GT	OVRL	SIG	BAK	ACC(%)
HuBERT-512	✓	3.23	3.54	3.99	100.00
	×	3.22	3.54	3.96	48.14
HuBERT-4096	✓	3.18	3.50	3.94	100.00
	×	3.18	3.51	3.93	41.29
Vq-wav2vec	✓	3.19	3.52	3.93	100.00
	×	3.27	3.56	4.02	30.44
EnCodec	✓	2.91	3.29	3.74	100.00
	×	1.94	2.20	3.35	15.73

When the accuracy of the prediction is greater than 30%, the non-intrusive metrics are essentially similar to when the ground truth tokens are used.

The discrete vocoder has some fault tolerance.

Experiments – Ablation Study

Second, we use re-synthesized speech of the target speaker from our method as the enrollment for the discriminative TSE model.

The purpose is to show that our reconstructed speech contains the target speaker information.

We have 2 training settings: TSE model trained with and without synthesized speech

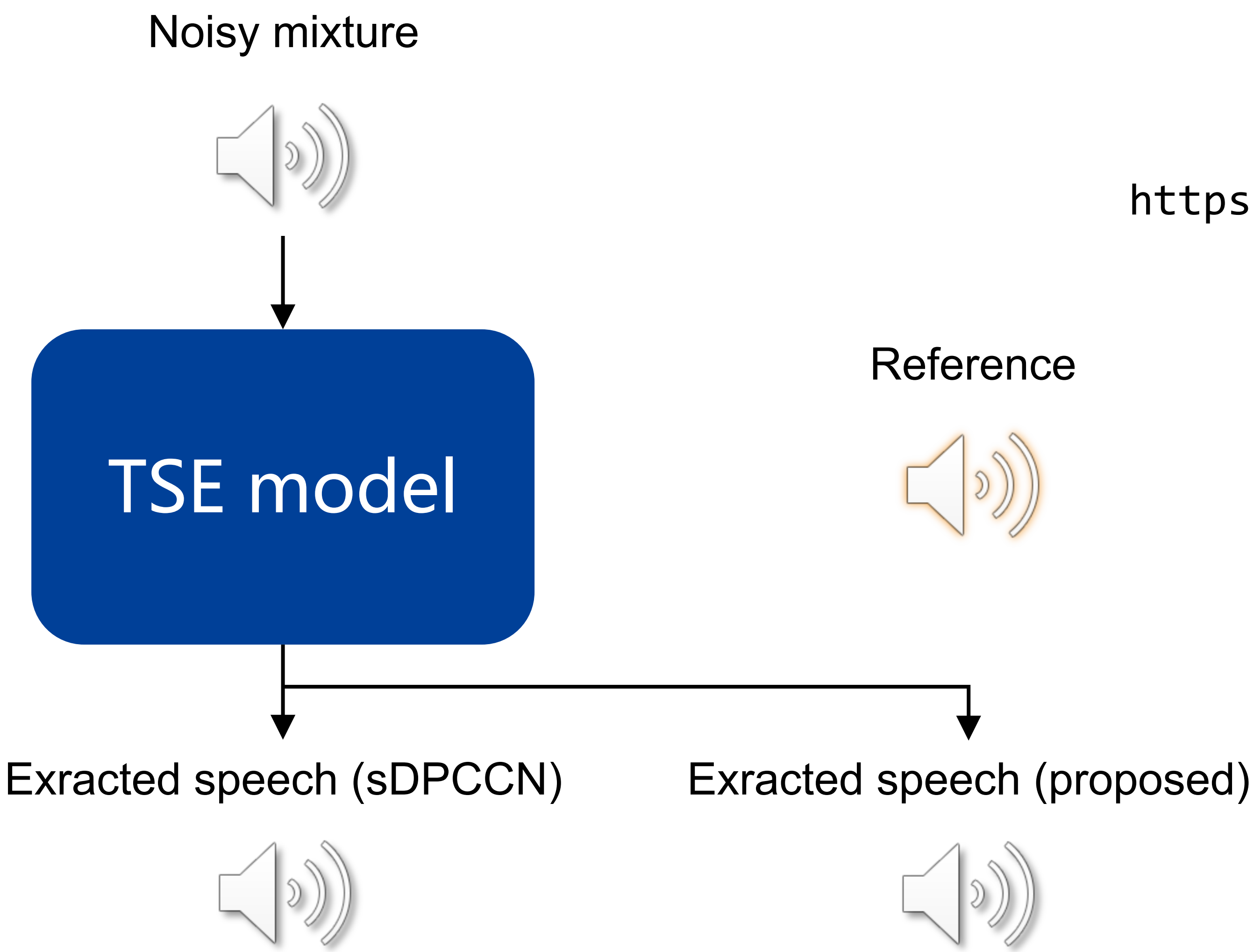
Training Setting	Enrollment	SI-SDR	OVRL
w/o syn	Original	9.36	3.00
	HuBERT-512	8.99	2.99
w/ syn	HuBERT-512	9.41	2.96

Directly using synthesized speech as the enrollment for the discriminative model that has not seen synthesized speech in training will degrade the performance.

The model trained with synthesized speech as the enrollment can achieve comparable performance as the discrimination-based model.

- Synthesized speech from our proposed architecture contains information about the target speaker.

Experiments – Main Results (demo)



https://earthmanylf.github.io/DiscreteTSE/

Conclusion

We proposed a new generation-based method for TSE task based on discrete token prediction and discrete vocoder. This is the first discrete token based method in audio-only TSE.

Experiments on both clean and noisy benchmark datasets in different settings show that our method can synthesize high-quality and human-hearing friendly target speech without any interference.

THANK YOU!

ylf2017@sjtu.edu.cn