大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

Improving curriculum learning for target speaker extraction with synthetic speakers

Yun Liu, Xuechen Liu, Junichi Yamagishi @NII, Tokyo, Japan

Motivation:

- Target Speaker Extraction (TSE) isolates a specific speaker's voice from Mixed speech.
- Mixed speech is usually created by directly adding the target speaker and interference speakers.
- **Increase the diversity** of speaker characteristics can improve the TSE results.
- This study use a **speech generative model** to generate speech of diverse interference speakers, and then uses the generated data as part of the curriculum learning.

TSE with 3-stage Curriculum Learning (CL)

- **Improving curriculum learning for TSE with synthetic speakers**

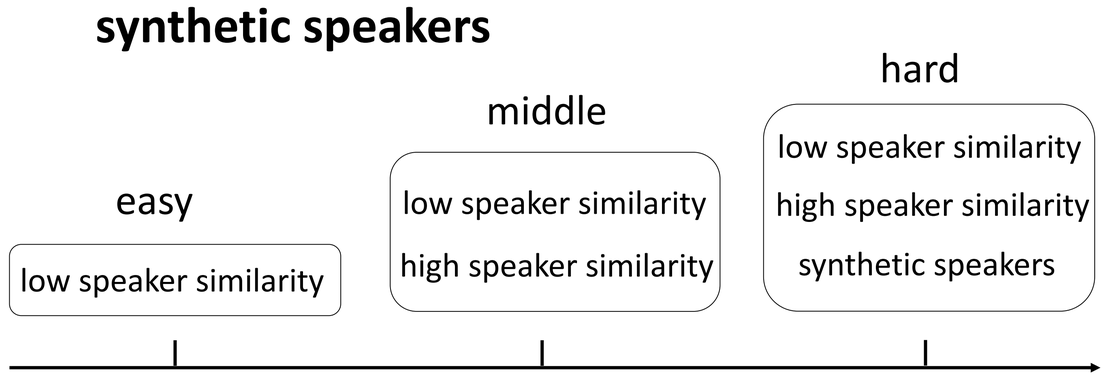


Fig. 1. Framework of the 3-stage Curriculum Learning.

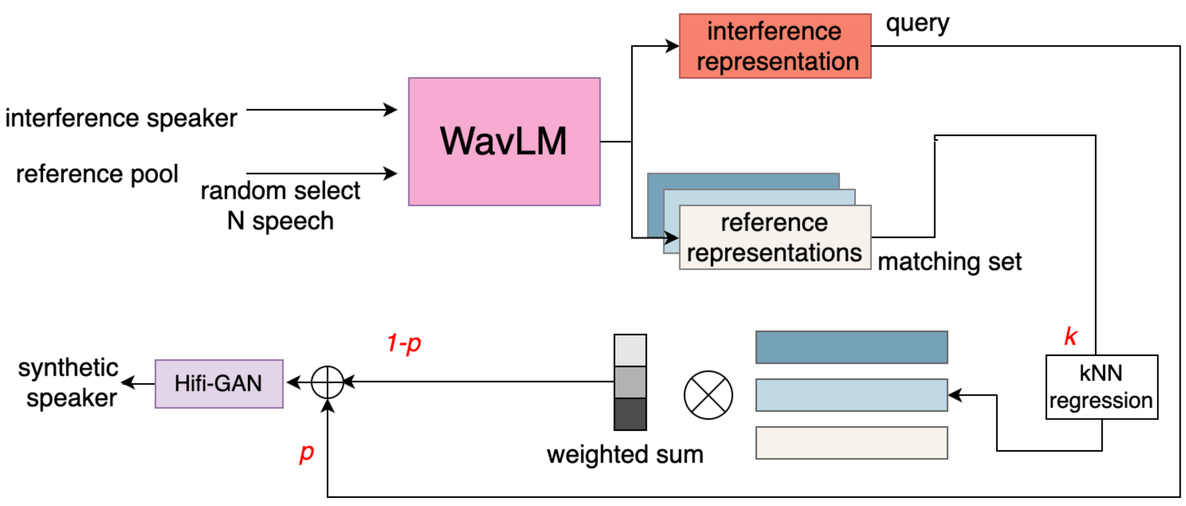


Fig. 2. Speech generative model [2] for generating synthetic speaker.

- **Step 1:** Use the k-NN algorithm to find similar reference speakers in the acoustic feature space.
- **Step 2:** Perform weighted mixing of the features from the reference speakers.
- **Step 3:** Generate the synthetic interference speaker and use a vocoder to produce the speech waveform.

Best configuration for experiment results:

k: the number of selected k-nearest neighbors, **k=4**

p: the preservation factor of retaining the original interference speaker, **p=4**

number of real data: 127k

number of synthetic data: 127k

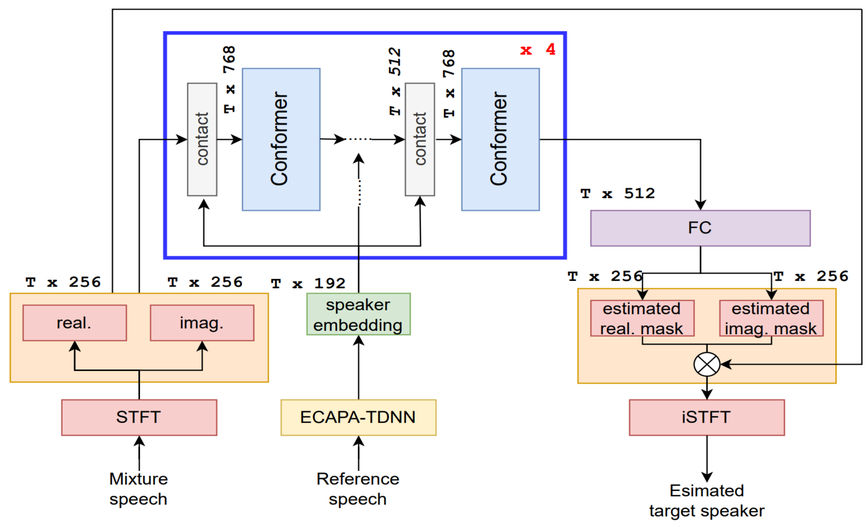


Fig.1 Framework of Conformer TSE

Experiment

Dataset[3]:

Dataset	Partition	#utterances	#speakers
Librisppech	train-360	50,800	921
	train-100	13,900	251
	test	3,000	40
Libri-2talker	train	127,056	1,172
	dev	2,344	1,172
	test	6,000	40

Extraction Model: Conformer **Speaker encoder**: ECAPA-TDNN

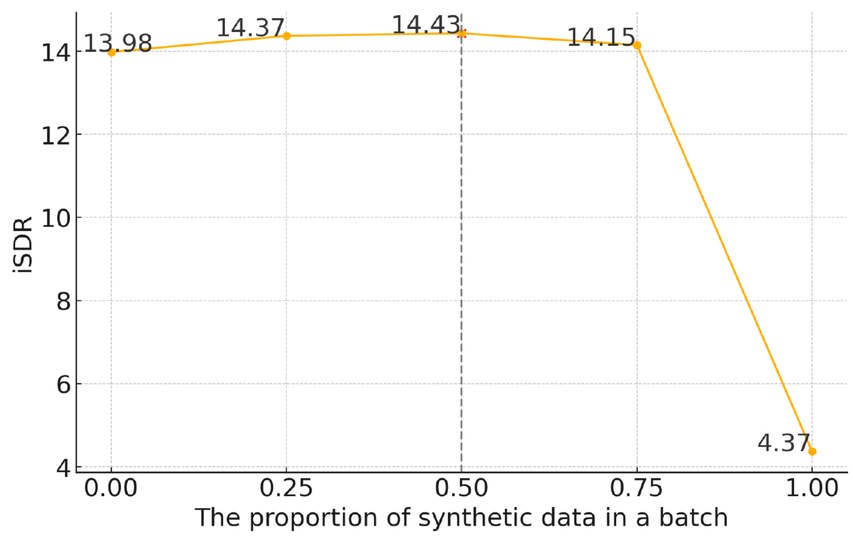
Loss function: SNR **Evaluation metric**: improved SDR(iSDR)

Results

- **iSDR(dB) results of using the Conformer TSE model**

TSE method	Stage2	Stage3	iSDR [dB]
Conformer w/o CL	×	×	12.57
Conformer	✓	×	13.44
Conformer (real only)	✓	✓	13.98
Conformer (real+syn)	✓	✓	**14.43**

- **iSDR(dB) results with different proportion of real and synthetic data**



- **iSDR(dB) results with different architectures**

TSE networks	base	real only	real+syn
Naive-BLSTM	8.30	9.79	**10.32**
SpeakerBeam [20]	9.78	10.10	**10.51**
VoiceFilter [21]	7.17	8.66	**9.32**

[1] Liu et al ., Target Speaker Extraction with Curriculum Learning. Interspeech 2024.
[2] Lv et al ., SALT: Distinguishable Speaker Anonymization Through Latent Space Transformation, ASRU2023
[3] Xu et al ., Target speaker verification with selective auditory attention for single and multi-talker speech. IEEE/ACM Transactions on audio, speech, and language processing