



Inter-University Research Institute Corporation /
Research Organization of Information and Systems
National Institute of Informatics



National University
SOKENDAI
The Graduate University for Advanced Studies

Target Speaker Extraction with Curriculum Learning

Yun Liu, Xuechen Liu, Xiaoxiao Miao, Junichi Yamagishi



- **1. Introduction**
- **2. Current problem**
- **3. Research Questions& Related work**
- **4. Experiments**
- **5. Results**



Target speaker extraction (TSE)

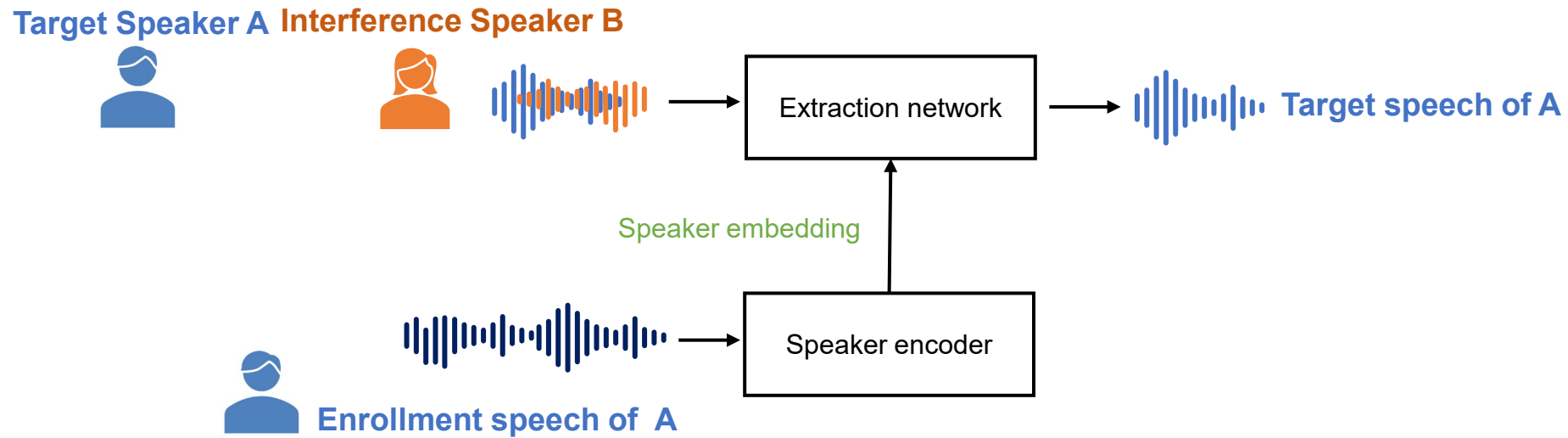
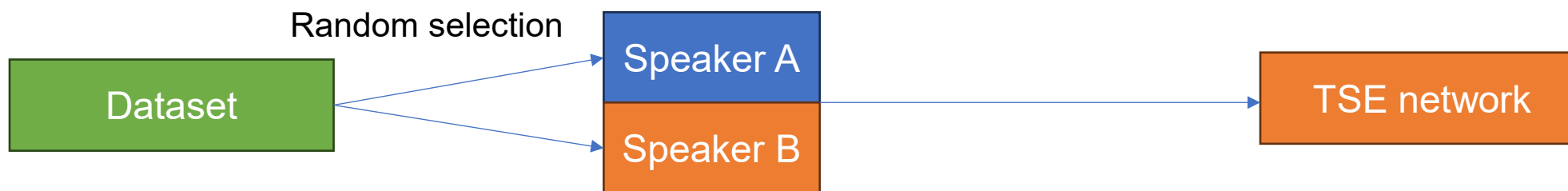


Fig1. Framework of TSE



Main Goal Improve the data efficiency and effectiveness for TSE training

Current Problem: less efficient and require too many effort for iteration!



However, training a model on all randomly selected data may not be optimal.

Some data is not suitable for use as the initial training data.



Many papers make the training data for TSE or speech separation more realistic and useful.

LibriCSS [1] achieves a more realistic training data by recording overlapped speech in a real setting, using multiple loudspeakers at different positions to play audio of a single speaker.

[2] suggests directly using real noisy training data with supervised learning for speech enhancement models to potentially reduce mismatches between training and inference.

Separation models trained with data augmentation [3] generalize better to unseen conditions and also enhance the robustness of the model.

when to use **what** data

[1] Chen, Zhuo, et al. "Continuous speech separation: Dataset and analysis." ICASSP 2020

[2] Z. Xu , et al. "Employing real training data for deep noise suppression," in ICASSP 2024

[3] A. Alex , et al. "Data augmentation for speech separation," Speech Communication 2023



RQ1 How to identify and select useful training data for more effective TSE training?

training data are considered as 'useful' when they are more effective to enhance the model's ability at the current stage of learning.

RQ2 How do we use the selected data at different training phase to improve TSE training?

Curriculum learning enhances TSE training by first using selected 'easy' data to establish foundational skills, then progressively introducing 'harder' data

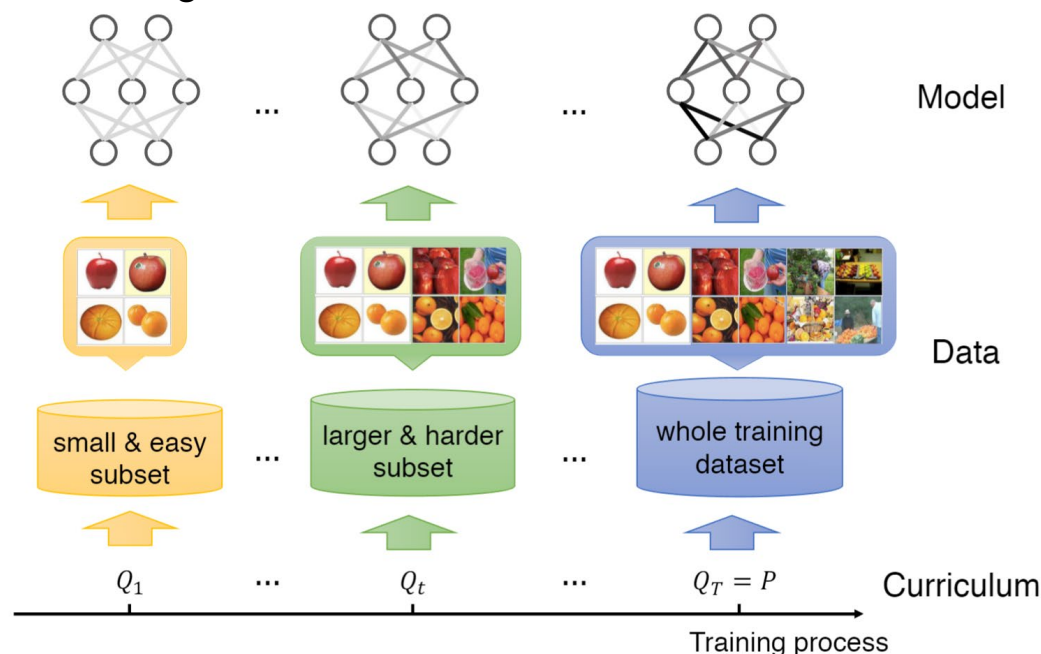


Figure1. Framework of the Curriculum Learning (CL)[1]

train from easier data to harder data



Dataset: Libri2talker(16 KHz)

Dataset	Partition	#utterances	#speakers
Libri-2talker	train	127,056	1,172
	dev	2,344	1,172
	test	6,000	40

Speaker encoder: ECAPA-TDNN pre-trained on the VoxCeleb2 dataset

Extraction network: 4 layer Conformer

Window length : 32 ms

Hop size : 8 ms

FFT length : 512

3 Model architecture



Frequency domain conformer

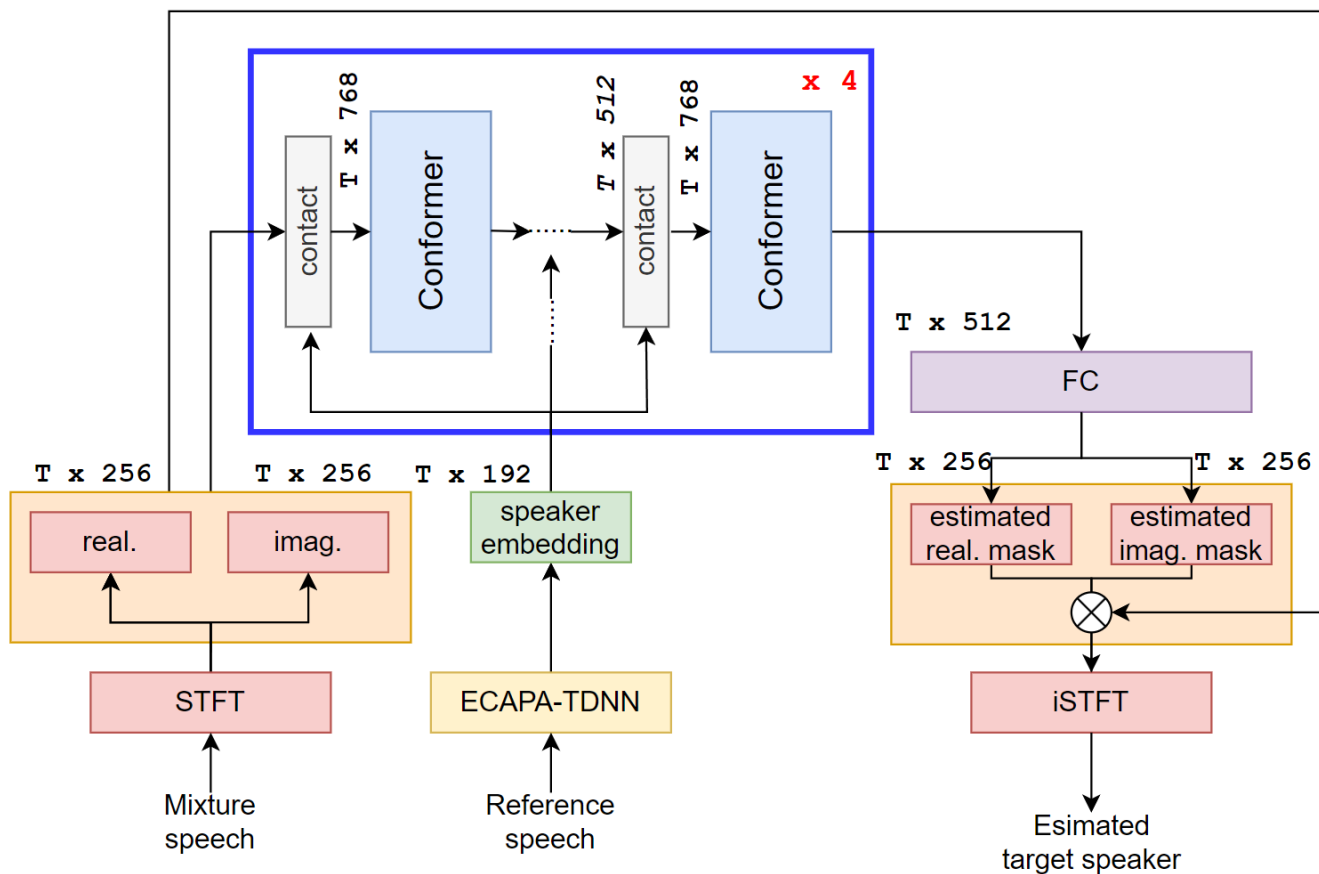
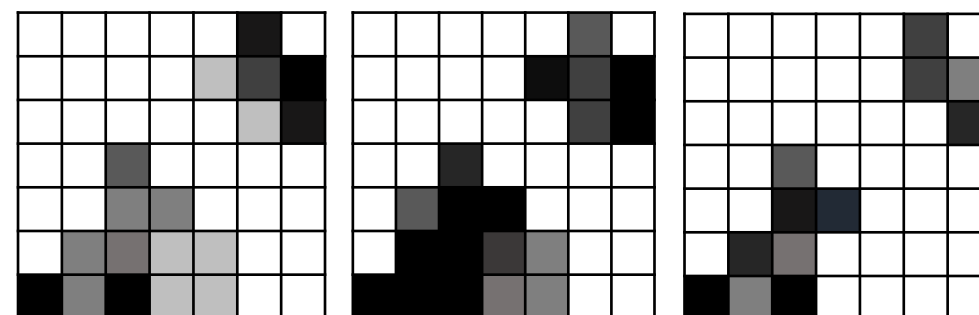


Fig 1. Network architecture of our CL-TSE



$$\text{Timex256} = \text{Timex256} \otimes \text{Timex256}$$

Estimated target speaker = Mixture \otimes Mask



RQ1

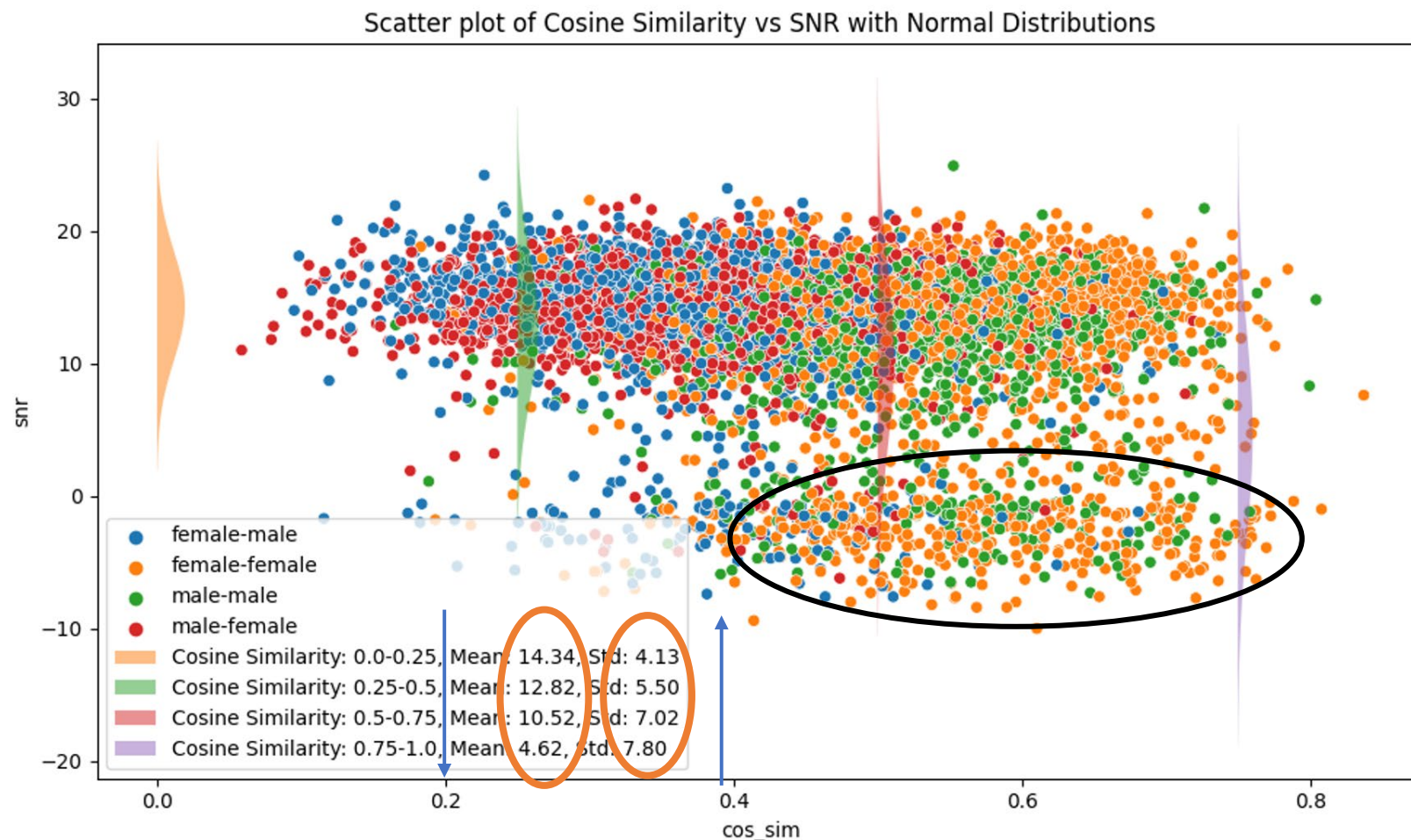
How to identify and select useful training data for more effective TSE training?

Difficulty measure: Identifying training samples are inherently more difficult than other samples

Target Speaker A

Interference Speaker B

- Speaker similarity
- Gender






RQ1

How to identify and select useful training data for more effective TSE training?

Difficulty measure: Identifying training samples are inherently more difficult than other samples

- Signal-to-noise ratio (SNR) $SNR = 10 \log_{10} \frac{||\text{Signal}||^2}{||\text{Noise}||^2} \longrightarrow SNR = 10 \log_{10} \frac{||\text{Target}||^2}{||\text{Interference}||^2}$
- Signal-to-distortion ratio (SDR)


output of model



\hat{S}

← Error = $\hat{S} - S$ →

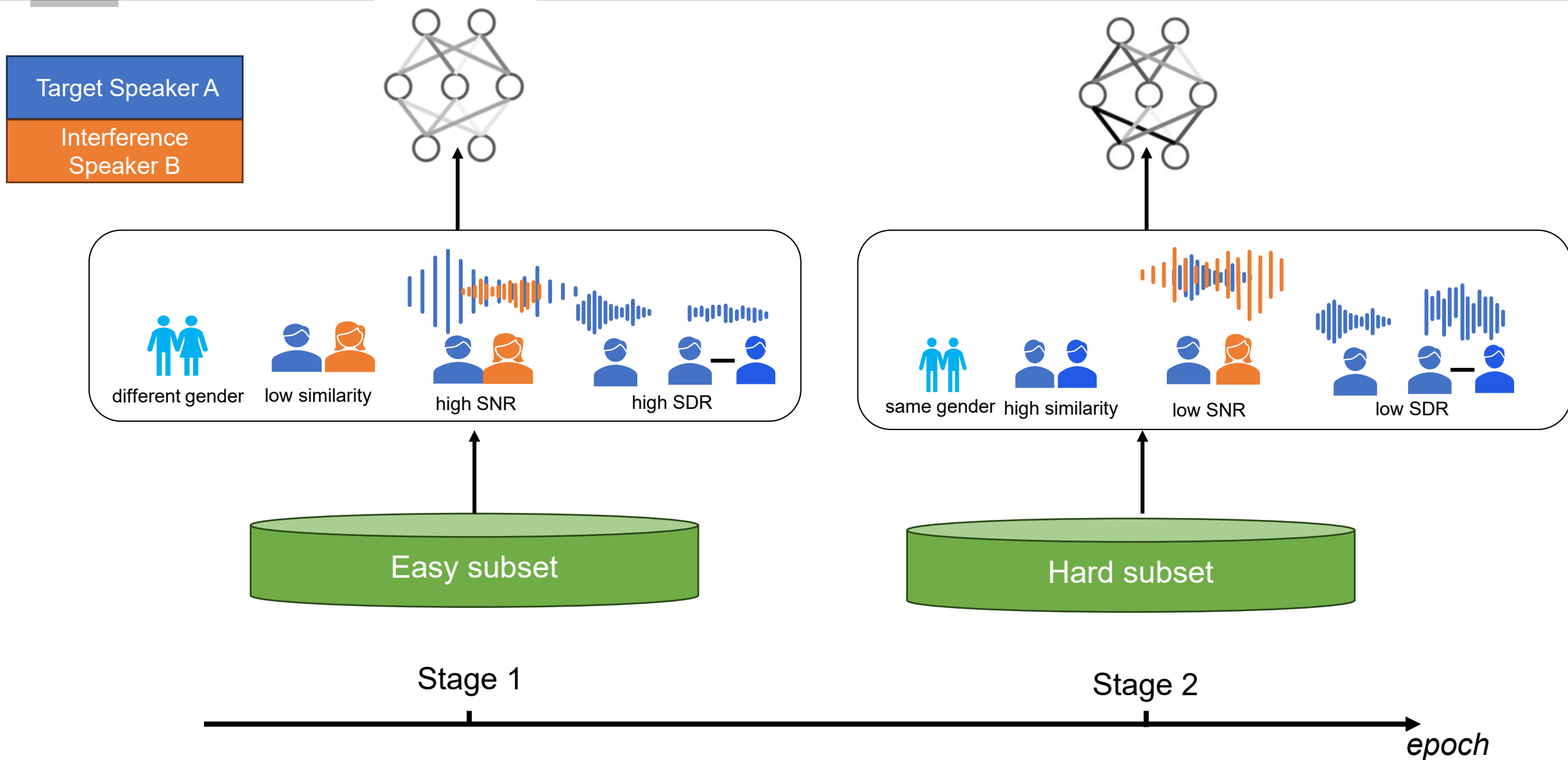
ground truth



S

$SDR = 10 \log_{10} \frac{||\text{Target}||^2}{||\text{Error}||^2}$

Loss function (SNR) : **negative** Signal-to-noise ratio



Training scheduler: Determining when and which hard data should be introduced to the training?



RQ2 How do we use the selected data at different training phase to improve TSE training?

Evaluation metric(dB): improved SDR = $SDR_{(estimated,target)} - SDR_{(mixture,target)}$

	Optimal τ	Used data	1st phase(100 epoch)	2nd phase(5 epoch)
random	-	100%	12.50	12.57
τ gender(same/different)	Different	50.0%	7.64	11.40
τ spk \in [0.5, 0.6, 0.7]	0.6	83.8%	13.03	13.44
τ SDR \in [-5, 5]dB	1	44.3%	8.75	13.40
τ SNR \in {0,5,10}dB	10	82.4%	12.79	12.99

Table 1: iSDR(dB) results(**high is better**) on the test set using the ***predefined*** CL-based TSE.

1st phase	2nd phase	3rd phase	4th phase	5th phase
3.30	7.79	10.93	12.66	13.54

Table 2: iSDR(dB) results on the test set using the ***self-paced*** CL-based TSE.



RQ2

How do we use the selected data at different training phase to improve TSE training?

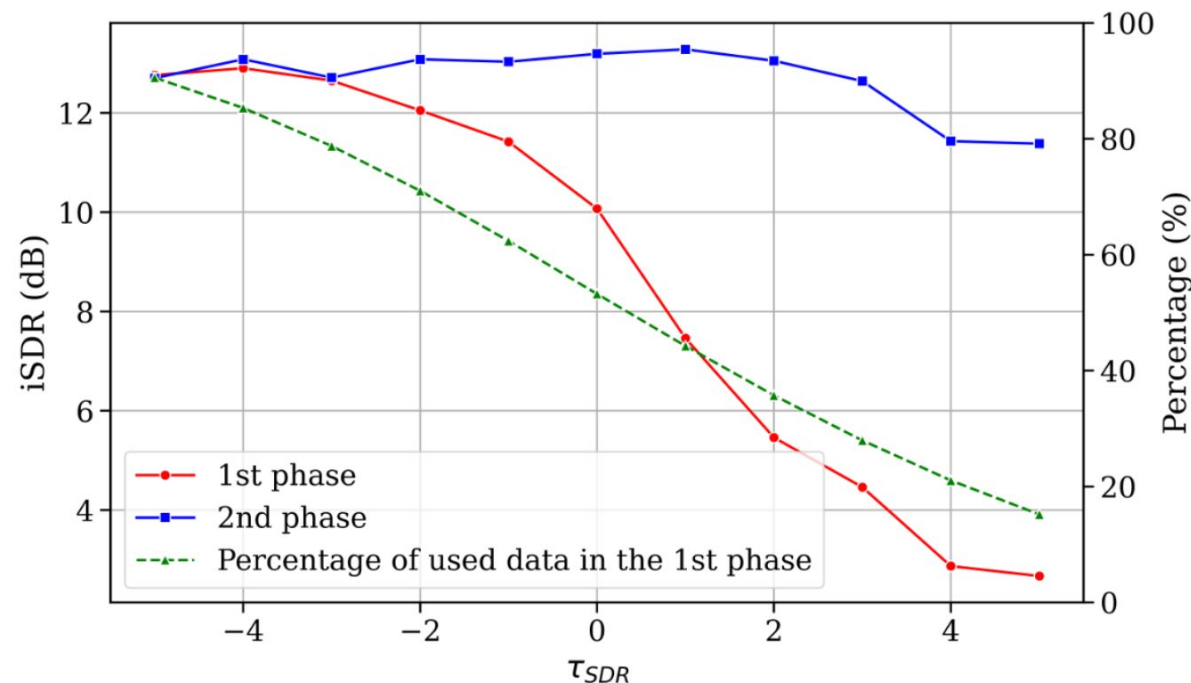


Figure 1: iSDR changes on the dev set with τ_{SDR} in two training phases, along with data usage percentage in the 1st phase.

The selection of the optimal threshold to balance between 'easier' and more complex data is sensitive and crucial



Networks	random	similarity(0.6)	self-paced
Naive-BLSTM	8.09	8.30	9.60
SpeakerBeam[28]	9.47	9.78	9.88
VoiceFilter [29]	6.90	7.17	7.54

Table 1:iSDR(dB) of CL methods with different architecture.

Q&A