Training Dynamics-Aware Multi-Factor Curriculum Learning for Target Speaker Extraction

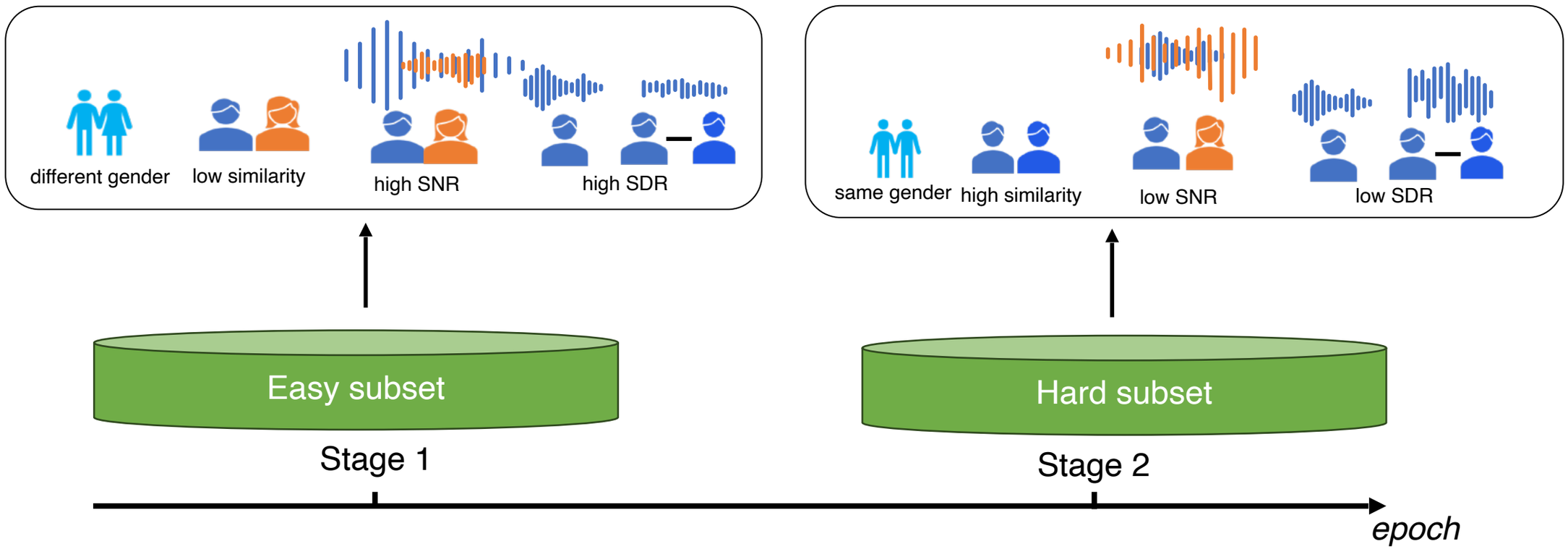
Yun Liu, Xuechen Liu, Xiaoxiao Miao, Junichi Yamagishi
National Institute of Informatics
Japan

Agenda

- Curriculum learning for TSE
- Curriculum learning with synthetic speakers
- Research Questions
- RQ1: Multi-factor curriculum learning
  - Database
  - Model
  - Experiments
- RQ2: Can we design a good curriculum with fewer thresholds?
  - Data Map
  - Experiments

Curriculum learning for target speaker extraction

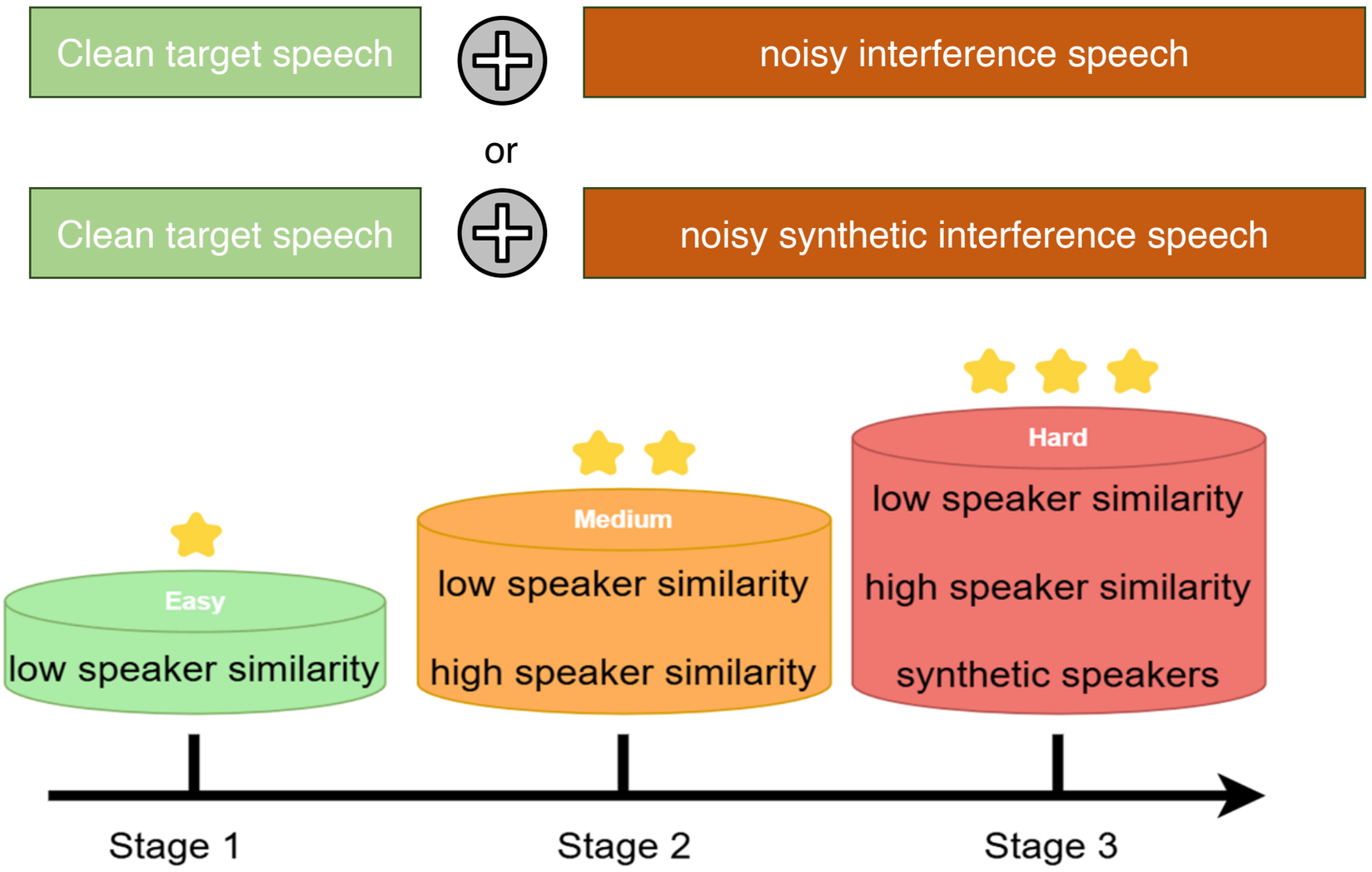
- In our previous study [1], we confirmed that creating a curriculum for training a target speaker extraction (TSE) model based on <u>one of the following factors</u> improves the TSE performance
  - target-interfering speaker similarity
  - Signal-to-Noise Ratio (SNR)
  - Signal-to-Distortion Ratio (SDR)
- The curriculum could also use other factors, such as the overlap ratio and the number of interference speakers in the mixture



[1] Yun Liu, Xuechen Liu, Xiaoxiao Miao, Junichi Yamagishi, “Target Speaker Extraction with Curriculum Learning” Interspeech 2024 4348-4352

Curriculum learning with synthetic speakers

- It has also been confirmed that introducing data partially synthesized by the latest speech generative models (e.g. synthetic interfering speakers in the mixture) is effective in curriculum learning for TSE [2][3]



[2] Yun Liu, Xuechen Liu, Junichi Yamagishi, “Improving Curriculum Learning For Target Speaker Extraction With Synthetic Speakers” 2024 IEEE Spoken Language Technology Workshop (SLT) 364-370, 2024
[3] Yun Liu, Xuechen Liu, Xiaoxiao Miao, Junichi Yamagishi, ``Target Speaker Extractor Training with Diverse Speaker Conditions and Synthetic Data,'' APSIPA Transactions on Signal and Information Processing: Vol. 14: No. 1, e30, 2025

Two research questions

- **RQ1**
  - *In curriculum learning, when multiple factors are controlled simultaneously rather than just one, does TSE performance improve compared to controlling only a single factor?*
  - The expected answer is ‘yes’.
  - However, the manual setup is complicated and time-consuming 😅.

- **RQ2**
  - *Is it possible to design the learning curriculum based on the main objective function only with fewer thresholds?*
  - *How much worse is the resulting model compared to one obtained from manually designed multi-factor curricula?*
  - Ideally, the degradation should be minimal.

Factors investigated for RQ1

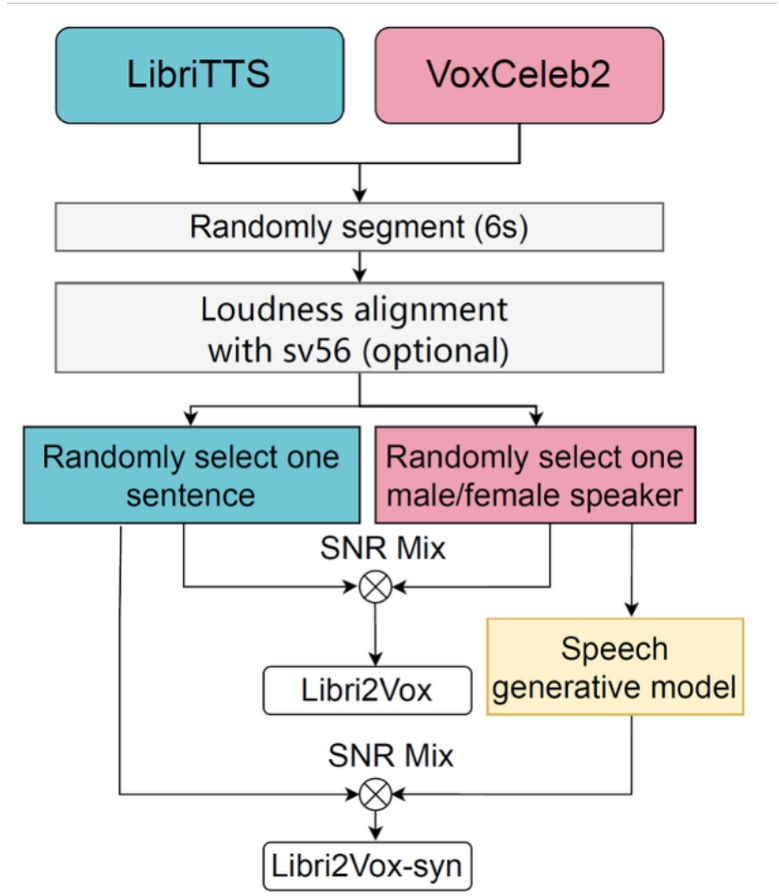
- Control the following factors individually or simultaneously for curriculum:
  - **SNR**
    - {0, 5, 10, 15 dB}
  - **Number of interfering speakers**
    - {1, 2, 3}
  - **Temporal overlap ratio of target and interfering speakers**
    - {0, 0.2, 0.4} where 0 means fully overlap and 1 means no overlap
  - **Synthetic and real interfering speakers**
    - {real, synthetic, real & synthetic}

- The first author manually adjusted and defined the difficulty level for each factor for the three-stage curriculum learning
  - Details are available in the paper

Dataset: the Libri2vox dataset

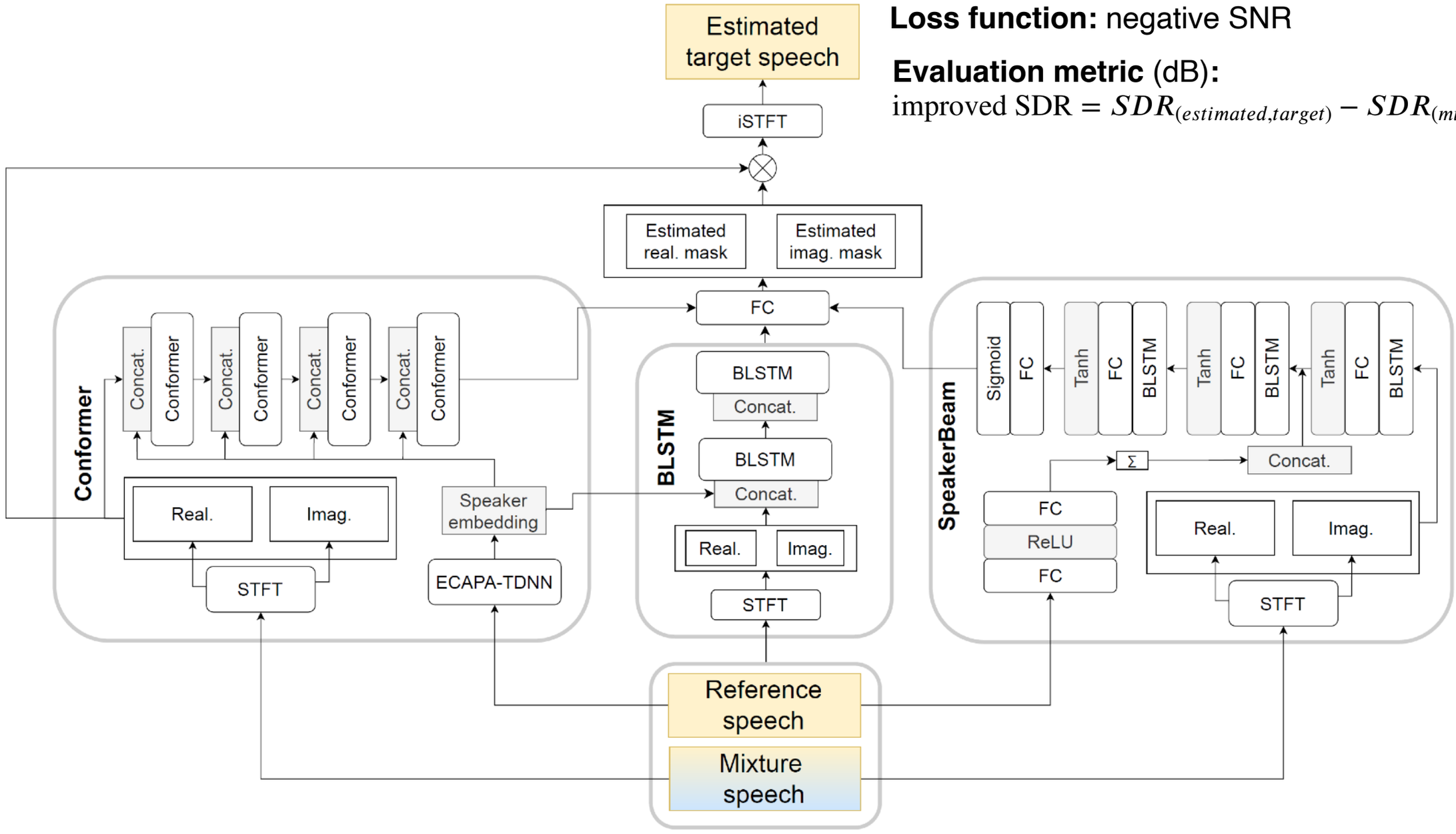
- Curriculum learning and its evaluation utilize the ***Libri2Vox dataset*** [3] that we previously constructed
- The target speakers were selected from LibriTTS, while the interfering speakers were selected from either the VoxCeleb dataset or a synthesized version of the VoxCeleb dataset.
- The number of interfering speakers in the evaluation set ranges from 1 to 3, and they are referred to as follows:
  - **2-speaker mixtures**: target + 1 interferer
  - **3-speaker mixtures**: target + 2 interferers
  - **4-speaker mixtures**: target + 3 interferers

Program to construct the Libri2Vox dataset



[3] Yun Liu, Xuechen Liu, Xiaoxiao Miao, Junichi Yamagishi, ``Target Speaker Extractor Training with Diverse Speaker Conditions and Synthetic Data,'' APSIPA Transactions on Signal and Information Processing: Vol. 14: No. 1, e30, 2025

TSE models



Loss function: negative SNR

Evaluation metric (dB):

improved SDR $= SDR_{(estimated,target)} - SDR_{(mixture,target)}$

Since the BLSTM, Conformer, and SpeakBeam models showed similar trends in our analysis of curriculum learning, only the BLSTM based TSE results are presented in the following slides

Experimental results for RQ1

- As expected, controlling multiple factors simultaneously for curriculum design yields better performance than learning the TSE model by controlling individual factors alone
- The same trend holds when the number of interfering speakers increases

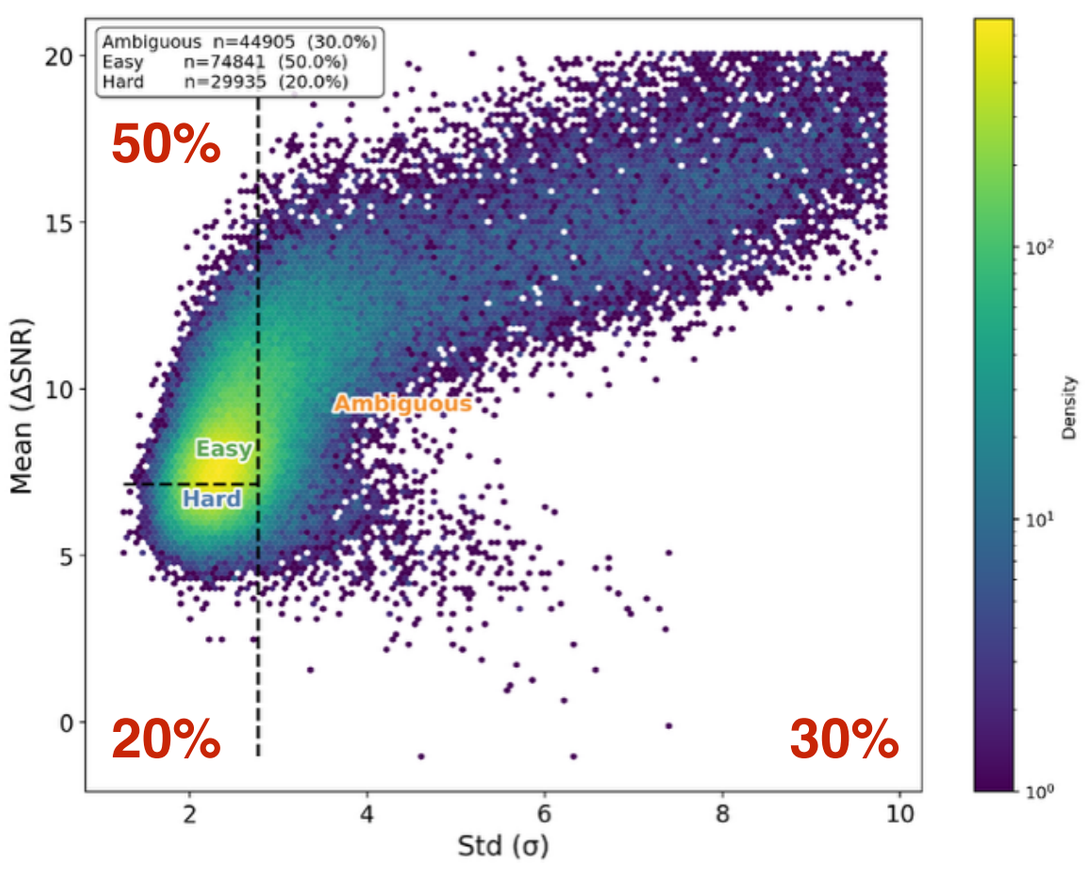
Number of speakers in mixtures

Control individual factors alone

Exp.	Curriculum Stages	iSDR 2spk	iSDR 3spk	iSDR 4spk
baseline	random sampling	12.38	8.56	7.16
baseline-CL	[[1,0,10,0,real], [2,0,10,0,real], [3,0,10,0,real]]	12.58	8.78	7.40
single factor: #snr	[[1,5,10,0,real], [2,0,10,0,real], [3,0,5,0,real]]	13.04	9.71	8.62
single factor: #overlap ratio	[[1,0,10,0,real], [2,0,10,0.2,real], [3,0,10,0.4,real]]	12.62	9.87	8.76
single factor: #inter_source_syn	[[1,0,10,0,syn], [2,0,10,0,syn], [3,0,10,0,syn]]	11.22	8.43	9.19
single factor: #inter_source_real/syn	[[1,0,10,0,real/syn], [2,0,10,0,real/syn], [3,0,10,0,real/syn]]	12.73	9.51	8.96
single factor: #speaker count	[[1,0,10,0,real], [1,0,10,0,real], [1,0,10,0,real]]	12.80	7.79	6.73
multiple factors	[[1,5,10,0,real/syn], [2,0,10,0.2,real/syn], [3,0,5,0.4,real/syn]]	**13.22**	**10.08**	**9.21**

RQ2: Design the curriculum using data map

- Next, we would like to define the good curriculum based solely on the information of the objective function (SNR) with fewer thresholds
- Data Map
  - The mean and standard deviation of the SNR improvements for each sample over the past 50 epochs
  - Define the top 30% of samples with a large standard deviation as **Ambiguous**, the top 50% of samples by mean value from the remaining as **Easy**, and the remaining samples as **Hard**.



Interpretations of the three regions

- **Easy:** easy to extract the target speaker
- **Ambiguous**: TSE results may be accurate or inaccurate
- **Hard**: hard to extract the target speaker

Experimental results for RQ2

- Select training samples from each region at different stages of curriculum learning and add them to the training set
- Example: E / A / H
  - 1st stage: Easy samples, 2nd stage: +Ambiguous samples, 3rd stage: +Hard samples.
- The results are naturally worse than when multiple factors are manually controlled, but they are close when training samples are selected from the appropriate region during the curriculum

Order	iSDR 2spk	iSDR 3spk	iSDR 4spk
baseline	12.38	8.56	7.16
multiple factors	**13.22**	**10.08**	9.21
E / A / H	13.15	9.85	**9.32**
E / H / A	12.93	9.63	9.18
A / E / H	12.82	9.61	9.22
A / H / E	12.96	9.72	9.17
H / E / A	12.77	9.54	**9.32**
H / A / E	12.90	9.63	9.10
E / A / H (forgetting)	8.83	5.43	5.52
single factor: #snr	[[1,5,10,0,real], [2,0,10,0,real], [3,0,5,0,real]]	13.04	9.71	8.62

- **Easy: E**
- **Ambiguous: A**
- **Hard: H**

Conclusion

- Deep analysis of curriculum learning in TSE
- RQ1
  - Designing a curriculum by considering multiple factors simultaneously, rather than just one, results in better-performing TSE models.
- RQ2
  - To reduce the labor required for manual threshold adjustment, we proposed a method that records the SNR variation—the objective function for TSE learning—for each training sample
  - Using the mean and standard deviation of the SNR variation, the data is divided into Easy, Ambiguous, and Hard regions.
  - It was found that TSE curriculum training in the order of "***Easy*** → ***+Ambiguous*** → ***+Hard***" can achieve performance close to one, with a curriculum that considers multiple factors simultaneously