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

National University
SOKENDAI
The Graduate University for Advanced Studies

Training and Data Strategies for Target Speaker Extraction

PhD Candidate: Yun Liu

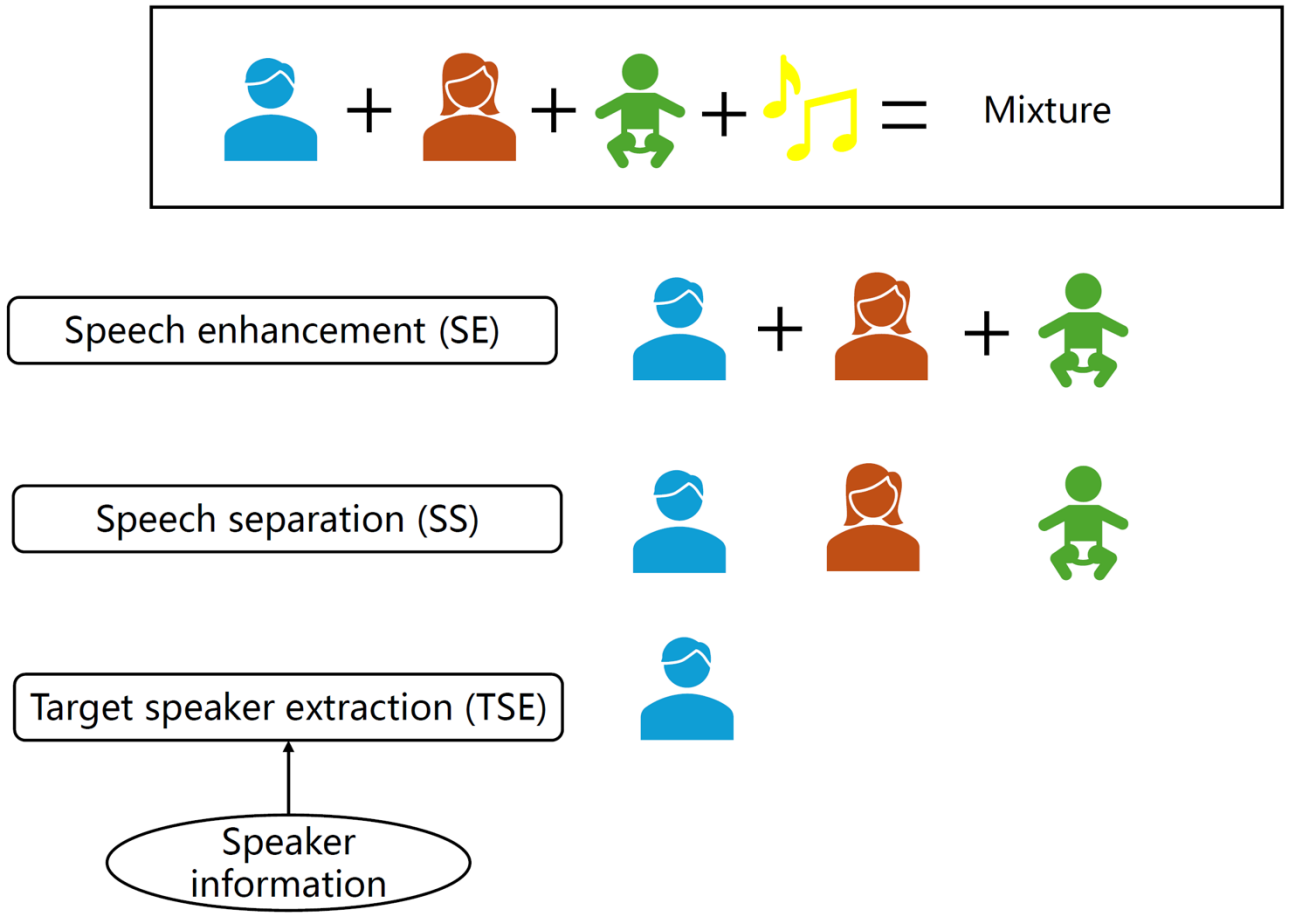
Supervisor: Prof. Junichi Yamagishi

Outline

1. Introduction
2. Research questions
3. Curriculum Learning to Select & Schedule Training Data (Chapter 3)
4. Curriculum Learning with Synthetic Speakers (Chapter 4)
5. Libri2Vox: More Diverse Speaker Conditions (Chapter 5)
6. Multi-factor Curriculum Learning with Training Dynamic (Chapter 6)
7. Conclusion and Future work

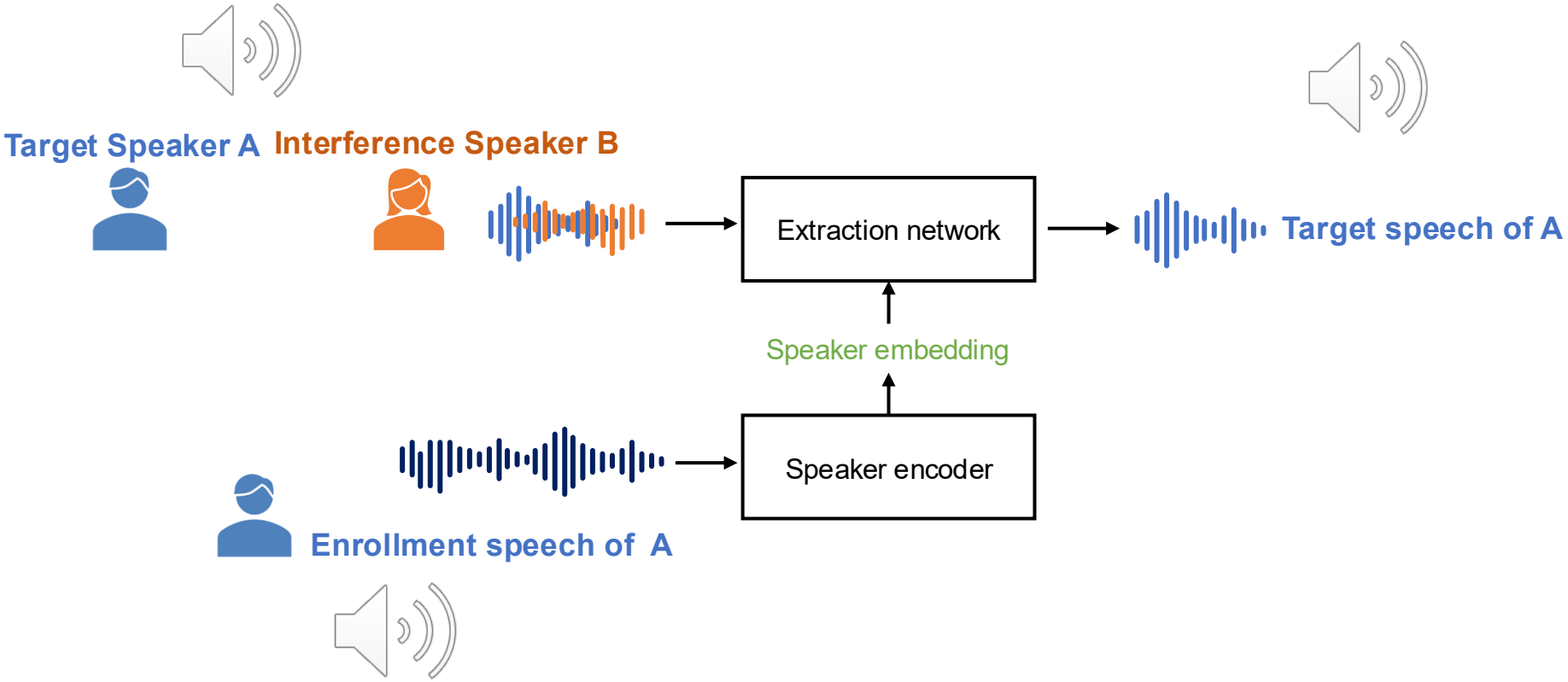
1. Introduction

Cocktail party problem



1 Introduction

Target speaker extraction (TSE)



Society impact of TSE

For down-stream tasks

- **Data pre-processing for speech synthesis using in-the-wild data**
- **Front-end for automatic speech recognition (ASR)**

For end-users

- **Hearing assistance**
- **Teleconferencing for remote work**

Limitations of Current TSE Training

Data-side limitation

- **Limited number of speakers** / **recording conditions**.
- Acoustic conditions: clean target speech + simple additive mixing with noise/interferers.
- **Poor generalization** to unseen speakers and real-world mixtures

Training-side limitations

- Current training relies on **random sampling** from all mixtures (no curriculum).
- No **stage-wise progression** from during training.
- No controlled design: speaker similarity, SNR, number of interferers, overlap ratio, or real vs. synthetic.

Limitations of Current TSE Training

Limits of current synthetic data use

- mainly replaces clean target and interference speech → not match the case: real speech in-the-wild
- mixed with real data without strategy → little gain
- distribution gap: synthetic vs real → easy to separate → No contribution for training

2. Research questions

2 Research Questions

Main question **Can selecting and creating more effective training data improve TSE?**

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

RQ2 **How do we use the selected data at different training stages to improve TSE performance?**

[1] Y. Liu, X. Liu, X. Miao, and J. Yamagishi, “Target speaker extraction with curriculum learning,” in Proc. Interspeech, 2024, pp. 4348-4352.

RQ3 **Can we use speech generative models to generate useful data for training TSE?**

[2] Y. Liu, X. Liu, and J. Yamagishi, “Improving curriculum learning for target speaker extraction with synthetic speakers,” 2024 IEEE SLT, 2024. pp. 364-370

[3] Yun Liu, X. Liu, X. 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.

RQ4 **How can we coordinate multiple complexity factors in challenging multi-speaker scenarios?**

[4] Y. Liu, X. Liu, J. Yamagishi, “Training Dynamics-Aware Multi-Factor Curriculum Learning for Target Speaker Extraction,” Proc. ICASSP 2026 (**accepted**).

2 Research Questions (Challenge)

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

- Existing TSE training uses **random sampling**, ignoring that samples have different difficulty levels.
- Models often **fail on hard cases**, especially when **target and interference speakers sound very similar**

RQ2 **How do we use the selected data at different training stages to improve TSE performance?**

- The order of training examples affects learning, but current methods mix easy and hard samples randomly.
- Without a structured schedule, the model can not well learn progressively from easy to hard cases.

RQ3 **Can we use speech generative models to generate useful data for training TSE?**

- Standard TSE datasets have **limited speaker diversity** (usually < 1,200 speakers).
- This small speaker set hurts generalization to unseen speakers, compared with ASR trained on datasets with thousands of speakers.

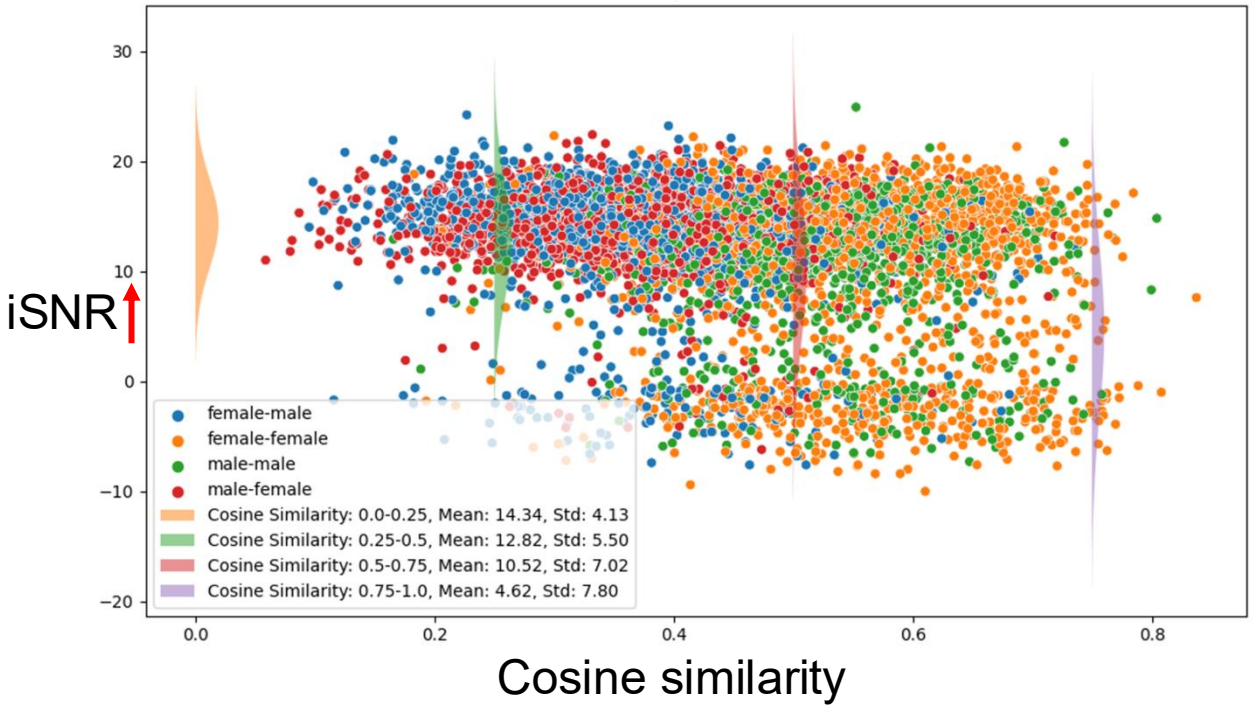
RQ4 **How can we coordinate multiple complexity factors in challenging multi-speaker scenarios?**

- Previous frameworks **only handle one difficulty factor** at a time.
- Hard to design curricula without knowing **how the model actually perceives difficulty**.

3. Curriculum Learning to Select & Schedule Training Data

3 Curriculum Learning to Select & Schedule Training Data

Motivation – Why Curriculum Learning for TSE?



Scatter plot of cosine similarity vs iSNR.

$$\text{iSNR} = SNR_{(estimated,target)} - SNR_{(mixture,target)}$$

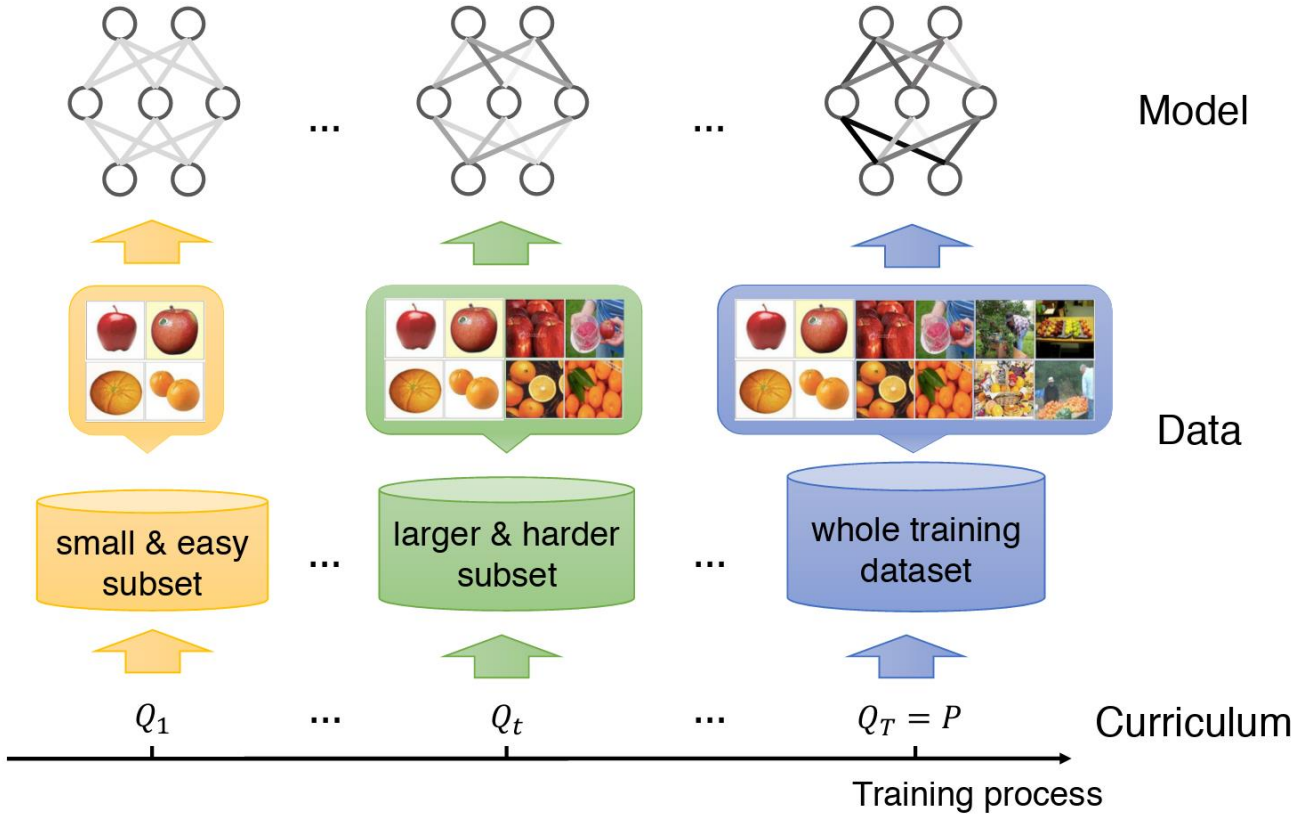
1.TSE mixtures have a **natural difficulty structure**
– Different-gender, high-SNR mixtures are easier;
– Same-gender, high speaker similarity,
low SNR mixtures are much harder.

2. **Error analysis** shows most failures happen on these **hard mixtures**
– high speaker similarity;
– low SNR / complex overlaps.

3. Standard training with **random sampling** mixes easy and hard samples
– causes unstable training;
– model cannot gradually build skills from easy to hard cases.

3 Curriculum Learning to Select & Schedule Training Data

Curriculum Learning



Flowchart of the Curriculum Learning (CL) [1][2]

[1] Wang, X., Chen, Y., & Zhu, W. (2021). A Survey on Curriculum Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 4555-4576.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. ICML 2009, 41–48.

3 Curriculum Learning to Select & Schedule Training Data

What is a Difficulty Measure?

A **difficulty measure** quantifies *how hard a training sample is* for a given model.

Defined **at the sample level**

Can be:

- **Knowledge-driven**
  - *(e.g., SDR, speaker similarity, overlap ratio)*
- **Data-driven / model-driven**
  - *(e.g., training loss, ΔSNR, model confidence)*

Purpose:

→ Rank, filter, or weight training samples by difficulty

What is a Training Scheduler?

A **training scheduler** determines *when and how different samples are used during training*.

Defined **at the training process level**

Controls:

- sample inclusion / exclusion
- sample weighting
- stage transitions (easy → hard)

Purpose:

→ Organize the learning process over time

3 Curriculum Learning to Select & Schedule Training Data

RQ1 **How to select training data for more effective TSE training? [1]**

We can select the training data with gender, speaker similarity, Signal-to-Noise Ratio (SNR) and Signal-to-Distortion Ratio (SDR)

RQ2 **How can we use the selected data at different training stage to improve TSE training? [1]**

We can use curriculum learning to enhances TSE training by first using selected 'easy' data to establish foundational skills, then progressively introducing 'harder' data.

Speaker-dependent factors:

- **Cosine similarity**: Target vs Interference $\text{sim} = \cos\left(E(s_0(n)),\ E(s_i(n))\right)$
- **Gender**: Same gender vs Different gender

Mixture-dependent factors

- **SDR** : Target vs Interference

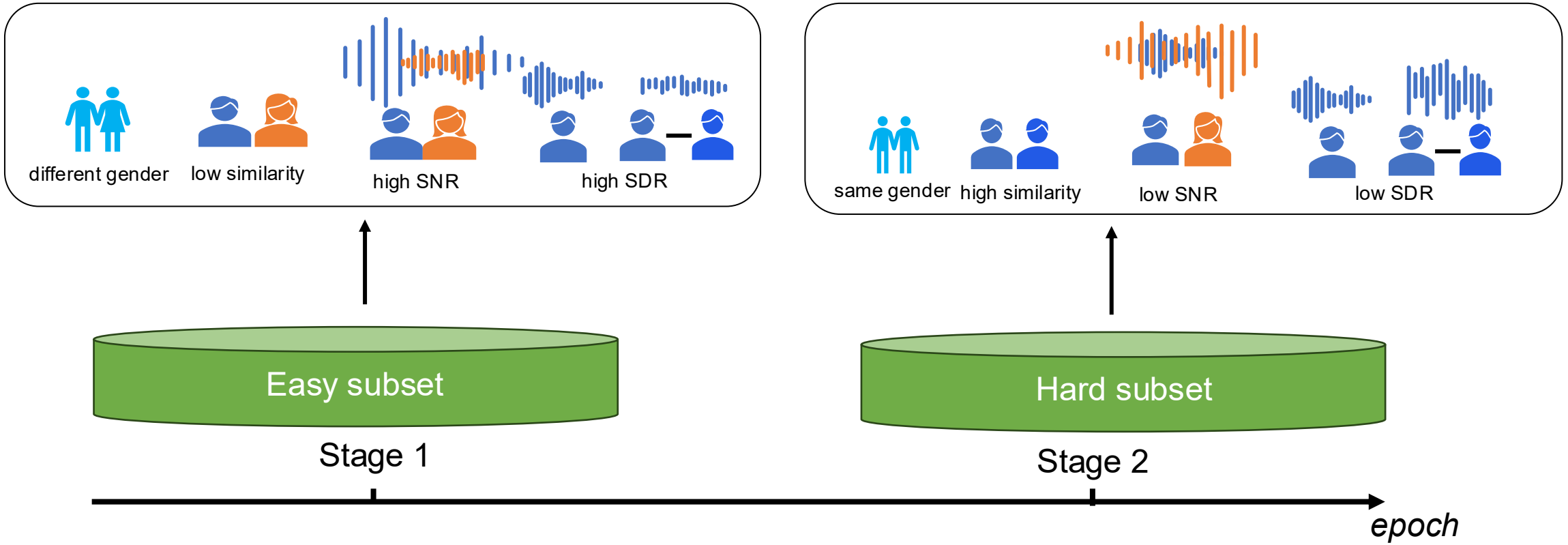
$$\text{SDR} = \frac{1}{N}\sum_{n=1}^{N} 10\log_{10}\left(\frac{\|s_0(n)\|^2}{\|s_i(n)\|^2}\right)$$

- **SNR** : computed using a pre-trained TSE model
  - Model-dependent measure capturing **current weaknesses** of the extractor.

$$\text{SNR} = 10\log_{10}\frac{\sum_{n=1}^{N}|s_0(n)|^2}{\sum_{n=1}^{N}|\hat{s}_0(n) - s_0(n)|^2}$$

[1] Y. Liu, X. Liu, X. Miao, and J. Yamagishi, "Target speaker extraction with curriculum learning," in INTERSPEECH2024, 2024, pp.4348–4352.

3 Curriculum Learning to Select & Schedule Training Data



Flowchart of CL-based TSE

3 Curriculum Learning to Select & Schedule Training Data

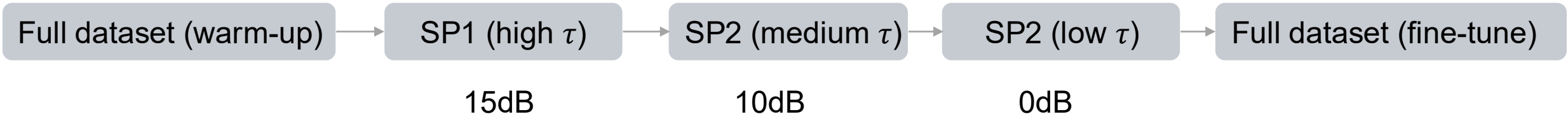
Two-stage CL



Self-paced CL:

- Use the **training loss** as a dynamic measure.
- In each mini-batch, only samples with loss > τ are back-propagated.
- Gradually decrease τ to include harder samples over time.

Self-paced (SP) CL



3 Curriculum Learning to Select & Schedule Training Data

Experiments setting

Datasets

Dataset	Partition	#utterances	#speakers
Libri2talker	train	127,056	1,172
	dev	2,344	1,172
	test	6,000	40

Two-speaker mixtures:
1 target + 1 interferer in each mixture.

Evaluation and Models

- Evaluate CL on **several TSE Models** (Conformer/BLSTM/SpeakerBeam/VoiceFilter).
- Speaker embedding was extracted by a pre-trained ECAPA-TDNN
- Front-end, loss function, and optimization are **kept the same**,

only the **data selection and training schedule** are changed by CL.

Training Strategy

- Baseline: **random sampling** over all training mixtures.
- CL methods:
  - build **easy vs. hard subsets** using speaker similarity, SNR, and SDR;
  - train the model with **two-stage** or **self-paced** curriculum from easy to hard.

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

3 Curriculum Learning to Select & Schedule Training Data

Results

	Optimal τ	Used data	1st phase	2nd phase
random	-	100%	12.50	12.57
τ_{gender}	Different	50.0%	7.64	11.40
τ_{spk}	0.6	83.8%	13.03	**13.44**
τ_{SDR}	1	44.3%	8.75	13.40
τ_{SNR}	10	82.4%	12.79	12.99

iSDR(dB) results on the test set using the **pre-defined** CL-based TSE.

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

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

4. Curriculum Learning with Synthetic Speakers

4 Curriculum Learning with Synthetic Speakers

RQ3 **Can we use speech generative models to generate useful data for training TSE? [2][3]**

By incorporating synthetic speakers through voice conversion into our curriculum learning approach, we enhanced both data diversity and model performance.

Motivation for RQ3

• TSE datasets still have limited speaker diversity and recording conditions; collecting more clean, multi-condition data is expensive.

• Speech generative model (VC / TTS) can create many controlled synthetic speakers without extra recordings.

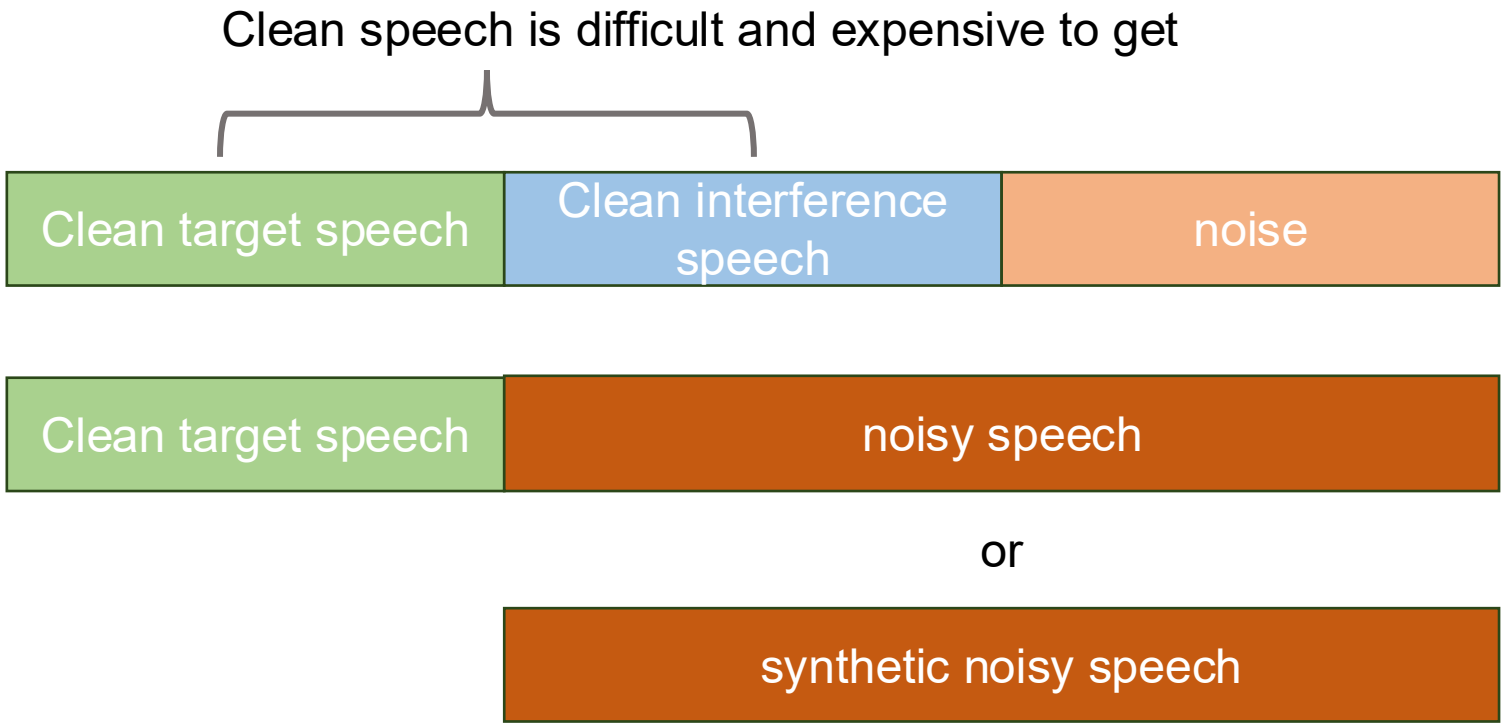
• We extend the CL framework by adding a synthetic-speaker stage: first train on real speakers, then use synthetic speakers as harder curriculum to further improve robustness.

[2] Y. Liu, X. Liu, and J. Yamagishi, “Improving curriculum learning for target speaker extraction with synthetic speakers,”2024 IEEE SLT, 2024. pp. 364-370
[3] Yun Liu, X. Liu, X. 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.

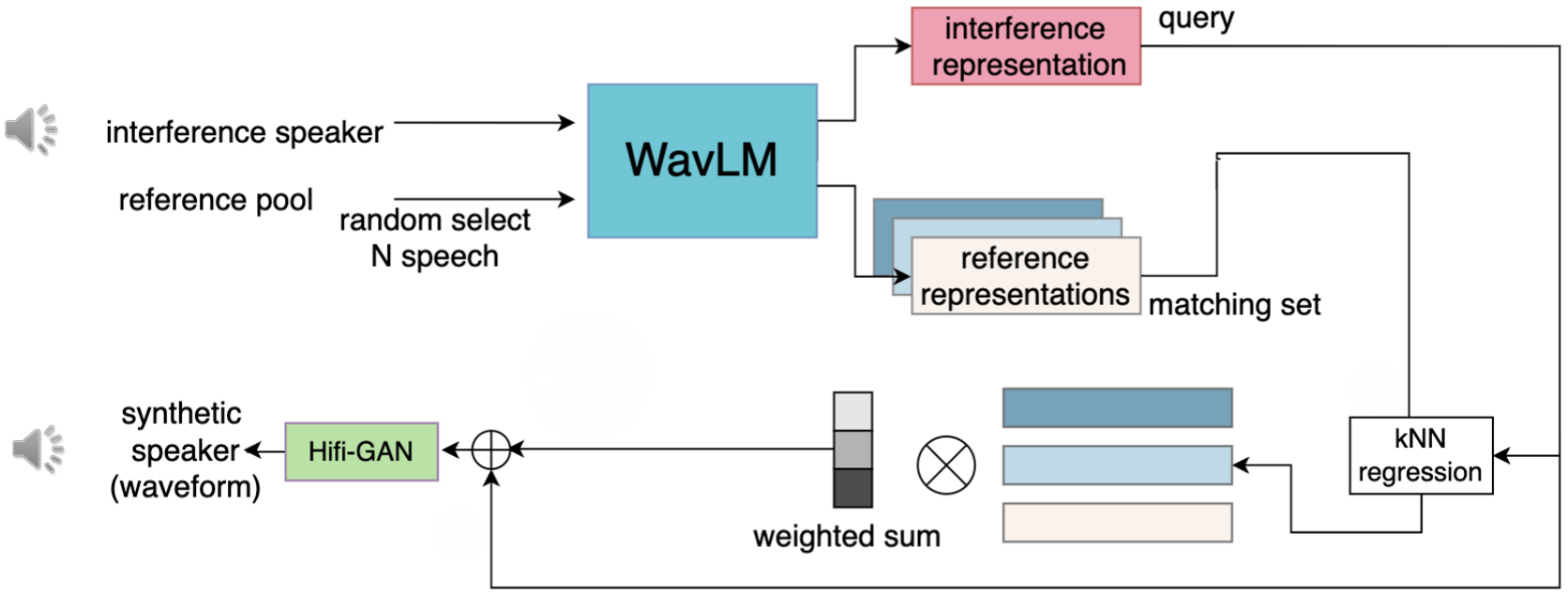
4 Curriculum Learning with Synthetic Speakers

Traditional methods: clean speech + interference speech + noise

Our methods: clean speech + noisy speech data in the wild



4 Curriculum Learning with Synthetic Speakers



Data generation framework by SALT. [1]

[1] Y. Lv, J. Yao, P. Chen, H. Zhou, H. Lu, and L. Xie, “SALT: Distinguish-able speaker anonymization through latent space transformation,” in2023 IEEE Automatic Speech Recognition and Understanding (ASRU),2023.

Why voice conversion (VC) not Text-to-Speech (TTS)?

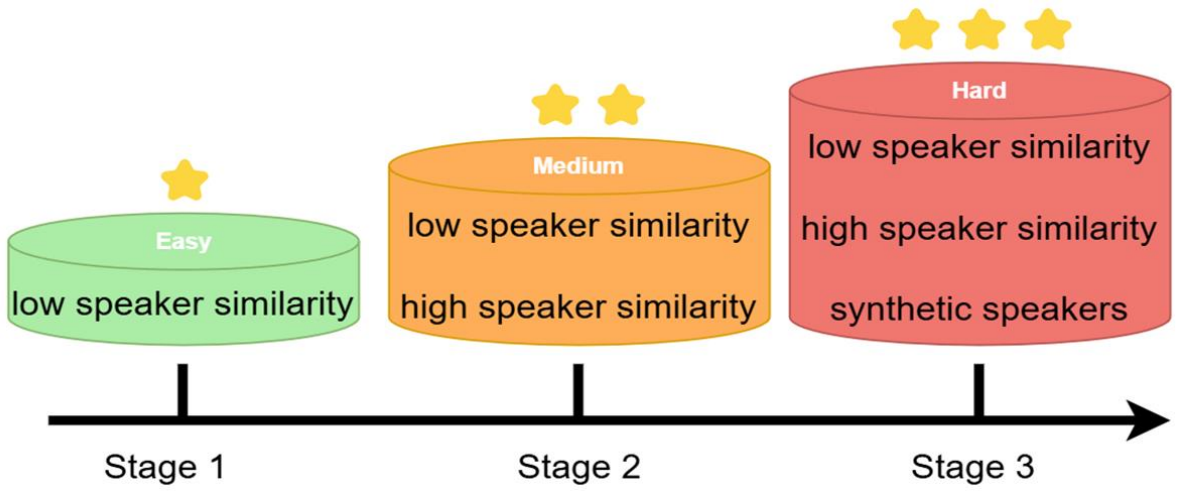
VC keeps the original linguistic content and timing of real recordings,
while only changing the speaker identity → more realistic interferers for TSE.

- VC does not require text; we can directly convert
large amounts of existing speech data.

- VC naturally preserves background noise, and channel effects
in the source mixtures, which are hard to recreate with TTS.

- Compared to TTS, VC makes it easier to control "how similar" the
synthetic speaker is to a given real speaker (by adjusting latent interpolation).

4 Curriculum Learning with Synthetic Speakers



Three-stage curriculum learning framework with synthetic interference speakers.

4 Curriculum Learning with Synthetic Speakers

Experiments setting

Key difference: interference speakers

•In Chapter 3: interferers are **real speakers** from Libri2talker.

•In Chapter 4: we **replace the interference speaker** with a **synthetic interference speaker** generated by voice conversion.

•Pre-processing, feature extraction, model architectures, and optimization follow the same setup

TSE method	Stage2	Stage3	iSDR [dB]
Conformer w/o CL (Chapter 3)	×	×	12.57
Conformer (Chapter 3)	✓	×	13.44
Conformer (real only)	✓	✓	13.98
Conformer (real+syn)	✓	✓	**14.43**

Experimental results before and after adding Stage 3, which uses synthetic speakers

5. Libri2Vox: More Diverse Speaker Conditions

5 Libri2Vox: More Diverse Speaker Conditions

Limitation:

- CL + synthetic speakers clearly improve TSE, but all experiments still use **artificially constructed mixtures**.
- Training data lacks **real "in-the-wild" acoustic conditions**, so the true robustness of our methods in real deployments is still unknown.

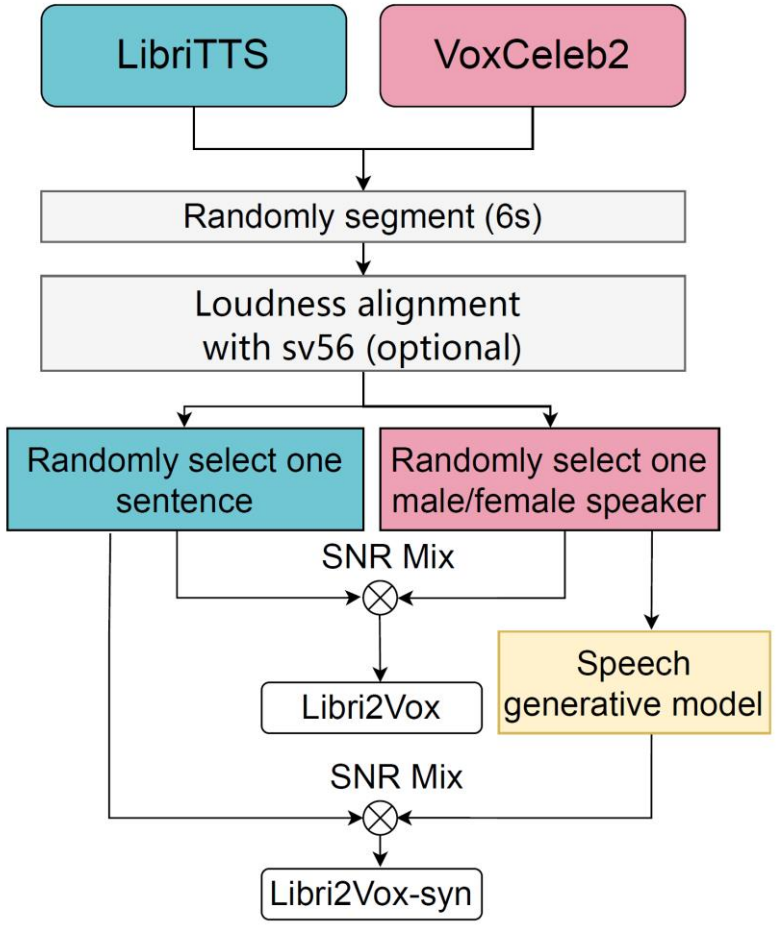
- **Motivation:**

  We need data that keeps clean target speech for supervision, but also exposes the model to real in-the-wild acoustic conditions.

Libri2Vox: LibriTTS clean target speaker + VoxCeleb2 in-the-wild interferer speaker

controlled experiments and real-world acoustic complexity.

5 Libri2Vox : More Diverse Speaker Conditions



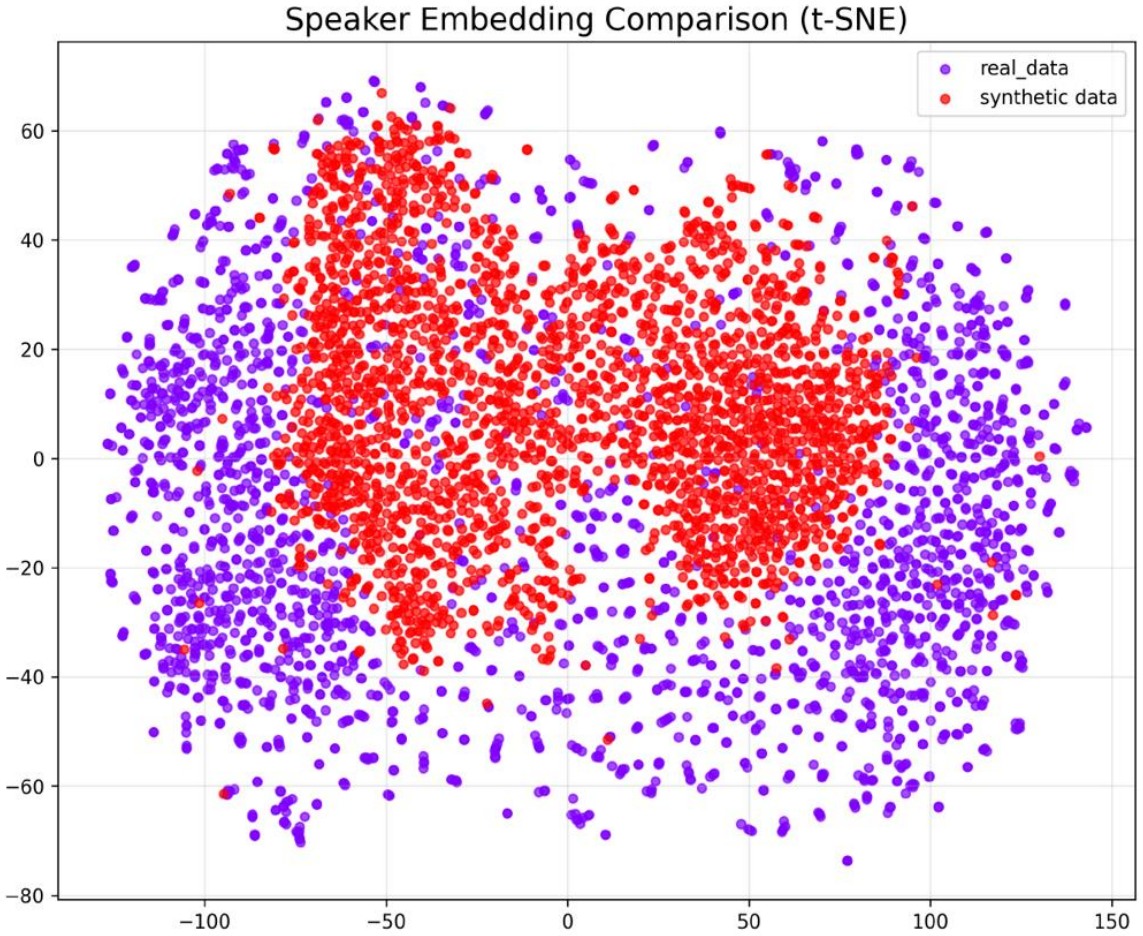
SALT[7]/Synvox2 [8]

Data generation framework of Libri2Vox and its synthetic version.

[7] Y. Lv, J. Yao, P. Chen, H. Zhou, H. Lu, and L. Xie, “SALT: Distinguish-able speaker anonymization through latent space transformation,” in2023 IEEE Automatic Speech Recognition and Understanding (ASRU),2023.
[8] X. Miao, X. Wang, E. Cooper, J. Yamagishi, N. Evans, M. Todisco,J.-F. Bonastre, and M. Rouvier, “Synvox2: Towards a privacy-friendly VoxCeleb2 dataset,” in ICASSP 2024

5 Libri2Vox : More Diverse Speaker Conditions

Why synthetic speakers?



Speaker embedding space visualization.

5 Libri2Vox : More Diverse Speaker Conditions

Experiments setting

Dataset	# Speakers	# Utterances	Duration(h)
WSJ0-2mix-extr	101 (train) 101 (val) 18 (test)	20,000 (train) 5,000 (val) 3,000 (test)	30 8 5
Libri2mix	921 (train-360) 251 (train-100) 40 (val) 40 (test)	50,800 (train-360) 13,900 (train-100) 3,000 val 3,000 test	212 58 11 11
Libri2talker	1,172 (train/val) 1,172 (val) 40 (test)	127,056 (train) 2,344 (val) 6,000 (test)	460 8 22

Comparison of WSJ0-2mix-extr, Libri2mix, and Libri2talker Datasets

Set	#. Speakers		#. Utterances	Duration (h)
	LibriTTS	VoxCeleb2	Total	
Training	1,151	5,900	149,691	250
Validation	40	94	7,200	8.97
Testing	39	118	6,000	8.56

Libri2Vox/Libri2Vox-syn statistics

The other settings, including the model, loss, and optimizer, are the same as in the previous experiments.

5 Libri2Vox : More Diverse Speaker Conditions

- PESQ: perceptual speech quality score (higher = better quality)
- STOI: short-time objective intelligibility (higher = better intelligibility)
- DNSMOS: neural-network based MOS predictor from DNS Challenge (higher = better perceived quality)

Model	Method	Stage 1	Stage 2	Stage 3	iSDR ↑	iPESQ ↑	iSTOI ↑	iDNSMOS ↑
Conformer	w/o CL (Real only)	✓			15.42 ± 0.076	1.325 ± 0.010	0.198 ± 0.002	0.650 ± 0.012
	w/o CL (Real+SynVox2)	✓			15.46 ± 0.073	1.295 ± 0.009	0.196 ± 0.002	0.640 ± 0.012
	w/o CL (Real+SALT)	✓			15.43 ± 0.069	1.287 ± 0.008	0.194 ± 0.002	0.635 ± 0.011
	w/ 1-stage CL (Real only)	✓			15.39 ± 0.075	1.401 ± 0.010	0.203 ± 0.002	0.664 ± 0.012
	w/ 2-stage CL (Real only)	✓	✓		15.89 ± 0.074	1.424 ± 0.010	0.205 ± 0.002	0.664 ± 0.011
	w/ 3-stage CL (Real only)	✓	✓	✓	16.01 ± 0.072	1.420 ± 0.010	0.205 ± 0.002	0.664 ± 0.011
	w/ 3-stage CL (Real + SynVox2)	✓	✓	✓	16.20 ± 0.072	1.470 ± 0.010	0.206 ± 0.002	0.692 ± 0.012
	w/ 3-stage CL (Real + SALT)	✓	✓	✓	16.20 ± 0.072	1.466 ± 0.010	0.207 ± 0.002	0.683 ± 0.012
BLSTM	w/o CL (Real only)	✓			12.34 ± 0.085	0.852 ± 0.008	0.157 ± 0.002	0.495 ± 0.012
	w/o CL (Real+SynVox2)	✓			12.53 ± 0.082	0.903 ± 0.009	0.164 ± 0.002	0.501 ± 0.012
	w/o CL (Real+SALT)	✓			12.54 ± 0.080	0.885 ± 0.009	0.161 ± 0.002	0.491 ± 0.011
	w/ 1-stage CL (Real only)	✓			11.73 ± 0.089	0.794 ± 0.008	0.149 ± 0.002	0.473 ± 0.012
	w/ 2-stage CL (Real only)	✓	✓		12.50 ± 0.084	0.922 ± 0.009	0.162 ± 0.002	0.520 ± 0.012
	w/ 3-stage CL (Real only)	✓	✓	✓	12.65 ± 0.083	0.932 ± 0.009	0.165 ± 0.002	0.531 ± 0.012
	w/ 3-stage CL (Real + SynVox2)	✓	✓	✓	13.07 ± 0.082	0.997 ± 0.010	0.171 ± 0.002	0.554 ± 0.012
	w/ 3-stage CL (Real + SALT)	✓	✓	✓	13.06 ± 0.082	0.986 ± 0.010	0.170 ± 0.002	0.544 ± 0.012

Results of three stage methods for Libri2Vox test set.

5 Libri2Vox : More Diverse Speaker Conditions

- WER: word error rate (lower value = better performance)

Architecture	w/o CL (Real only)	w/ 3-stage CL (Real + SynVox2)
Conformer	28.8	28.3
BLSTM	31.6	30.5
SpeakerBeam	41.4	37.1
VoiceFilter	30.8	30.1
Mixture (no TSE)	72.5	

Downstream ASR performance (faster-whisper large-v3)

5 Libri2Vox : More Diverse Speaker Conditions

	Training Data	iDNSMOS
Conformer	Libri2Talker	-0.722
	Libri2Talker+Noise	-0.534
	Libri2Vox	-0.047
	Libri2Vox+Noise	0.017
	Libri2Vox & 3-stage (Real + SynVox2)	**0.186**
BLSTM	Libri2Talker	-0.666
	Libri2Talker+Noise	-0.421
	Libri2Vox	-0.081
	Libri2Vox+Noise	0.139
	Libri2Vox & 3-stage (Real + SynVox2)	**0.190**
SpeakerBeam	Libri2Talker	-0.701
	Libri2Talker+Noise	-0.542
	Libri2Vox	-0.289
	Libri2Vox+Noise	-0.009
	Libri2Vox & 3-stage (Real + SynVox2)	**0.031**
VoiceFilter	Libri2Talker	-0.739
	Libri2Talker+Noise	-0.369
	Libri2Vox	-0.309
	Libri2Vox+Noise	-0.140
	Libri2Vox & 3-stage (Real + SynVox2)	**0.064**

DNSMOS results on **real-world recordings** from ICASSP 2021 DNS Challenge test set.

6. Multi-factor Curriculum Learning with Training Dynamic

6 Multi-factor Curriculum Learning with Training Dynamic

RQ4 **How can we coordinate multiple complexity factors in challenging multi-speaker scenarios?**

Present a training dynamics-aware multi-factor CL framework that moves beyond the single-factor approaches explored in earlier chapters to coordinate multiple complexity dimensions simultaneously.

Motivation for RQ4

- Previous chapters mainly handle one difficulty factor at a time. But real TSE has multiple factors interacting together.
- Existing CL uses handcrafted difficulty rules, which may not match how the model actually learns during training.

Design a training-dynamics-aware multi-factor CL that coordinates several complexity dimensions based on the model's learning behavior.

6 Multi-factor Curriculum Learning with Training Dynamic

Multiple factors Curriculum Learning:

- Number of interfering speakers
- SNR range of the mixture
- Temporal overlap ratio
- Real vs synthetic interference speakers

a **multi-factor curriculum** that can coordinate all these dimensions together.

Experiments (Libri2Vox multi-speaker mixtures)

- We use the **Libri2Vox** dataset as the base for all mixtures in this chapter.
- Compared with previous chapters (mainly 2-speaker mixtures), here we create a challenge **test set**:
  - **2-speaker** mixtures: target + 1 interferer,
  - **3-speaker** mixtures: target + 2 interferers,
  - **4-speaker** mixtures: target + 3 interferers.

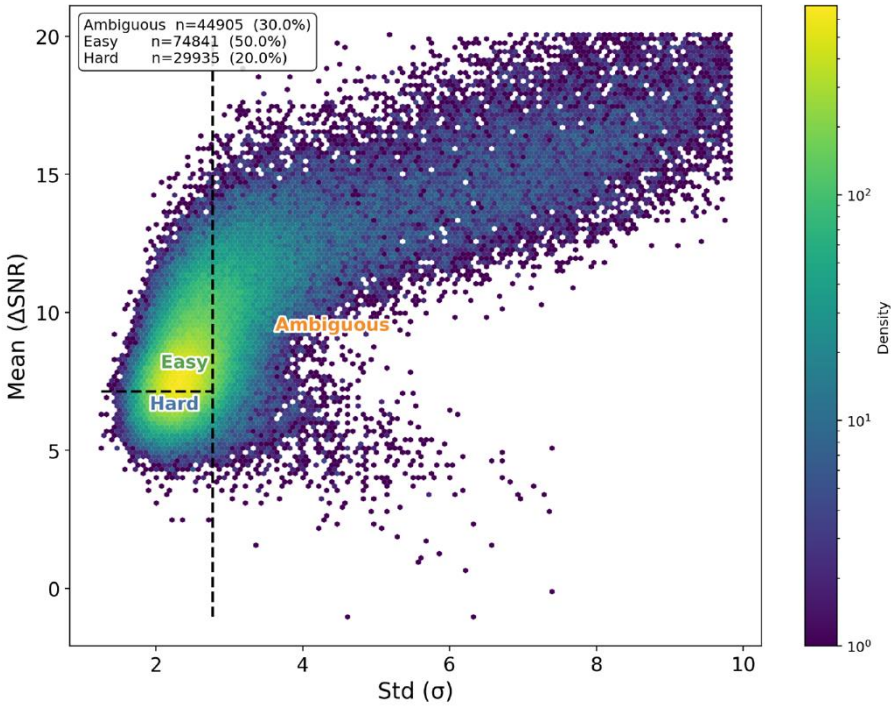
	# speakers	SNR range	Overlap ratio	Type of speakers
Threshold	{1, 2, 3}	([−5,0,5] dB	{0.2, 0.4, 0.6}	{real,syn,real+syn}

6 Multi-factor Curriculum Learning with Training Dynamic

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**

Impact of single-factor and multi-factor CL.

TSE-Datamap: Seeing Training Dynamics in 2D



TSE-Datamap

x-axis: std of improved SNR
y-axis: mean improved SNR

$$M_i(s_{0,i}, \hat{s}_{0,i}) = -10 \log_{10} \frac{\|s_{0,i}\|^2}{\|s_{0,i} - \hat{s}_{0,i}\|^2}$$

- Track **ΔSNR for each sample over 50 epochs**
- Map training dynamics to 2D space

Regions

- **Easy**: high mean, low std → consistently well learned
- **Ambiguous**: high std → predictions change a lot, helpful for **decision boundary shaping**
- **Hard**: low mean, low std → **consistently bad**, extremely low SNR cases

6 Multi-factor Curriculum Learning with Training Dynamic

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

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

Performance comparison of **different stage** orderings in CL. Each stage takes one third of total epochs.

6 Multi-factor Curriculum Learning with Training Dynamic

Setting	2spk	3spk	4spk
ambi 70%	**11.67**	**8.67**	**8.61**
baseline 70%	11.37	8.19	7.17
hard 70%	10.10	6.75	6.99
easy 70%	9.71	6.50	6.40

Fixed data quantity drawn from Datamap regions

7. Conclusion and Future Work

7 Conclusion and Future Work

Contributions

- **Systematic view of training data difficulty for TSE**

– Analyze errors and difficulty factors in TSE mixtures.

– Show that *speaker similarity* is a strong and stable difficulty measure.

- **Curriculum learning strategies for TSE**

– Design two-stage and self-paced curricula that move from easy to hard mixtures.

– Achieve consistent gains over random sampling across several TSE models.

- **Using synthetic speakers as training data**

– Build generative speech pipelines to generate interference speakers.

– Show that mixing real and synthetic speakers (around 50%–50%) gives better robustness than real-only or synthetic-only training.

8 Conclusion and Future Work

Can TSE generalize to emotional speech?

Potential Impact

• Strong emotions or expressive styles *(e.g., shouting, crying, laughter, dramatic prosody)* may cause **style-induced mismatch**.

• This mismatch can **degrade speaker representations** and **extraction performance**.

Scope & Future Work

• Emotional and expressive speech is **outside the scope** of this thesis.

• Future work:

- Benchmark on **emotional speech corpora**
- Explore **style-robust speaker encoders**
- Design **curricula incorporating speaking-style variation**

8 Conclusion and Future Work

Knowledge-driven vs. Data-driven Curriculum Design

Knowledge-driven curriculum

- Difficulty defined by **predefined domain factors**
- e.g., SDR, speaker similarity, mixture conditions
- ✓ Interpretable, low cost
- ✕ May not reflect **model-specific weaknesses**

Data-driven curriculum

Difficulty inferred from **model feedback**

e.g., SNR-based model difference, training loss (self-paced)

✓ Adaptive to the current model

✕ Higher computational cost

Beyond Loss-based Signals (Future Direction)

•Use **automatic speech quality predictors** as difficulty signals e.g., DNSMOS, UTMOS

•Enable:

- •filtering low-quality synthetic samples
- •quality-aware easy → hard curricula
- •Especially useful for **large-scale, in-the-wild, or synthetic data**

Thank you!