Adaptive Hard-Pair Sampling via Curriculum Learningfor Speech Separation

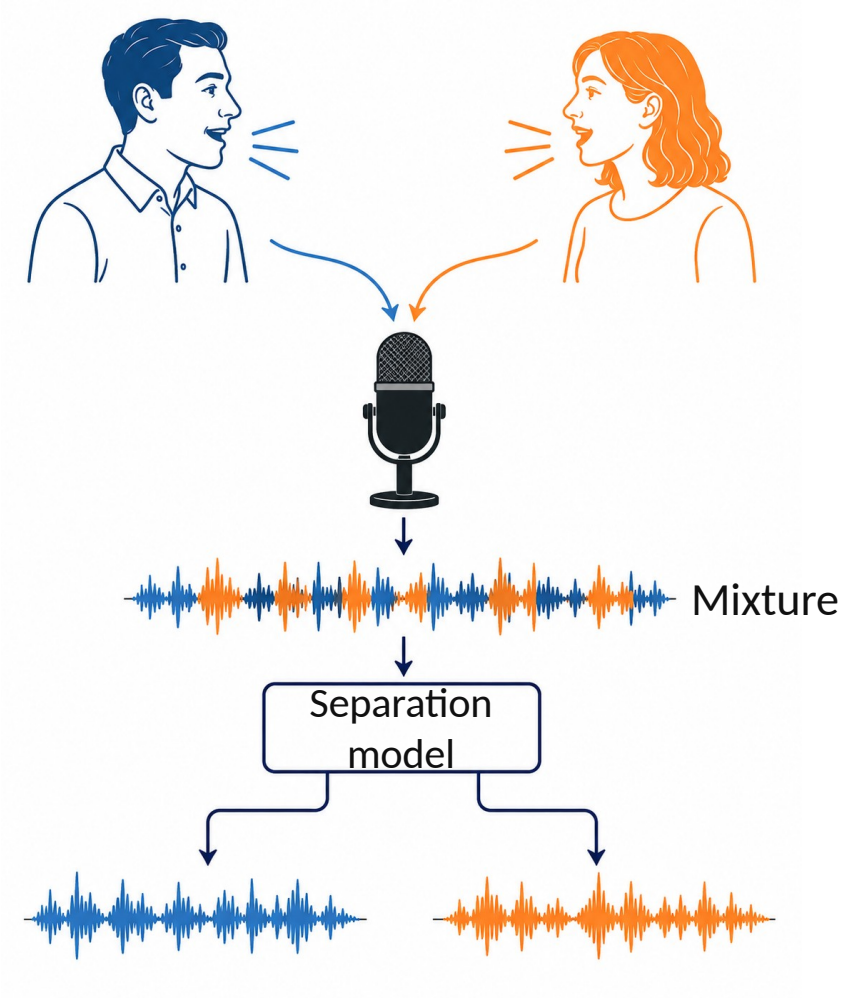
Pengjie Shen[1,2], Xueliang Zhang[1], and Zhong-Qiu Wang[2]

[1] Inner Mongolia University, China

[2] Southern University of Science and Technology, China

shenpengjie@mail.imu.edu.cn

Speech Separation



- Separate two speakers from a single-channel mixture

$$x(t) = s_1(t) + s_2(t)$$

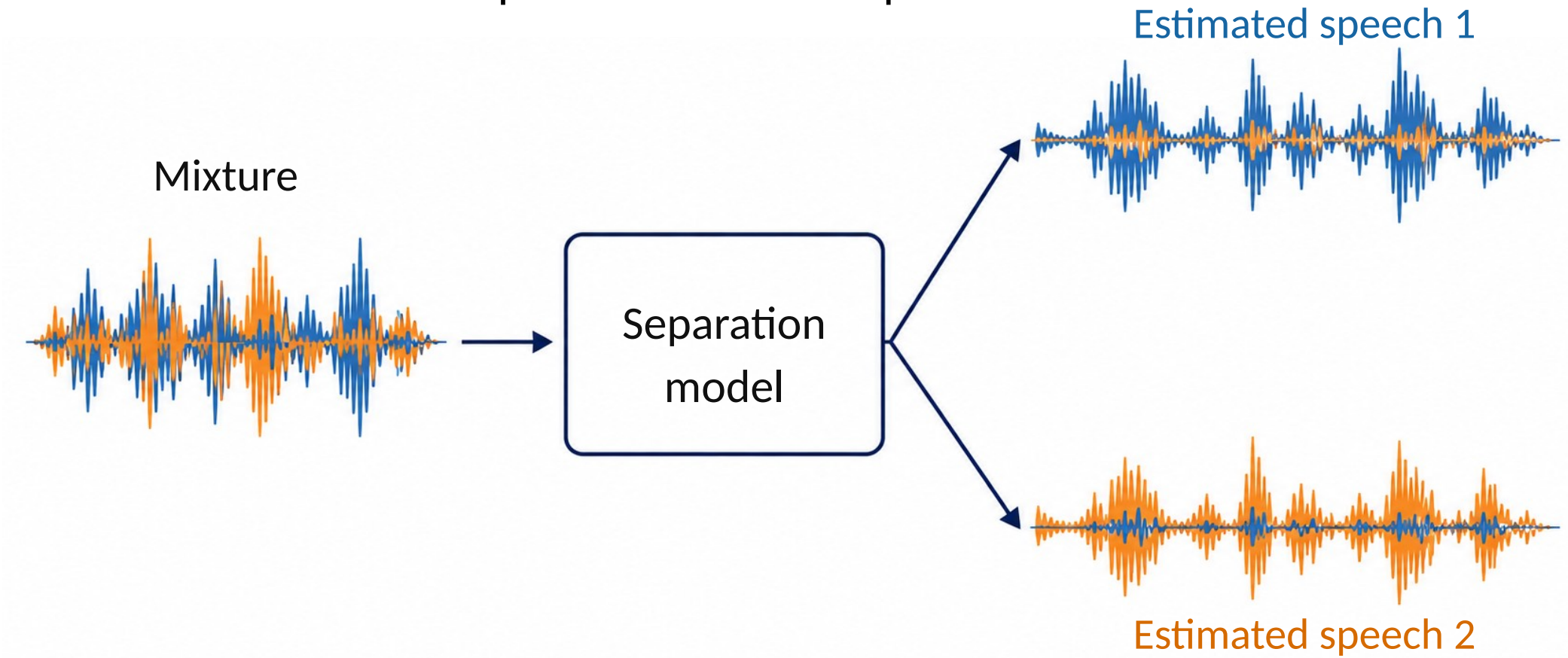
x: recorded mixture

s_1, s_2: clean speech signals

t: time

Difficult Speaker Pairs

- Speakers with similar voices can be difficult to separate
  - We call these difficult speaker pairs **hard pairs**
- Residual interference:speech from other speaker remains



Uniform Pair Sampling

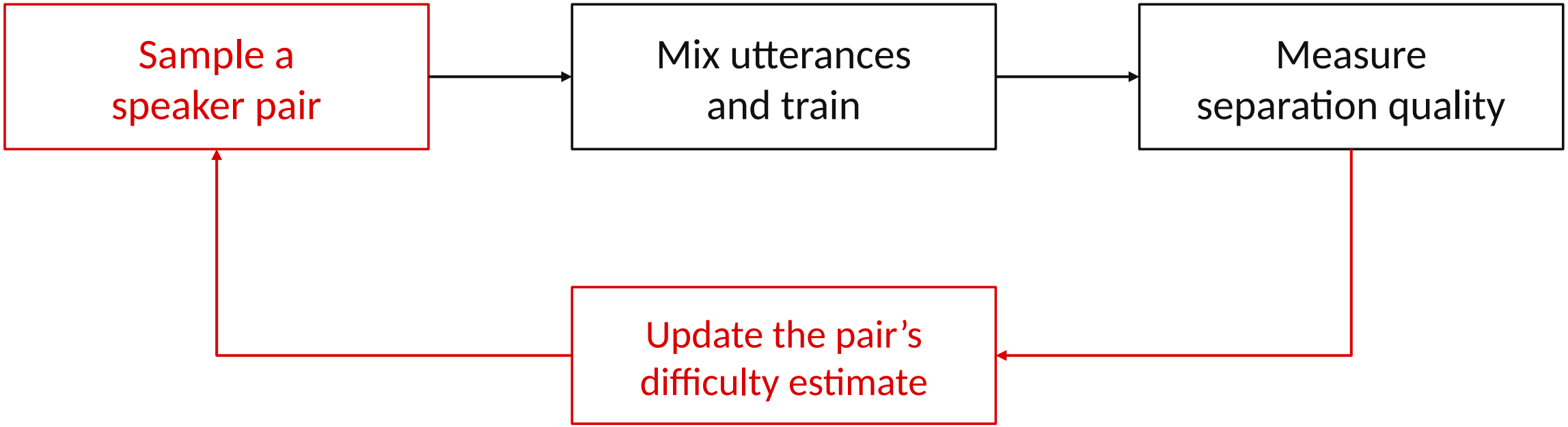
- Dynamic mixing [1] creates training mixtures from individual utterances
- Each speaker pair has same sampling probability
- Hard pairs make up only a small fraction of all speaker pairs



[1] N. Zeghidour and D. Grangier, “Wavesplit: End-to-End Speech Separation by Speaker Clustering,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 29, pp. 2840–2849, 2021

Adaptive Hard-Pair Sampling

- Measure separation quality **on the fly**
- Updated estimates determine the next sampling probabilities



Estimating Pair Difficulty

- SI-SDR compares each output with its clean speech signal
- Mixture difficulty= −SI-SDR
- $H(A, j)$: difficulty of speaker pair (A, j)
  - Updated from the scores of training mixtures containing the same two speakers

Pair (A, E) has the highest difficulty estimate

Partner j	**B**				**C**				**D**				**E**							
Partner j	B	C	D	E	Partner j	B	C	D	E	Partner j	B	C	D	E	Partner j	B	C	D	E	Partner j	B	C	D	E
H(A, j)	-20	-15	-10	-5	H(A, j)	-20	-15	-10	-5	H(A, j)	-20	-15	-10	-5	H(A, j)	-20	-15	-10	-5	H(A, j)	-20	-15	-10	-5

Selecting the Second Speaker

- Select first speaker uniformly (A in this example)
- Select the second speaker j from speakers previously paired with A
- Higher pair difficulty gives a higher sampling probability

$$\rho(j \mid A) = \frac{\exp(H(A, j)/\tau)}{\sum_k \exp(H(A, j)/\tau)}$$

Probability of selecting speaker j given A

Temperature

Partner j **B** **C** **D** **E**

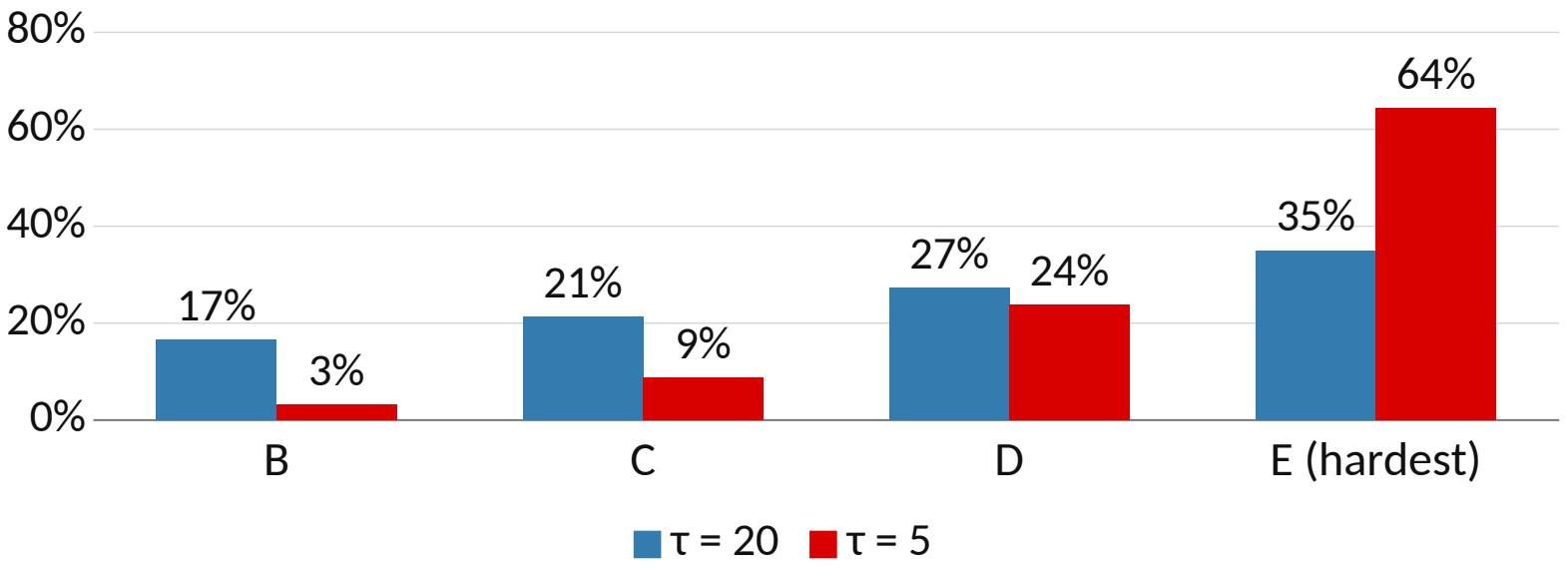
Partner j	B	C	D	E
H(A, j)	-20	-15	-10	-5
Probability (τ = 20)	17%	21%	27%	35%

Partner j	B	C	D	E
H(A, j)	-20	-15	-10	-5
Probability (τ = 20)	17%	21%	27%	35%

Temperature and Sampling Probabilities

Partner j

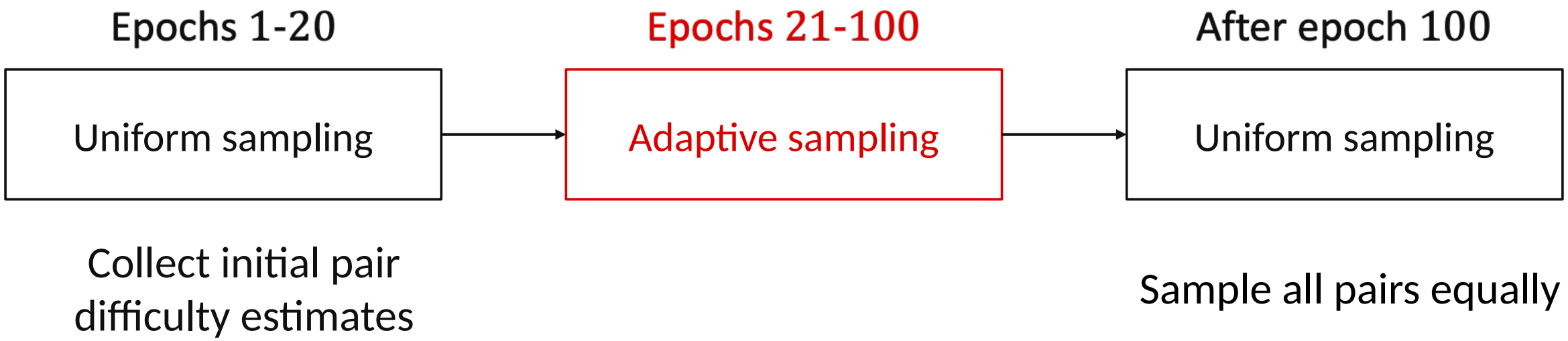
Partner j	B	C	D	E
H(A, j)	-20	-15	-10	-5



- $\tau = 20$
  - Smaller differences between probabilities
- $\tau = 5$
  - Higher probability for hardest pair

Three Training Stages

- Dynamic mixing generates new mixtures throughout training.



Difficulty Updates and Temperature Schedule

- Update speaker-pair difficulty
  - For first mixture from pair (A, j): $H(A, j) = -\text{SI-SDR}$
  - For later mixtures from the same pair: $H(A, j) \leftarrow (1 - \alpha)H(A, j) + \alpha(-\text{SI} -$
- Gradually reduce temperature during adaptive sampling

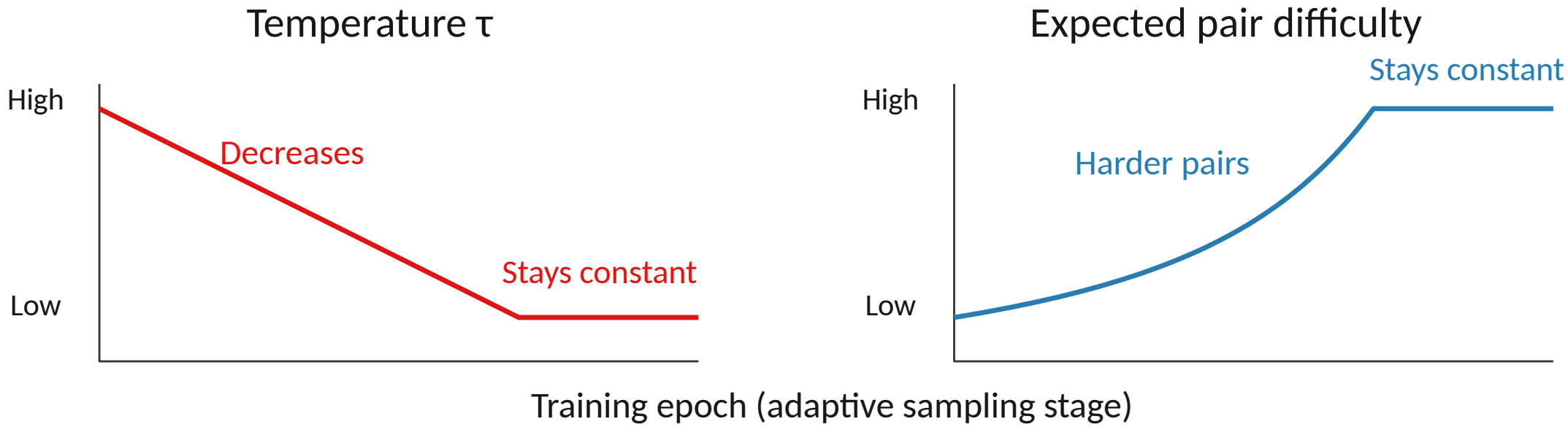


Illustration with fixed pair difficulty scores

Experimental Setup

- **Libri2Mix**
  - 1172 training speakers and 6000 test mixtures
    Two-speaker mixtures at 16 kHz; relative speech level: -5 to +5 dB
- **Three separation models**
  - Conv-TasNet, BSRNN and TF-GridNet
- **Uniform sampling versus proposed sampling**
  - Same separation model, training loss and training settings

Both sampling strategies use dynamic mixing

Average Separation Performance

- Mean SI-SDRi is unchanged or higher for all three models

Model	Uniform sampling	Proposed	Change
Conv-TasNet	16.7	16.7	+0.0
BSRNN	21.0	21.3	+0.3
TF-GridNet	21.9	**22.7**	**+0.8**

Model	Uniform sampling	Proposed	Change
Conv-TasNet	16.7	16.7	+0.0
BSRNN	21.0	21.3	+0.3
TF-GridNet	21.9	**22.7**	**+0.8**

Model	Uniform sampling	Proposed	Change
Conv-TasNet	16.7	16.7	+0.0
BSRNN	21.0	21.3	+0.3
TF-GridNet	21.9	**22.7**	**+0.8**

Model	Uniform sampling	Proposed	Change
Conv-TasNet	16.7	16.7	+0.0
BSRNN	21.0	21.3	+0.3
TF-GridNet	21.9	**22.7**	**+0.8**

Results on the Lowest-Scoring Mixtures

- Bottom 30%: test mixtures with the lowest SDRi scores
- Adaptive sampling improves mean SDRi for all three models on the bottom 30%

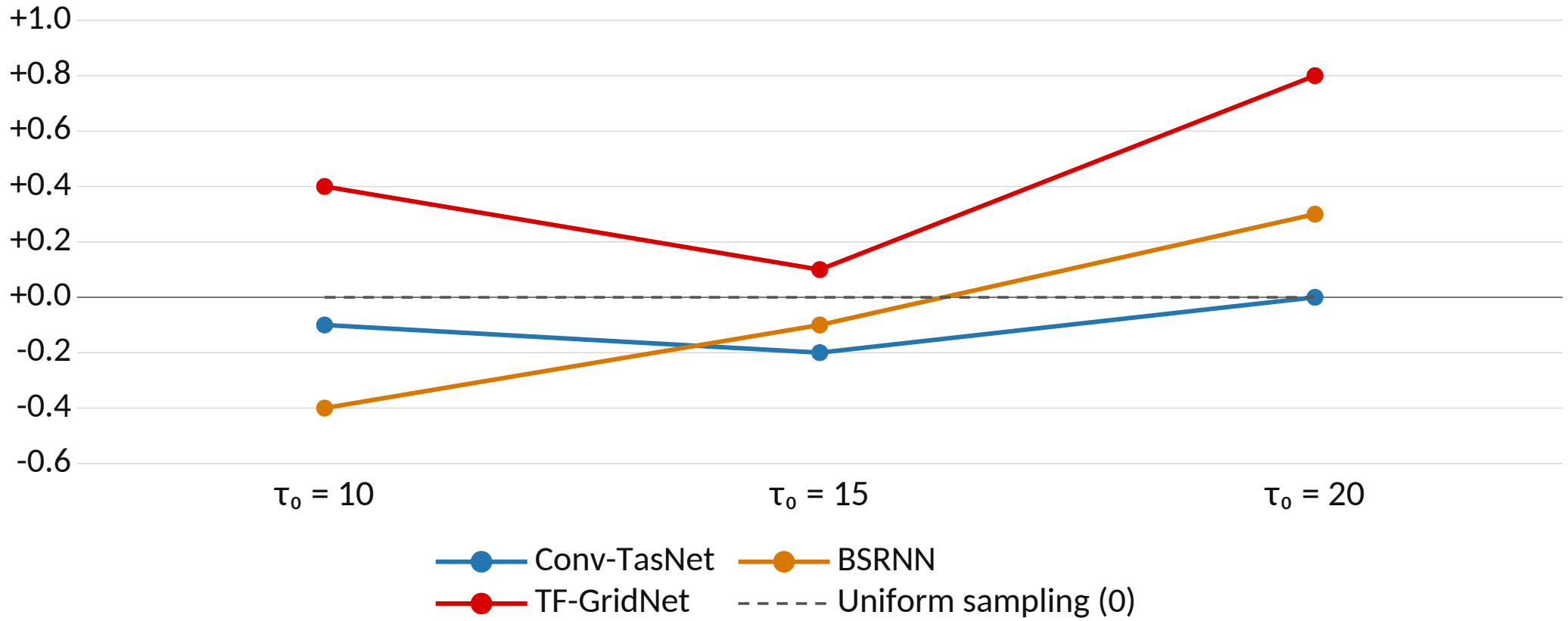
Model	Uniform sampling	Proposed	Change
Conv-TasNet	12.9	13.2	+0.3
BSRNN	17.4	17.6	+0.2
TF-GridNet	**18.3**	**18.8**	+0.5

Model	Uniform sampling	Proposed	Change
Conv-TasNet	12.9	13.2	+0.3
BSRNN	17.4	17.6	+0.2
TF-GridNet	**18.3**	**18.8**	+0.5

Model	Uniform sampling	Proposed	Change
Conv-TasNet	12.9	13.2	+0.3
BSRNN	17.4	17.6	+0.2
TF-GridNet	**18.3**	**18.8**	+0.5

Effect of Initial Temperature

- Among the three settings tested, τ_0 gives the highest mean SI-SDRi for each model



Conclusion

- Method
  - Update speaker-pair difficulty on the fly using −SI-SDR
  - Sample harder pairs more often

Keep the separation network and training loss unchanged

- Results on Libri2Mix
  - Mean SI-SDRi is maintained or improved for all three models
  - Mean SDRi on the bottom 30% improves for all three models

Thank you!

Questions?

Pengjie Shen

shenpengjie@mail.imu.edu.cn