@Conversational AI Reading Group, Feb. 13, 2025

Scalable and Efficient Speech Enhancement

Minje Kim, Ph.D.
Siebel School of Computing and Data Science
Visiting Academic at Amazon Lab126
https://minjekim.com
minje@illinois.edu

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

SIEBEL SCHOOL OF COMPUTING AND DATA SCIENCE
GRAINGER ENGINEERING

Outline

- Motivation
  - General model compression for SE
- Personalization for Model Compression
  - Knowledge distillation
  - Mixture of local experts
- Scalable and Efficient Models
  - BLOOM-Net
  - Cold diffusion
- Discussion

Sunwoo Kim

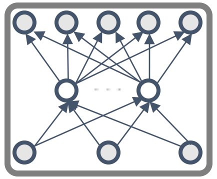
Aswin Sivaraman

Trausti Kristjansson

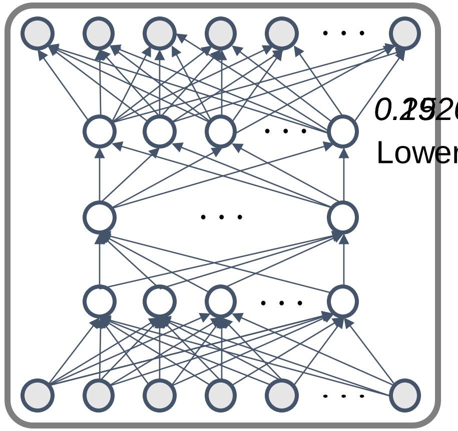
Motivation

- Speech enhancement as a benchmark task

- ○ Typical application goals of SE
  - ◻ High quality audio
    - Intelligibility, perceptual, subjectivity, etc.
  - ◻ Online/real-time processing
    - Low delay
  - ◻ On-device processing
    - Low complexity
- ○ General-purpose SE
  - ◻ Works well but is too complex
- ○ Model compression
- ○ Speaker-agnostic model compression
- ○ Scalable and efficient SE

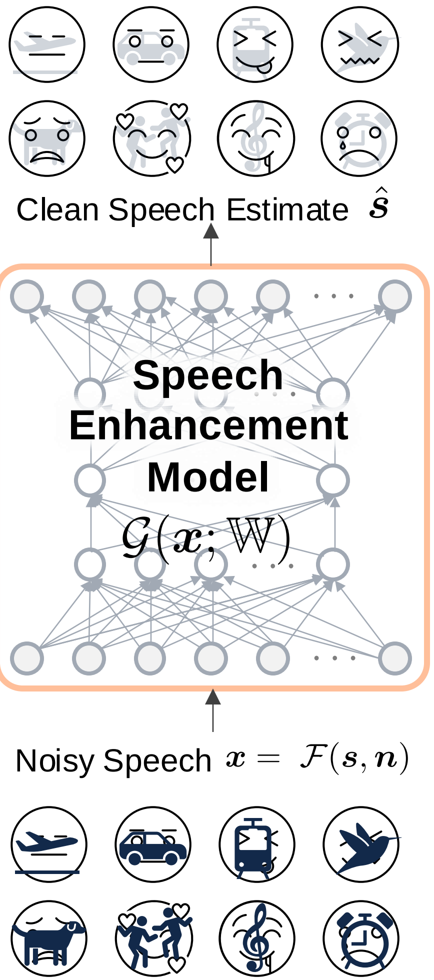


Smaller Architecture



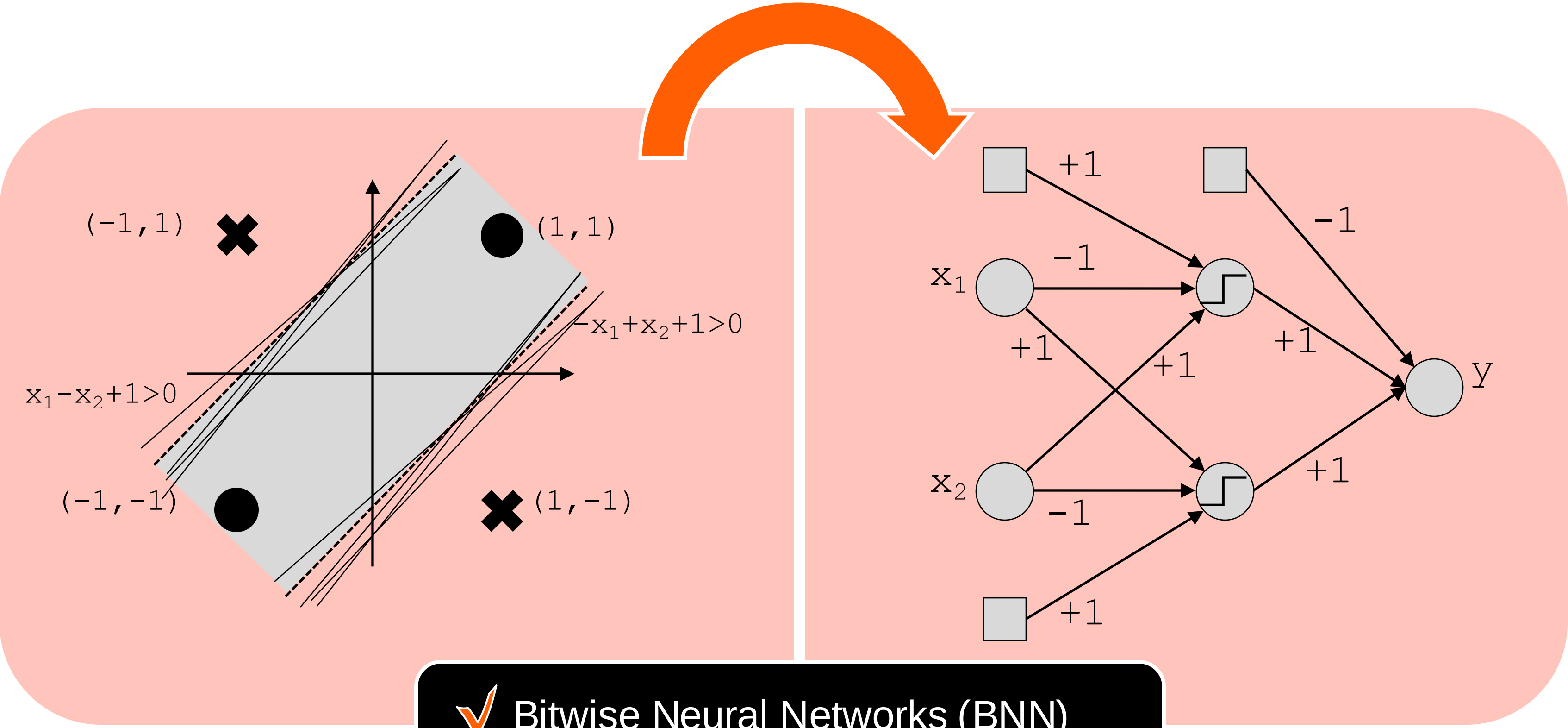
Lower-bit quantization

Pruning



Bitwise Neural Networks

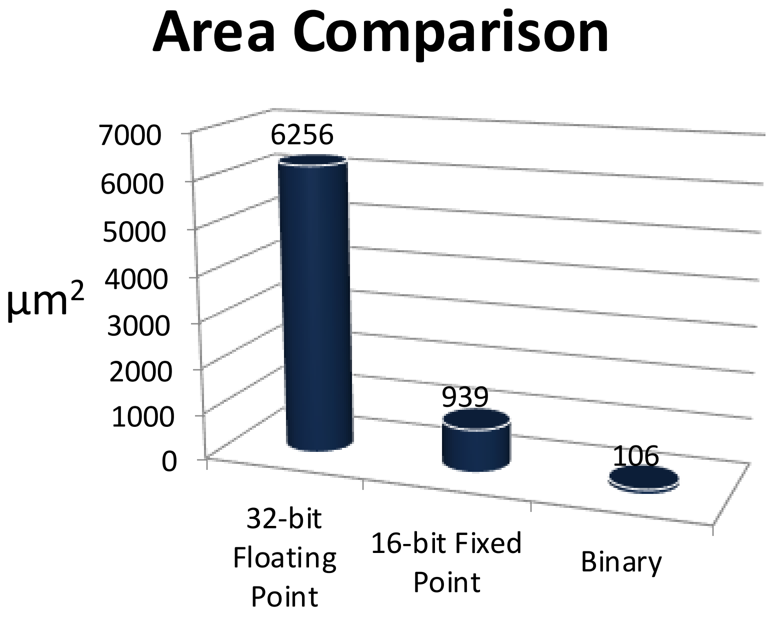
- The XOR Example

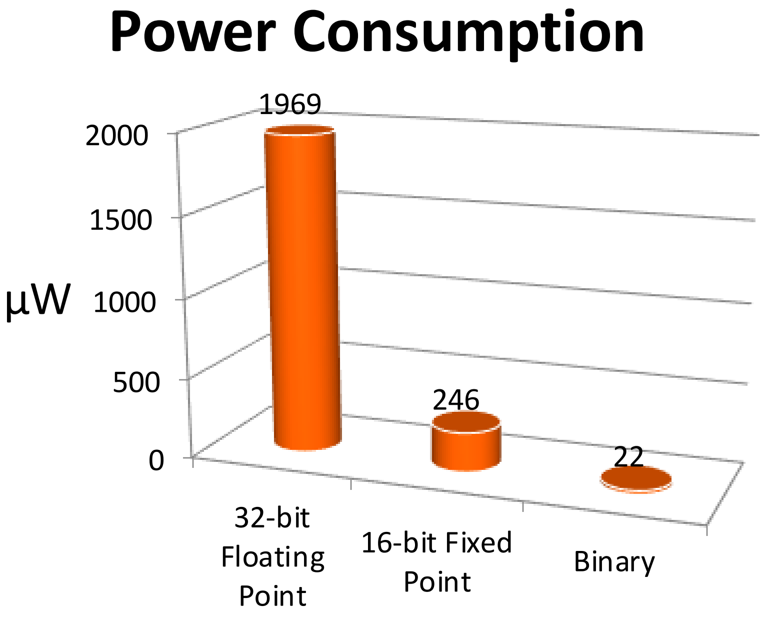


✓ Bitwise Neural Networks (BNN) CAN solve non-linear problems

Bitwise Neural Networks

- Efficiency in HW

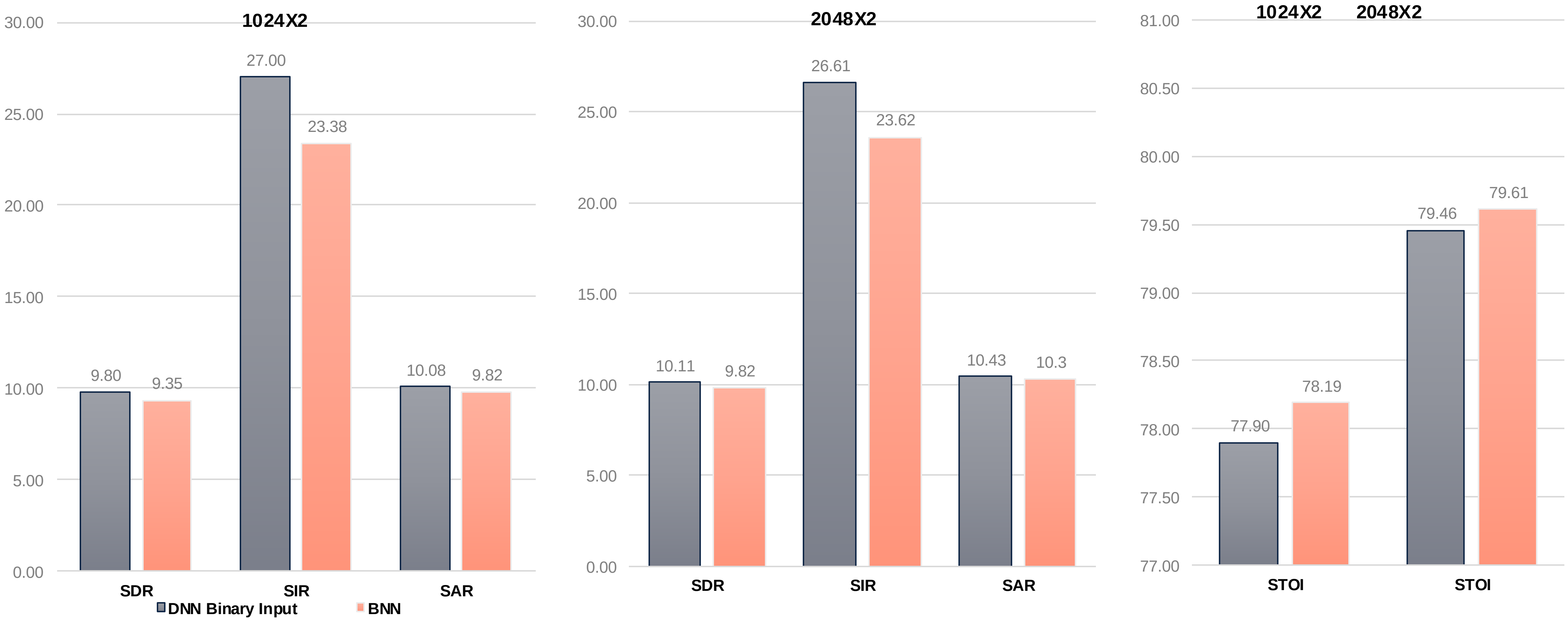




- Rough estimation without considering some constant overhead
- NanGate 45nm / DesignCompiler
- Per each node

BNN for Supervised Speech Denoising

- Compared to a single-precision network



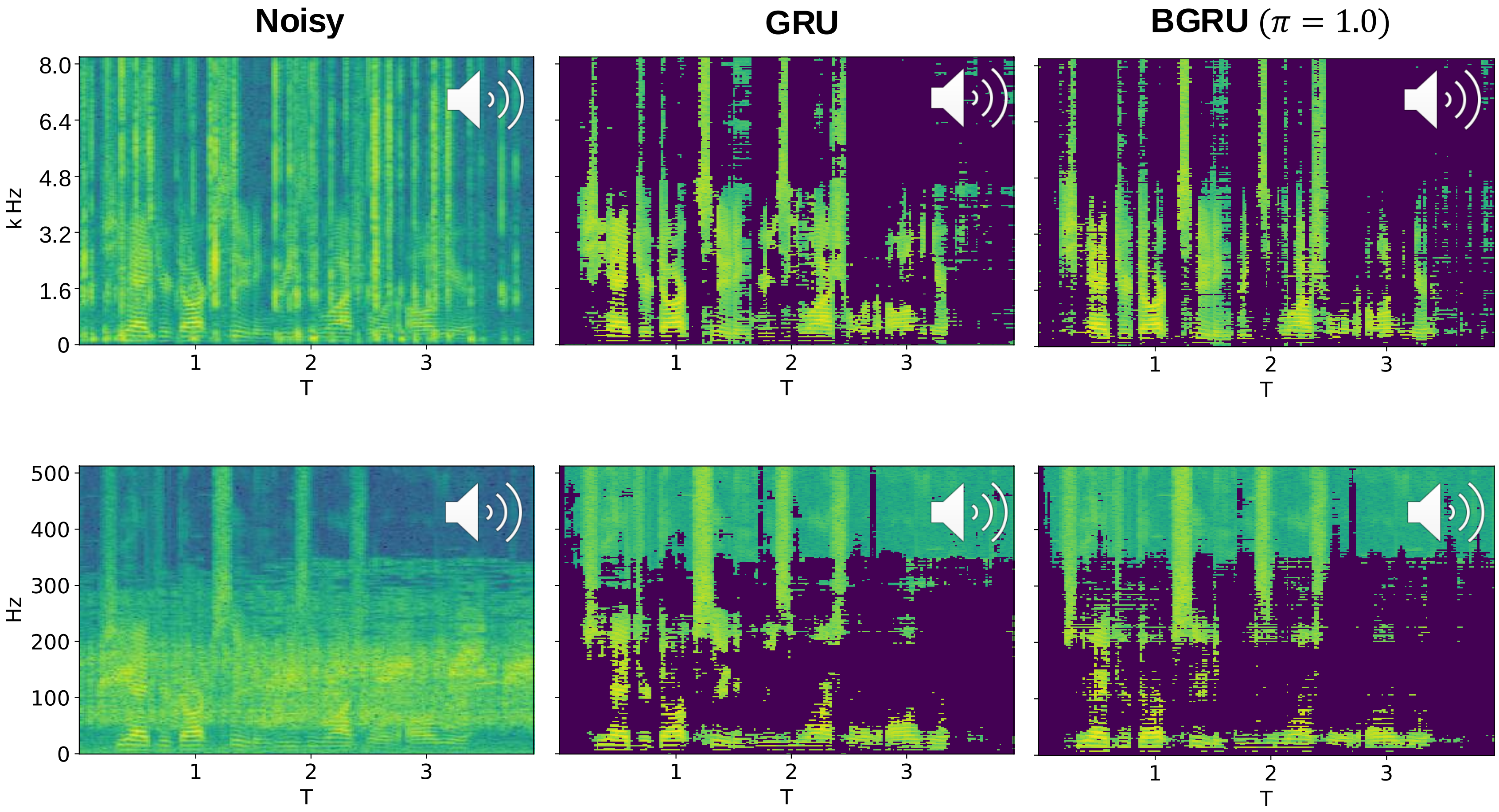
BNN for Supervised Speech Denoising

- Audio demo

	Input Noisy Speech	Deep Learning (Binary Input)	Bitwise
Female + Typing			
Female + Ocean			
Female + Frogs			
Male + Eating Chips			
Male + Jungle			

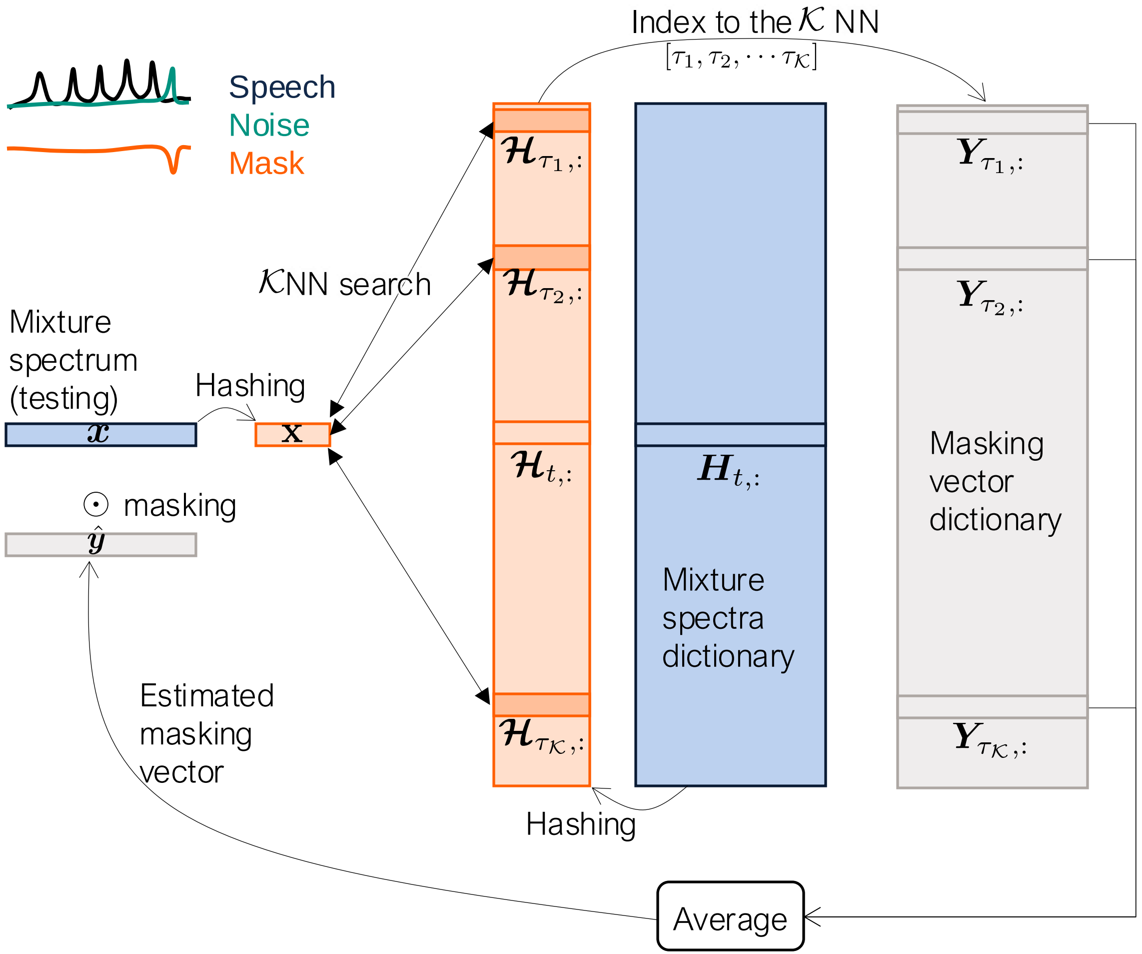
Bitwise Gated Recurrent Units

- Audio demo



Boosted Hashing for Bitwise Source Separation

- Overview

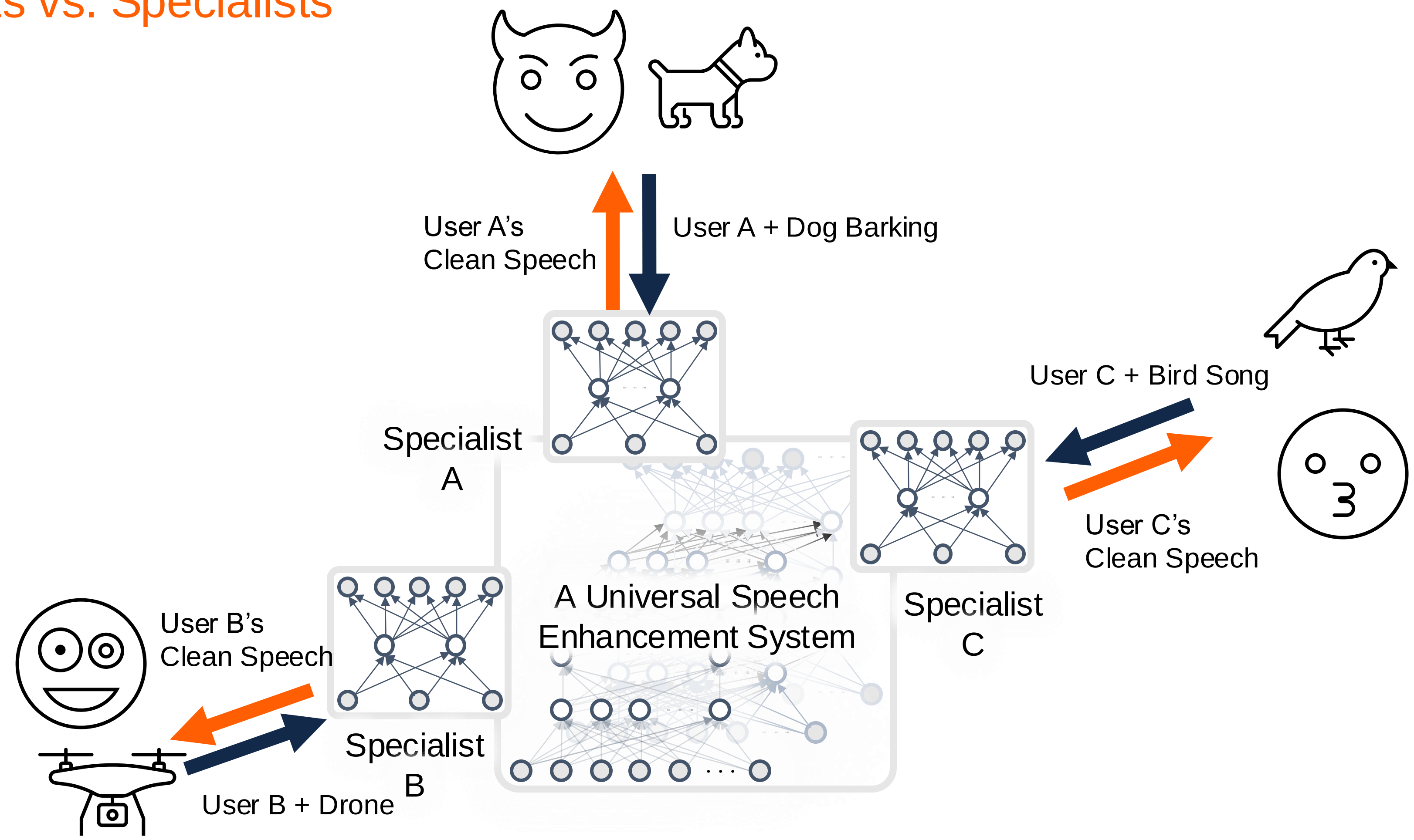


- Hashing can speed up the search
  - Search is based on Hamming similarity
- Hashing can degrade the performance
  - Hamming similarity vs. perceptual similarity
  - Needs some machine learning
- Adaboost + locality sensitive hashing

Personalized Speech Enhancement

Motivation

- Generalists vs. Specialists



M. Kolbæk, Z. H. Tan and J. Jensen, "Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems," *IEEE/ACM TASLP*, 2017.

Motivation

- Generalists vs. Specialists

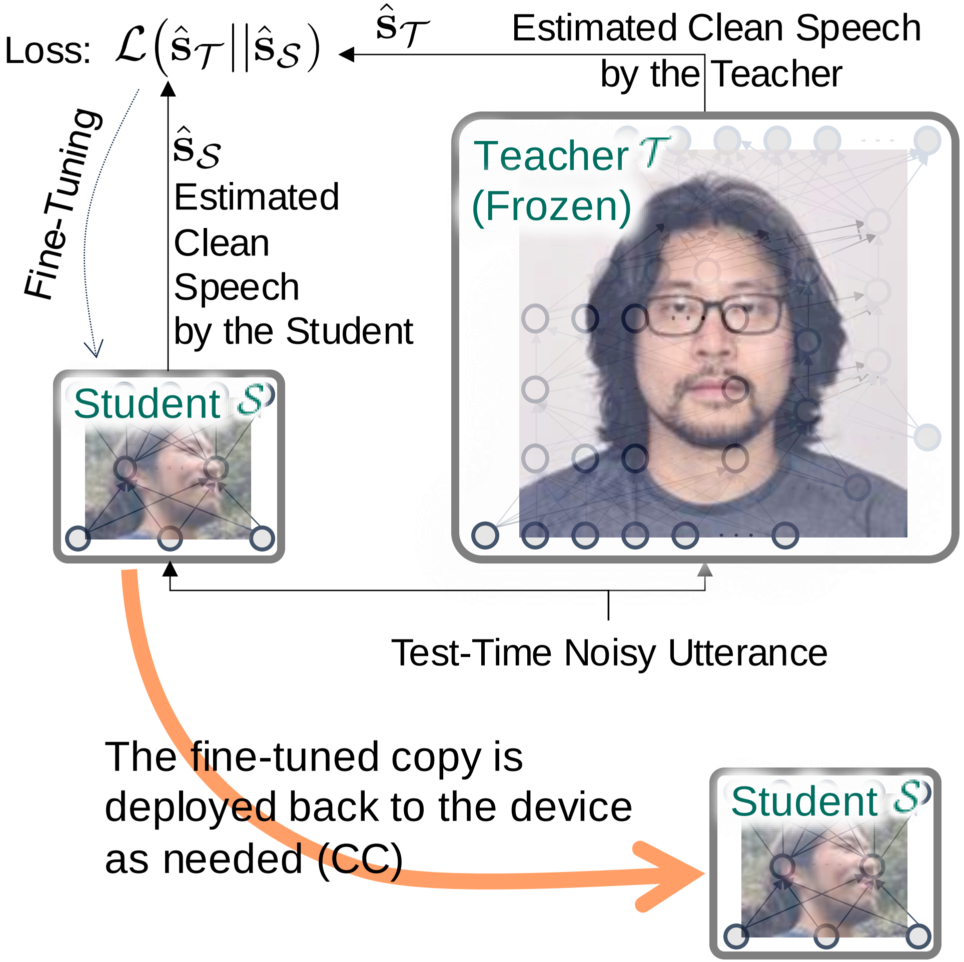
Noise Types	Mixture (Input)	Results from the Best Specialist	Results from the Worst Specialist
Bird Singing			
Typing			
Motorcycle			

- How to train a personalized SE system?
  - We don’t have access to clean personal speech

Test-Time Model Adaptation

- Knowledge distillation for PSE

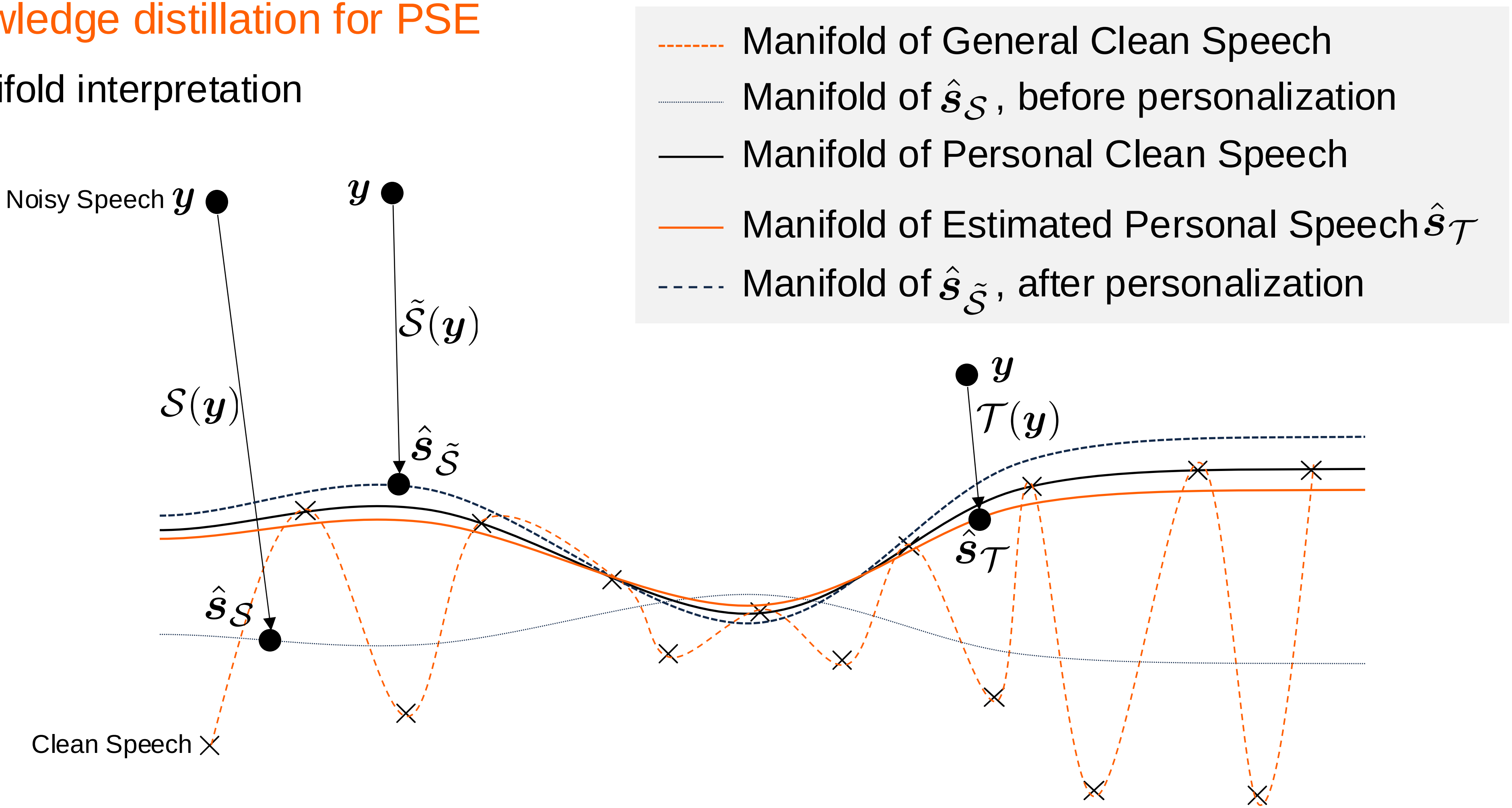
- Pre-train a large teacher model $\mathcal{T}$ for SE and freeze it
  - Generalizes well but is too big
- Pre-train a small, thus efficient student model $\mathcal{S}$
  - But can make a mistake
  - No way to fix it on its own (lack of GT clean speech)
- Test-time adaptation
  - Distill the teacher's outputs as pseudo-targets
  - Fine-tune the student
  - Assumption: teachers are better than students
    $\mathcal{L}(\mathbf{s}||\hat{\mathbf{s}}_{\mathcal{T}}) < \mathcal{L}(\mathbf{s}||\hat{\mathbf{s}}_{\mathcal{S}})$
- Use-case scenario:
  - Only the student model is used during inference on the device
  - Fine-tuning occurs either on a cloud server or on-device during idle time



Test-Time Model Adaptation

- Knowledge distillation for PSE

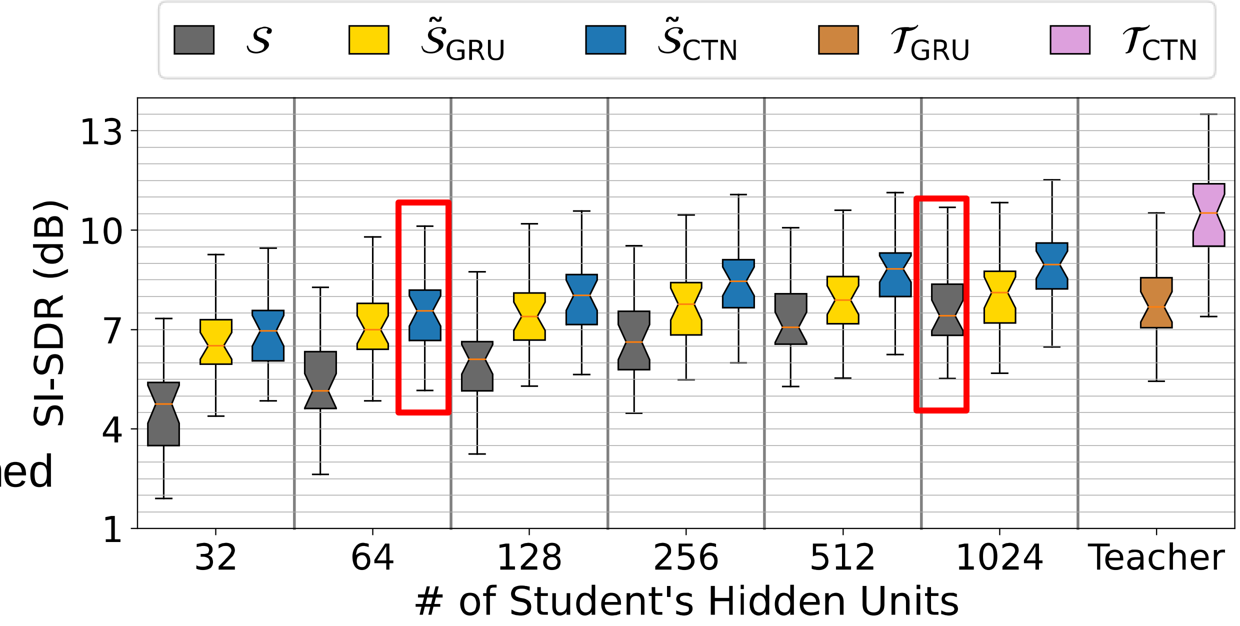
- Manifold interpretation



Test-Time Model Adaptation

- Knowledge distillation for PSE

Models		MACs (G)	Param. (M)
Student	GRU (2×32)	0.010	0.08
	GRU (2×64)	0.011	0.17
	GRU (2×128)	0.026	0.41
	GRU (2×256)	0.071	1.12
	GRU (2×512)	0.216	3.42
	GRU (2×1024)	0.729	11.55
Teacher	GRU (3×1024)	1.126	17.85
	ConvTasNet [28]	9.831	4.92



- PSE consistently outperforms all pre-trained student models
  - More improvement on smaller architectures
- $\tilde{S}_{CTN}$ always outperforms their corresponding $\tilde{S}_{GRU}$
- Lossless network compression
  - 2 x 64 $\tilde{S}_{CTN}$ vs. 2 x 1024 S
    - ~66x lower MACs and parameters

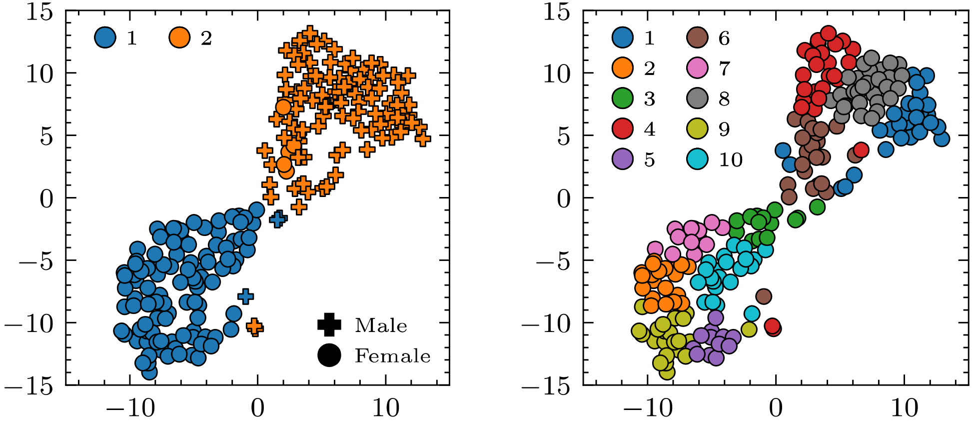
Test-Time Model Selection

- Speaker-Specific Sparse Ensemble of Specialists

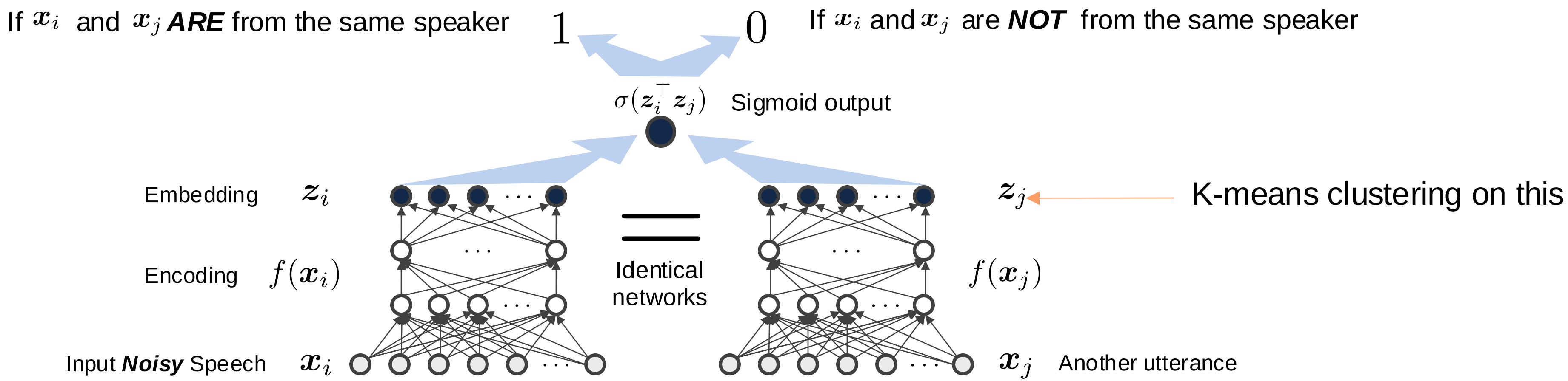
- Speaker-specific subproblem

Speech Enhancement Problem Space

Speaker Group A	Speaker Group B
Speaker Group C	Speaker Group D



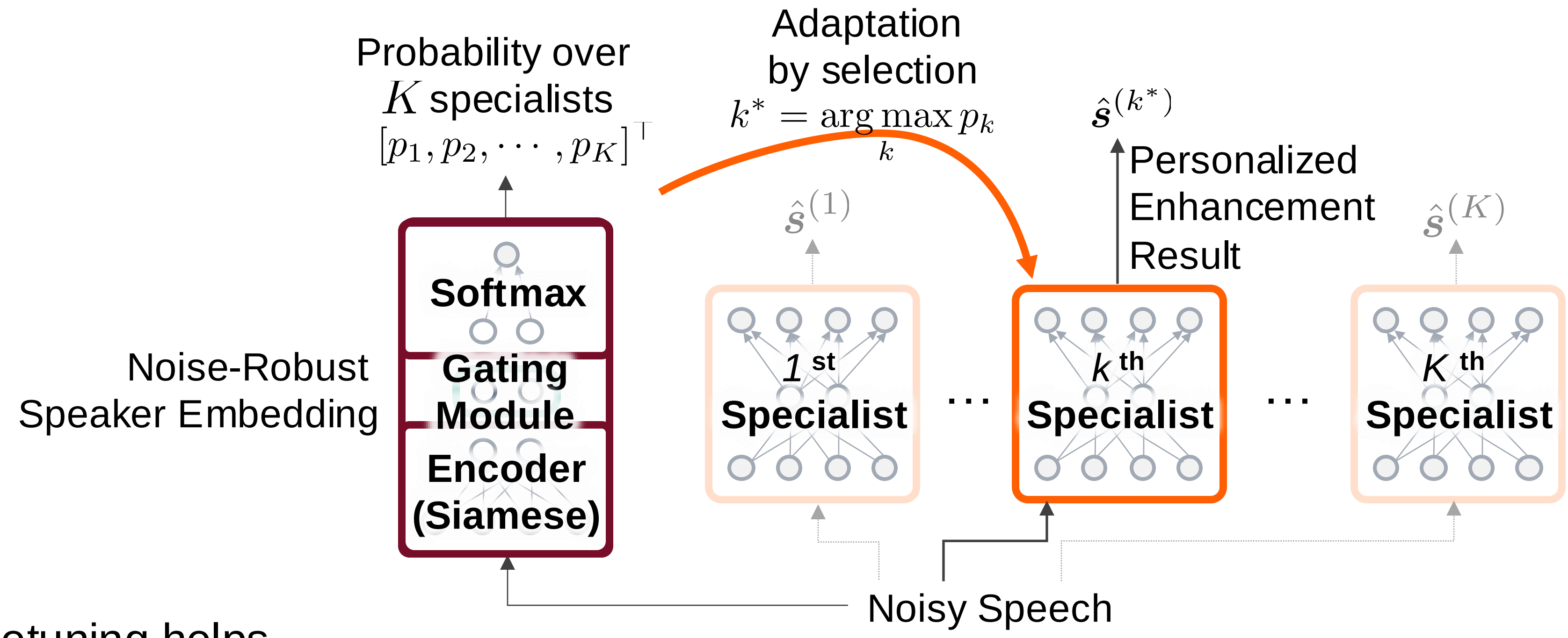
- Contrastive learning for noise-robust speaker embedding



Test-Time Model Selection

- Speaker-Specific Sparse Ensemble of Specialists

- Speaker-specific specialists

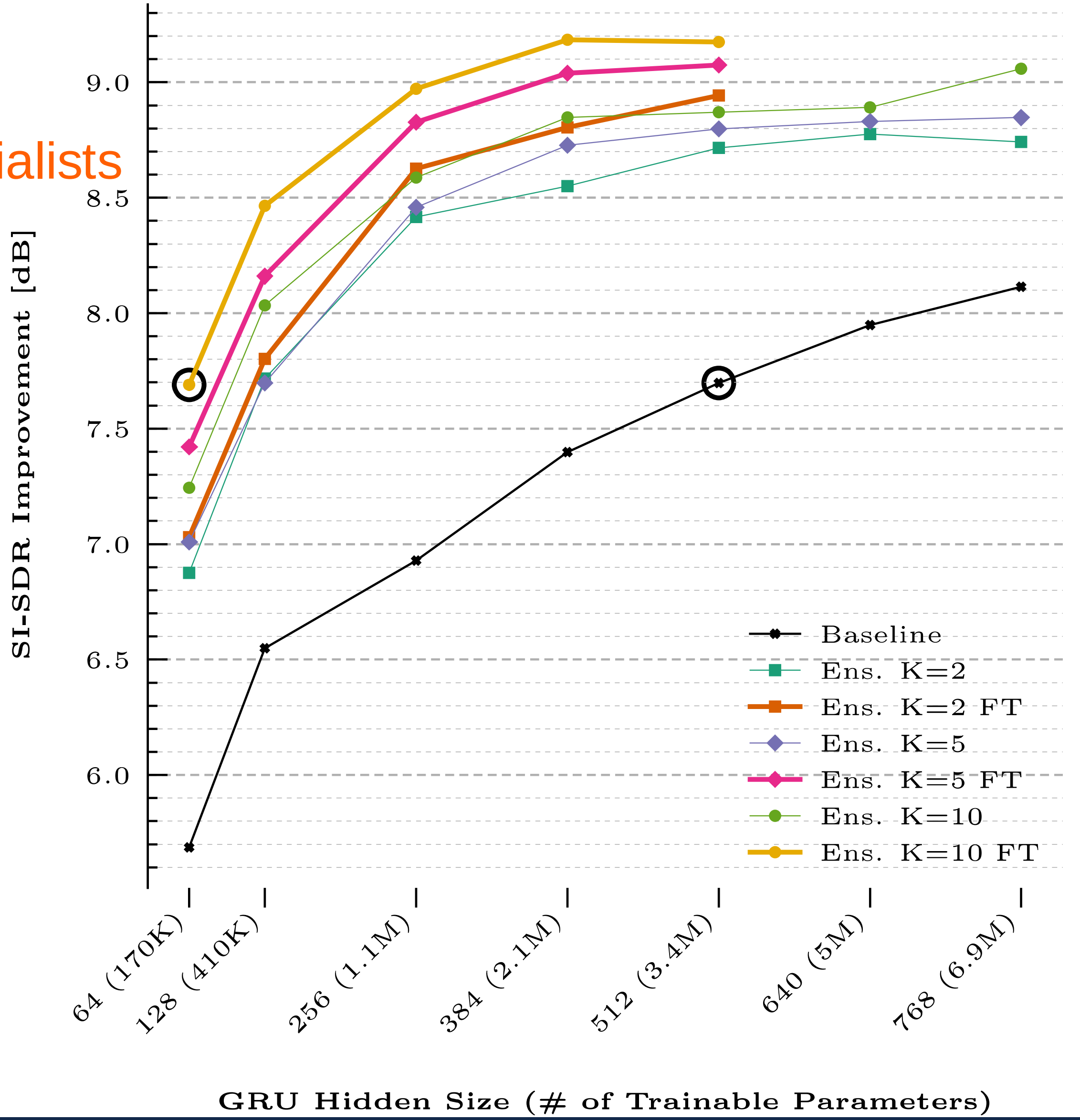


- Finetuning helps
  - Can refine speaker groups
  - Can make the gating module robust

Test-Time Model Selection

- Speaker-Specific Sparse Ensemble of Specialists

- Baseline: a generalist GRU model
- All proposed models outperform the baseline
- By increasing ***K***, performance increases
- Finetuning lifts the performance in all cases
- The smallest specialists is on par with a large generalist
    - A 95%-reduction in inference complexity
    - Plus a 50%-reduction in spatial complexity

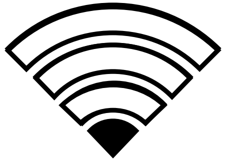


Scalable and Efficient Speech Enhancement

Motivation

- Scalability and Efficiency

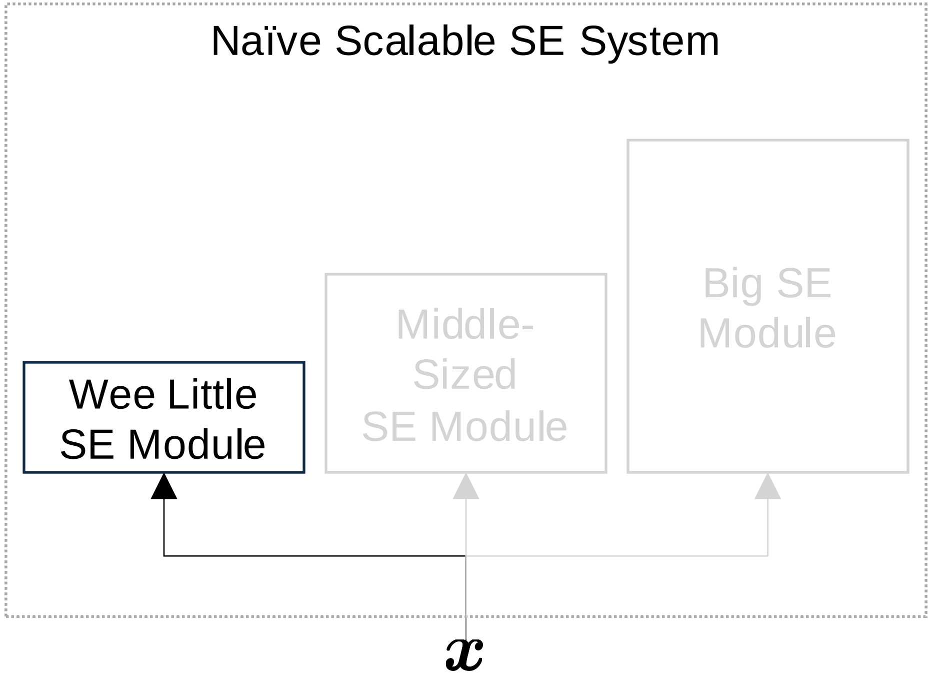
- Scalability in video coding
  - Video codec adjusts bitrate

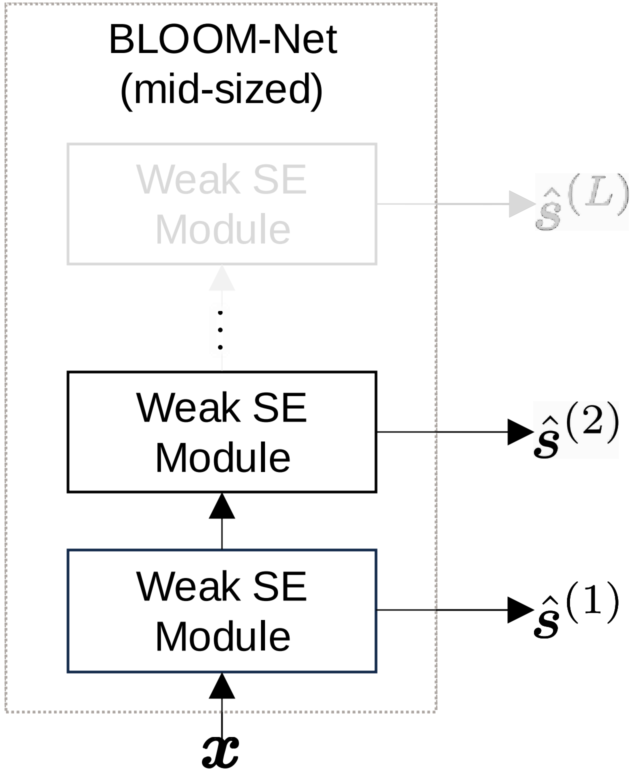


- Our goal: the trade-off between performance and resource usage
  - Speech enhancement quality vs model complexity

Motivation

- A naïve scalable SE system and BLOOM-Net



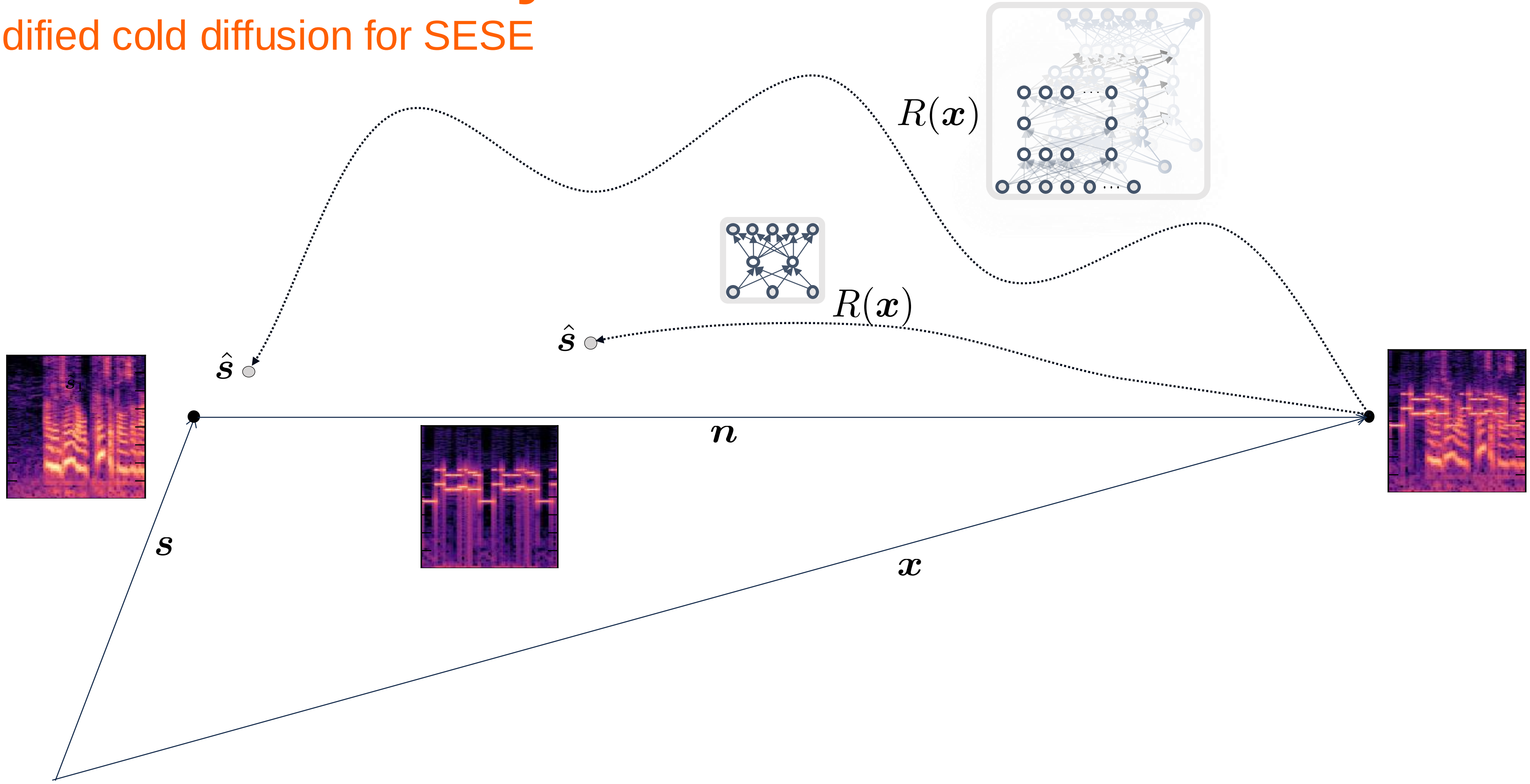


- $1M + 2M + 3M = 6M$ params
- Modules do not communicate, wasting computation

- BLOOM-Net is dependent on masking-based architectures
  - BLOOM-Net: BLOck-wise Optimization of Masking Networks
- Missing components: residual learning, milestone goals

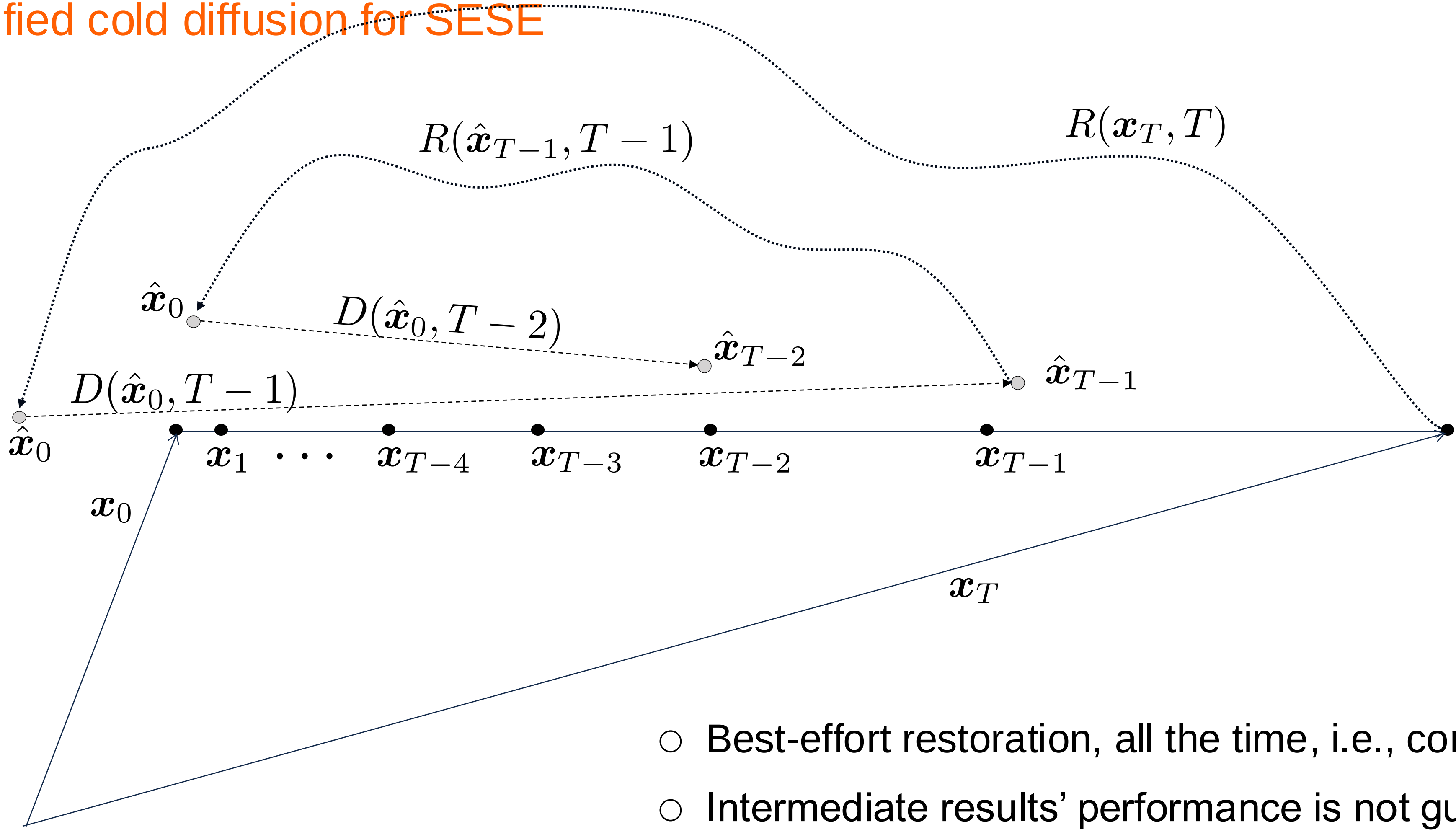
Baseline 1: Ordinary One-Shot Inference Model

- Modified cold diffusion for SESE



Baseline 2: Iterative Inference Model (Cold Diffusion)

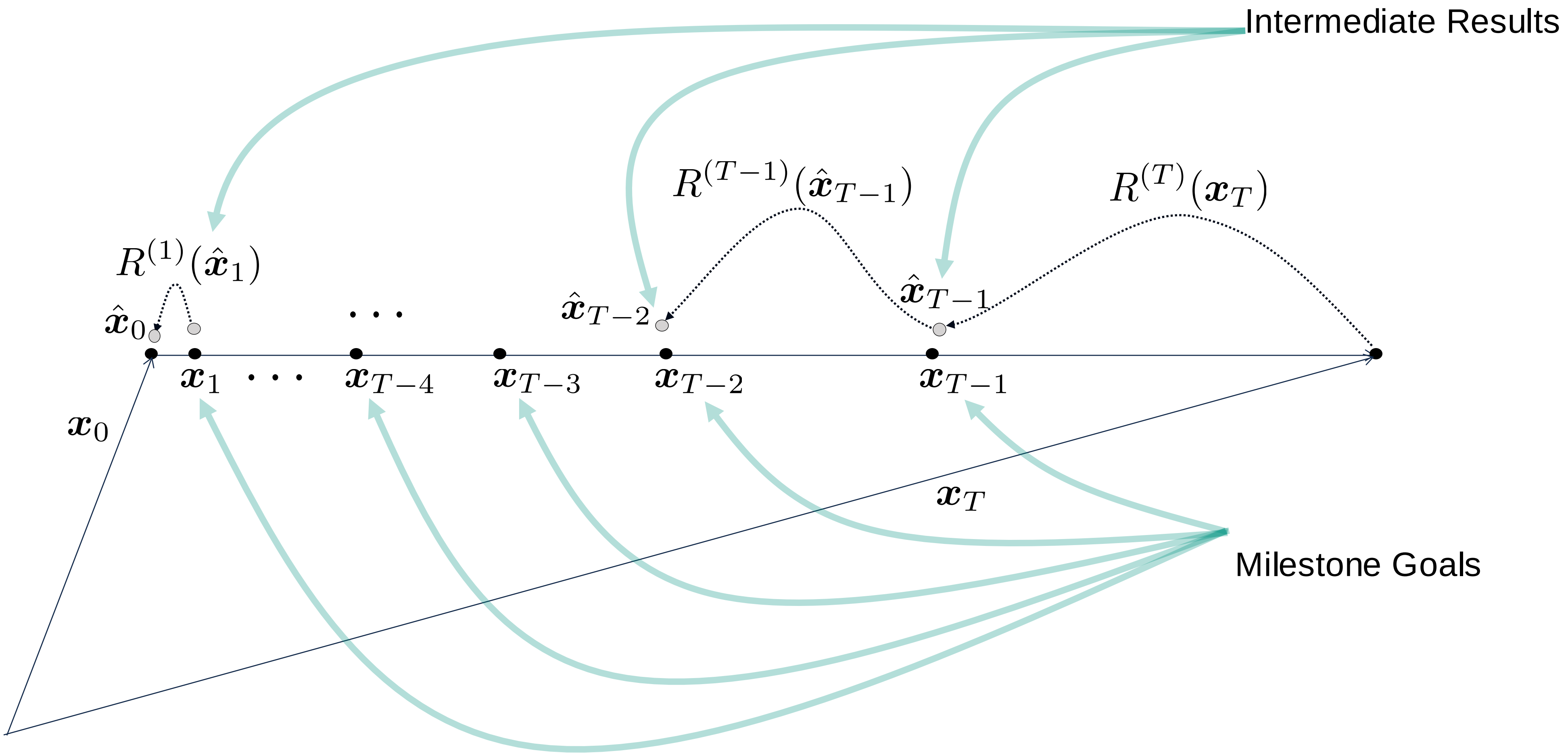
- Modified cold diffusion for SESE



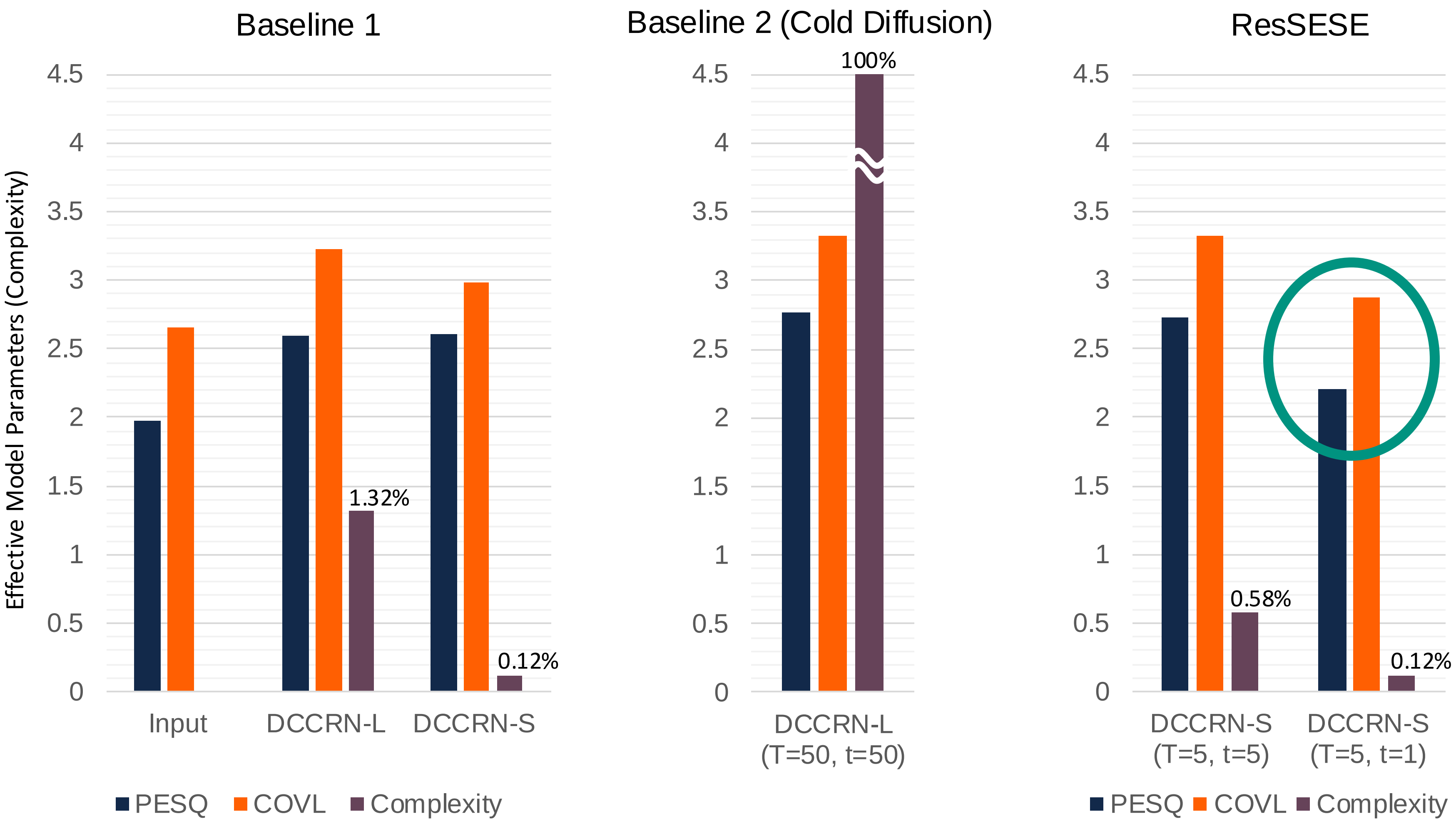
- ○ Best-effort restoration, all the time, i.e., complex
- ○ Intermediate results' performance is not guaranteed

SESE via Modified Cold Diffusion

- The Proposed Method



Experimental Results (Voicebank + DEMAND)

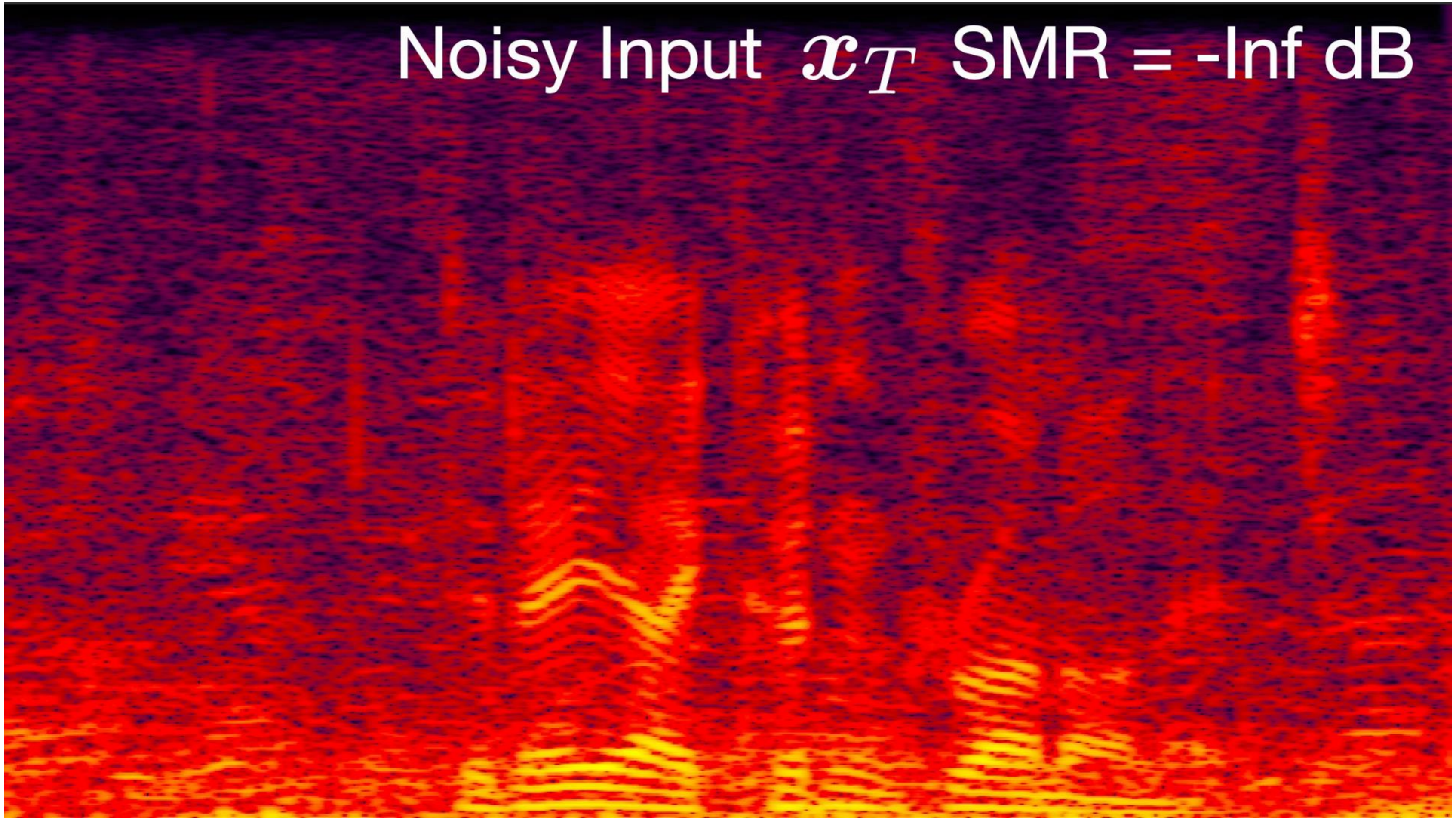


Unprocessed SIR: 8.66

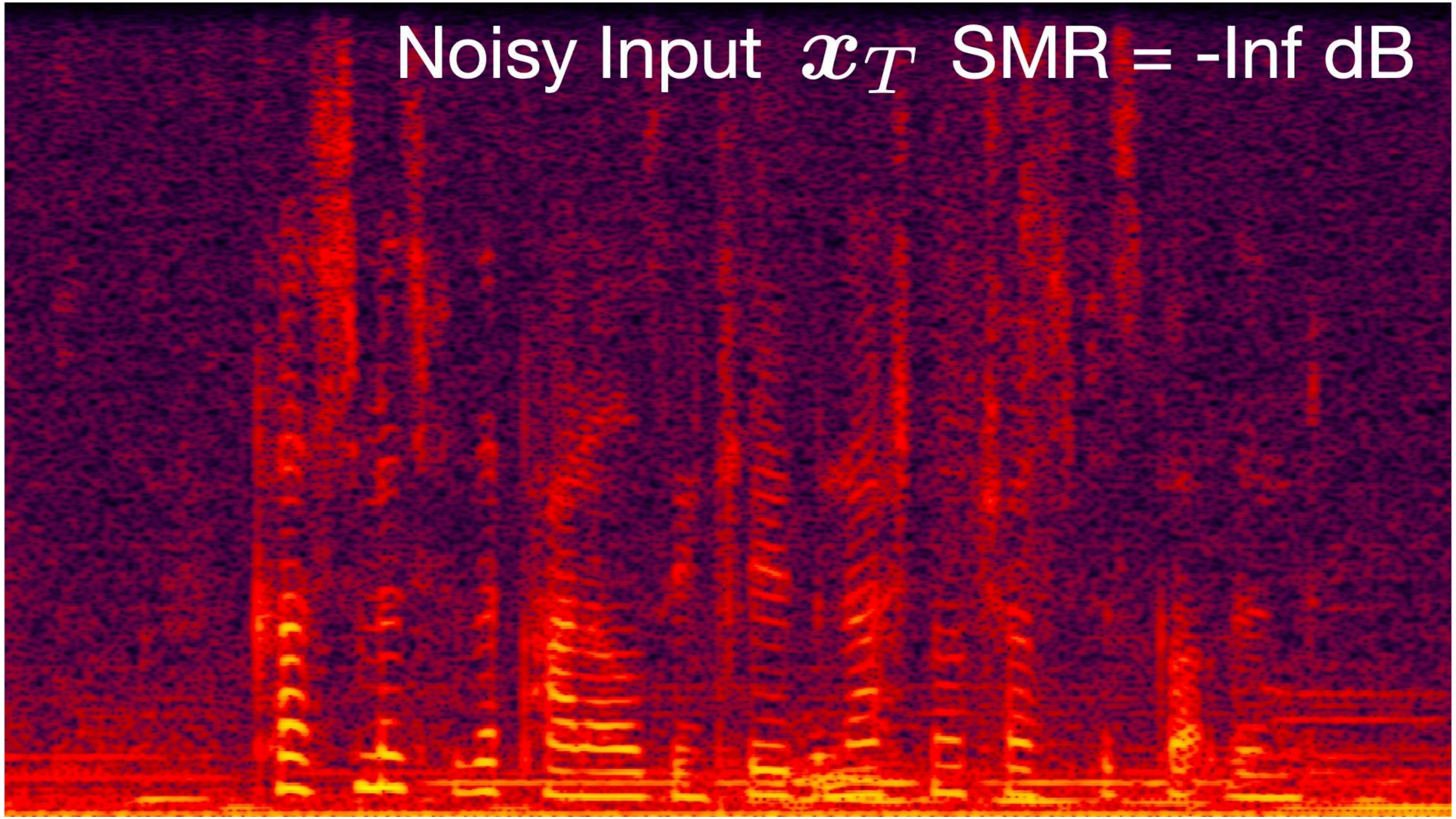
DCCRN-S (T=5, t=1)
SIR: 12.20 dB (+3.54)
SAR: 31.44 dB
Sounds gentle
Less oversuppression

ResSESE is
better than BL1
on par with BL2
still more compact than both

Audio Demo #1



Audio Demo 2

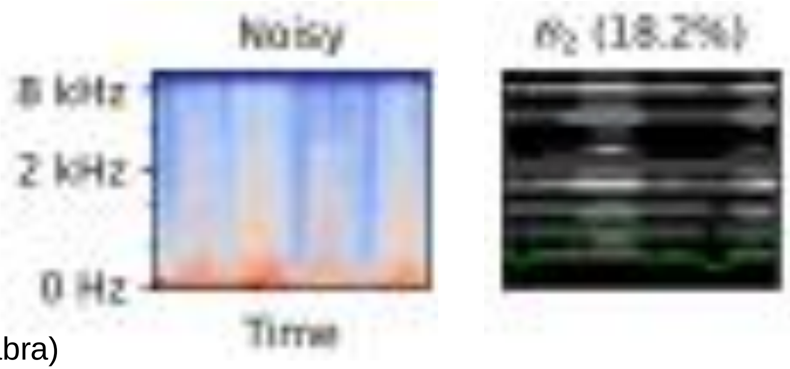


Discussion

- On-device inference is costly
  - Efficiency matters
- Task-aware adaptation can achieve high efficiency
  - e.g., personalized SE
  - Could be sensitive to domain mismatch
- Scalable models are underexplored

Riccardo Miccini et al.,
“Scalable Speech Enhancement with Dynamic Channel Pruning,”
ICASSP 2025
https://doi.org/10.48550/arXiv.2412.17121

Riccardo Miccini
(Technical University of Denmark; GN Audio/Jabra)



References

- *Minje Kim* and Paris Smaragdis, "**Bitwise Neural Networks**," ICML Workshop on Resource-Efficient Machine Learning, 2015 [pdf]
- *Minje Kim* and Paris Smaragdis, "**Bitwise Neural Networks for Efficient Single-Channel Source Separation**," ICASSP 2018 [pdf]
- Sunwoo Kim, Mrinmoy Maity, and *Minje Kim*, "**Incremental Binarization On Recurrent Neural Networks for Single-Channel Source Separation**," ICASSP 2019 [pdf, code]
- Sunwoo Kim, Haici Yang, and *Minje Kim*, "**Boosted Locality Sensitive Hashing: Discriminative Binary Codes for Source Separation**," ICASSP 2020 [pdf, demo, code, presentation video] **<Finalist for the Best Student Paper Award>**
- Sunwoo Kim and *Minje Kim*, "**Boosted Locality Sensitive Hashing: Discriminative, Efficient, and Scalable Binary Codes for Source Separation**," ***IEEE/ACM Transactions on Audio, Speech, and Language Processing***, vol. 30, pp. 2659-2672, Aug. 2022 [pdf, demo, code, presentation video]
- *Minje Kim*, "**Collaborative Deep Learning for Speech Enhancement: A Run-Time Model Selection Method Using Autoencoders**," ICASSP 2017 [pdf]
- Sunwoo Kim and *Minje Kim*,"**Test-Time Adaptation Toward Personalized Speech Enhancement: Zero-Shot Learning With Knowledge Distillation**," WASPAA 2021 [pdf, code, demo, presentation video]
- Sunwoo Kim, Mrudula Athi, Guangji Shi, *Minje Kim*, and Trausti Kristjansson, "**Zero-Shot Test-Time Adaptation Via Knowledge Distillation for Personalized Speech Denoising and Dereverberation**," ***Journal of Acoustical Society of America***, Vol. 155, No. 2, pp 1353-1367, Feb. 2024 [pdf]
- Aswin Sivaraman and *Minje Kim*, "**Sparse Mixture of Local Experts for Efficient Speech Enhancement**," Interspeech 2020 [pdf, demo, code, presentation video]
- Aswin Sivaraman and *Minje Kim*, "**Zero-Shot Personalized Speech Enhancement Through Speaker-Informed Model Selection**," WASPAA 2021 [pdf, code, presentation video]
- Sunwoo Kim and *Minje Kim*, "**BLOOM-Net: Blockwise Optimization for Masking Networks Toward Scalable and Efficient Speech Enhancement**," ICASSP 2022 [pdf, demo, code, presentation video]
- *Minje Kim* and Trausti Kristjansson, "**Scalable and Efficient Speech Enhancement Using Modified Cold Diffusion: a Residual Learning Approach**," ICASSP 2024 [pdf, demo].

Thank You!
(Q&A)

Minje Kim, Ph.D.
https://minjekim.com
minje@illinois.edu