Interspeech 2022 Tutorial (09/17/2022)

Personalized Speech Enhancement:

Data- and Resource-Efficient Machine Learning

Minje Kim

Associate Professor, Indiana Uinversity

Visiting Academic, Amazon Lab126

minje@indiana.edu

https://minjekim.com

https://minjekim.com/research-projects/pse/

Part of this material is based upon work supported by the National Science Foundation under Grant No. 2046963.

The papers introduced in this talk are not associated with Amazon.

Outline

- Motivation
  - Generalist vs. specialist
    - Data and resource efficiency
    - Performance
    - Fairness
    - Privacy preservation
- Zero-Shot PSE
  - Primitive NMF models
  - Test-Time Model Adaptation
  - Test-Time Model Selection
- Few-Shot PSE
  - Target Speaker Extraction as PSE
  - Self-Supervised Learning
  - Data Purification
  - Contrastive Mixtures
- Conclusion & Discussion

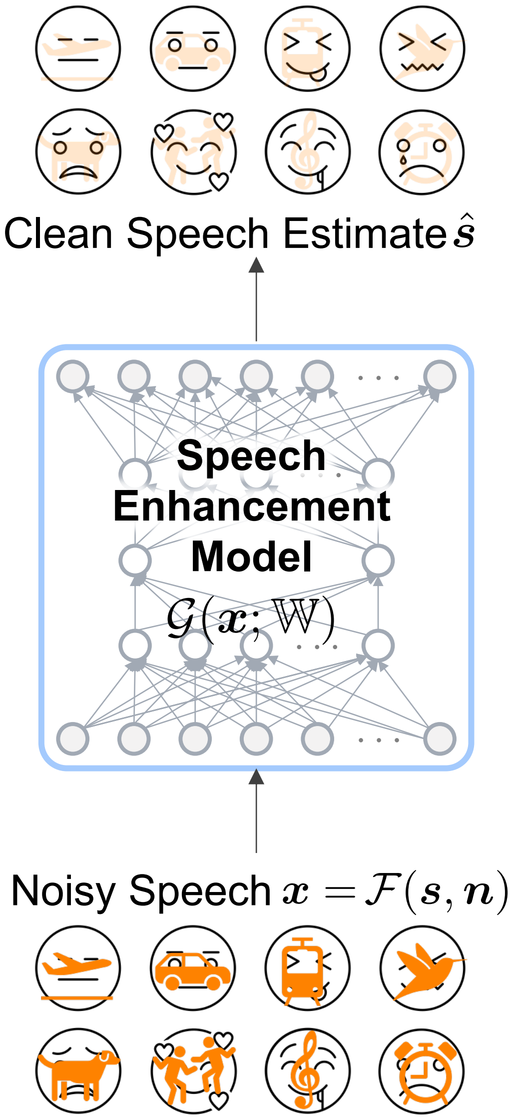
Motivation

- Machine learning-based speech enhancenent approaches

- A typical supervised setup
  - Artificial filtering $\boldsymbol{x} = \mathcal{F}(\boldsymbol{s}, \boldsymbol{n}) = \boldsymbol{s} + \boldsymbol{n}$
  - The goal is to learn another parametric function (e.g., a neural network)

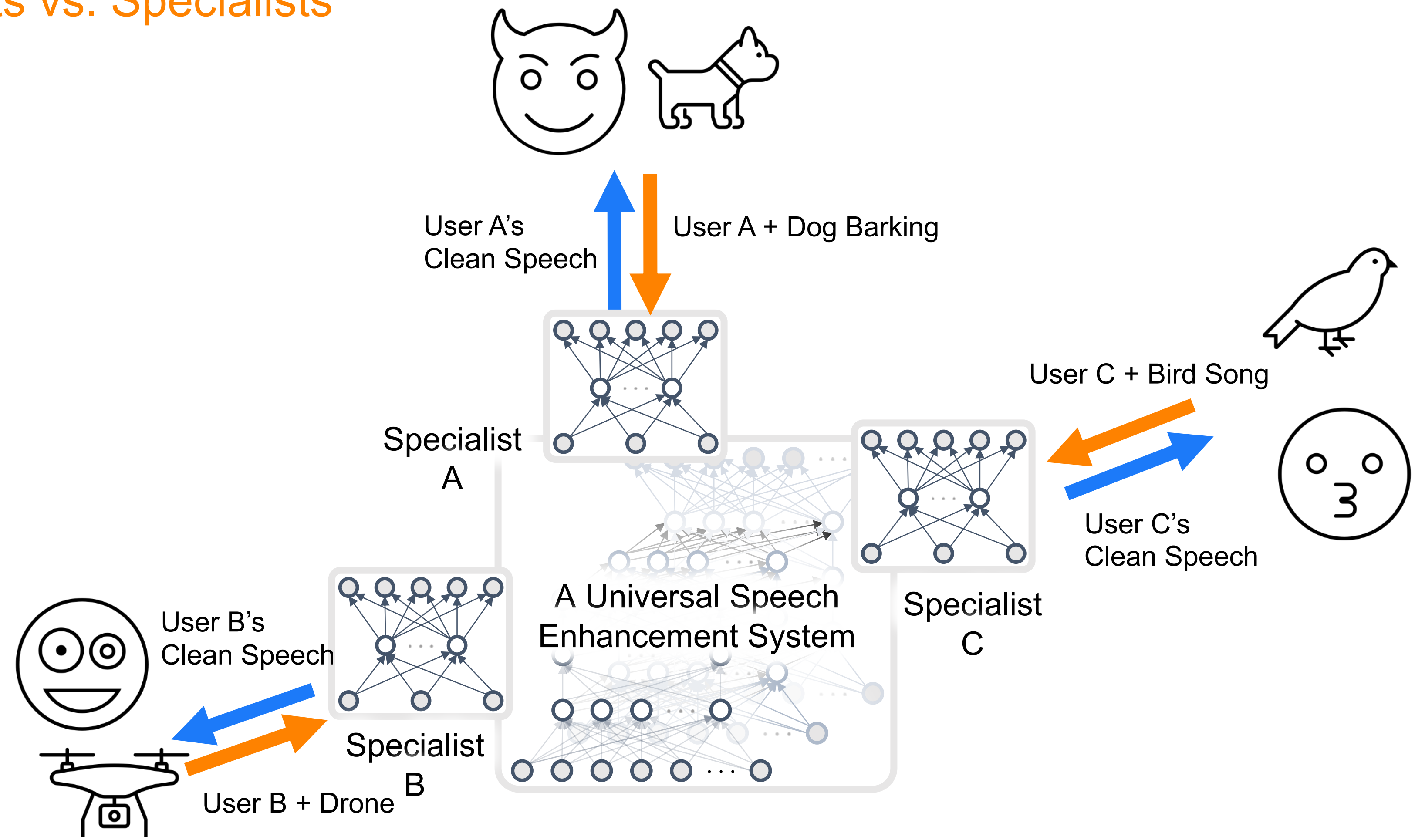
    $\boldsymbol{s} \approx \hat{\boldsymbol{s}} = \mathcal{G}(\boldsymbol{x}; \mathbb{W})$

- Issues
  - The deformation function $\mathcal{F}(\boldsymbol{s}, \boldsymbol{n})$ might be too artificial
    - Reverberation, band-pass filtering, etc.
  - Big data and big models
    - Deep learning advancements have relied on the big *labeled* data, i.e., $(\boldsymbol{x}, \boldsymbol{s})$
    - So the *big models generalize well*
  - Do we always need a big model?



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.

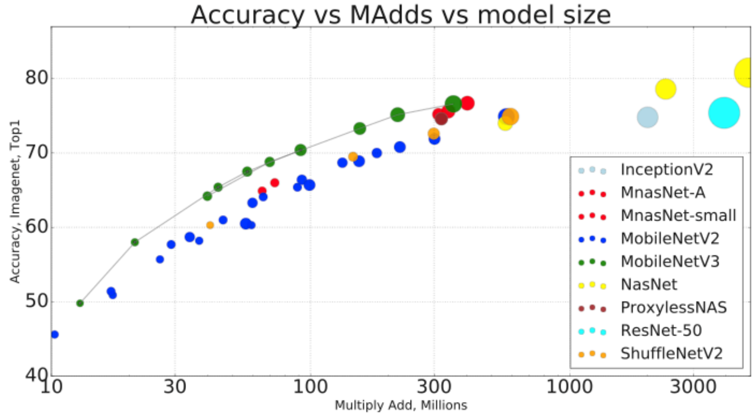
Motivation

- Generalists are heavy

○ DNNs are big

	# Weights	FLOP	WEIGHTS (%)	FLOP (%)
LeNet-300-100	266K	532K	8%	8%
LeNet-5	431K	4586K	8%	16%
AlexNet	61M	1.5B	11%	30%
VGG-16	138M	30.9B	7.5%	21%

S. Han et al., "Learning both Weights and Connections for Efficient Neural Networks," NIPS 2015.



A. Howard et al., "Searching for MobileNetV3" ICCV 2019

□ Lossless model compression?

○ Training DNNs is costly

Model	Hardware	CO2	CC Cost
Transformer (base)	P100x8	26	$41-$140
Transformer (big)	P100x8	192	$289-$981
ELMo	P100x3	262	$433-$1472
BERT (base)	V100x64	1438	$3751-$12,571
NAS	P100x8	626,155	$942,973-$3,201,722
Consumption			
Aire travel, 1 person, NY—SF		1984	
Human life, 1yr		11,023	
American life, 1yr		36,156	
Car, 1 lifetime		126,000	

E. Strubell et al., "Energy and Policy Considerations for Deep Learning in NLP," arXiv:1906.02243

□ A small model that just works well?

Motivation

- Bitwise neural networks for SE

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

Systems		Topology	SDR	STOI
FCN with original input		1024×2	10.17	0.7880
		2048×2	10.57	0.8060
FCN with binary input		1024×2	9.80	0.7790
		2048×2	10.11	0.7946
BNN		1024×2	9.35	0.7819
		2048×2	9.82	0.7861
GRU with binary input		1024×1	16.12	0.9459
BGRU	π=0.1	1024×1	15.50	0.9393
	π=0.2		15.17	0.9361
	π=0.3		14.90	0.9324
	π=0.4		14.58	0.9292
	π=0.5		14.32	0.9252
	π=0.6		14.02	0.9217
	π=0.7		13.66	0.9174
	π=0.8		13.30	0.9104
	π=0.9		12.70	0.9019
	π=1.0		11.76	0.8740

[Tan & Wang, ICASSP 2021]
[Luo et al., ICASSP 2021]

M. Kim and P. Smaragdis, "Bitwise Neural Networks for Efficient Single-Channel Source Separation," ICASSP 2018
S. Kim et al., "Incremental Binarization On Recurrent Neural Networks for Single-Channel Source Separation," ICASSP 2019

Motivation

- Generalists can be unfair

- Big data is easier to construct if you don't care about the fairness
- The consequence is an unfair model
- "Facial Recognition Is Accurate, if You're a White Guy" by Steve Lohr, New York Times (Feb 9, 2018)

  https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

Social Group	Classification Error (%)
Lighter-Skinned Males	1
Lighter-Skinned Female	7
Darker-Skinned Males	12
Darker-Skinned Female	35

- Personalization can be a solution!

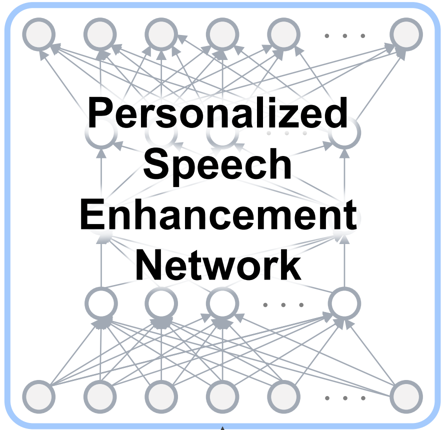
J. Buolamwini and T. Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification," *Conference on fairness, accountability and transparency*. 2018.

Motivation

- A naïve approach to personalized speech enhancement

- Supervised learning?

Clean Speech
of the Test-Time User



Noisy Speech
of the Test-Time User

- The clean speech of the test-time user is rare
  - **Privacy**: people are reluctant to share their clean voice

  - **Technical issues**: people might not be equipped to record their clean voice
    - Microphones, anechoic rooms, etc.
    - “Clean recordings” might not be clean enough

https://neosapience.com; https://typecast.ai

Motivation

- Summary of generalists vs. specialists

Property	Generalists	Specialists
Performance	Overall Good	Can Be Better
Generalization	High	Low
Computational Efficiency	Low	High
Data Efficiency	Low	High
Training	Heavy, But Straightforward	Light, But Complicated
Privacy Preservation	Potentially High	Potentially Low
Social Fairness	Low	High

- How do we specialize a model?
  - I mean, for a particular user
  - And, his/her test environment
  - And, during the test time?

Personalized Speech Enhancement

- Zero- or few-shot learning

- ○ Zero-shot PSE
  - □ Adaptively learns from the test-time environment
  - \+ No need to collect the enrollment signals
  - \+ Privacy-preserving (to some degree)
  - \- How do we do this?
- ○ Few-shot PSE
  - □ A finetuning method that adapts an existing SE system to the test-time environment
  - □ Does require test-time clean speech, but only a small amount
  - \+ Performance might be better than no-shot training
  - \- Still requires clean speech
  - \- Overfitting

Zero-Shot PSE

Primitive NMF Models

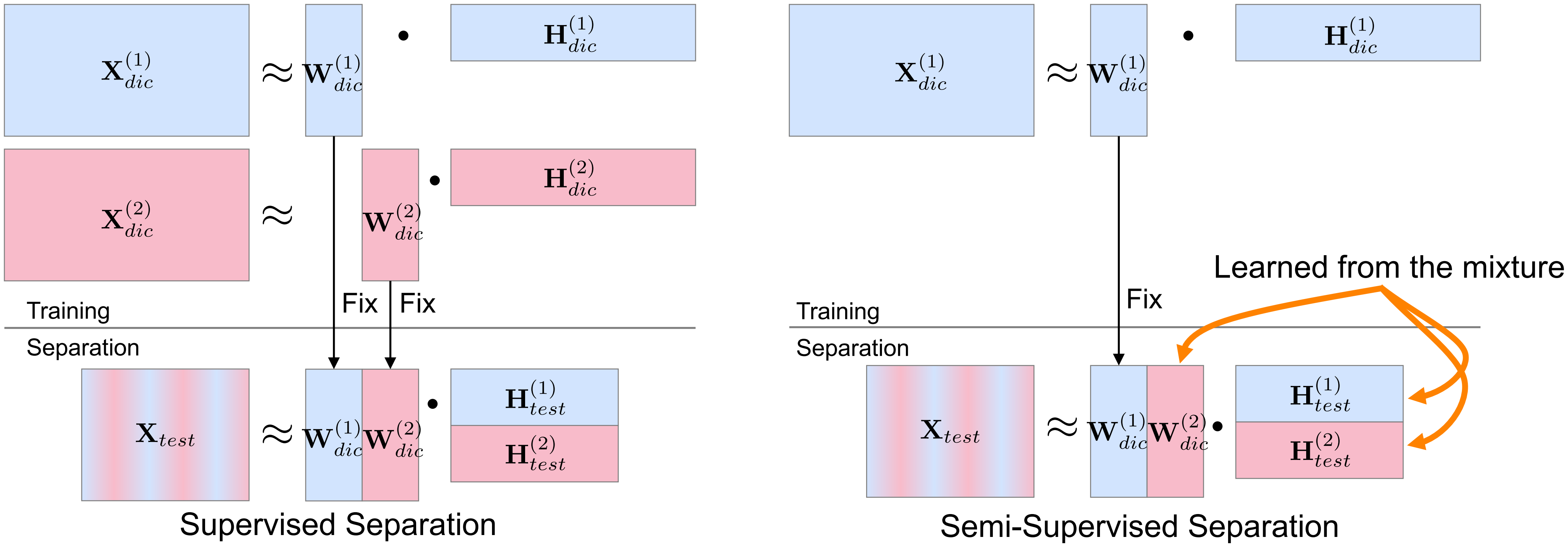
Test-Time Model Adaptation

Test-Time Model Selection

Semi-Supervised Nonnegative Matrix Factorization

- For Speaker Adaptation

- Traditional nonnegative matrix factorization (NMF) for SE



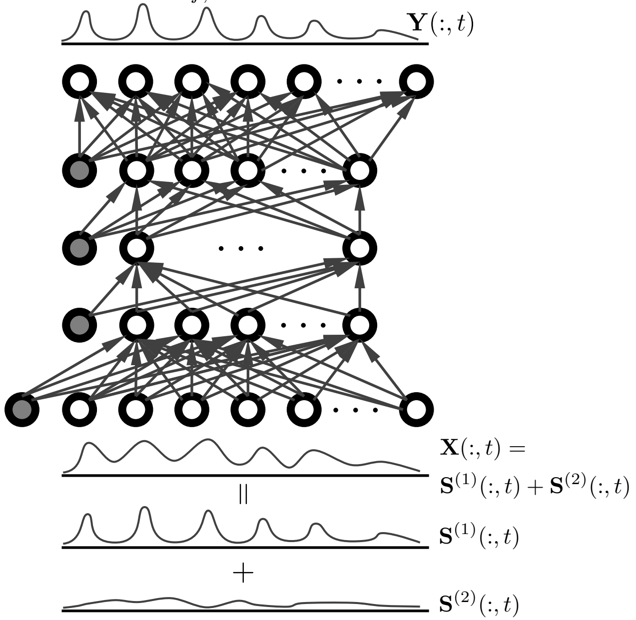
- Assumes that the noise source type is known
- Weak supervision (generative vs. discriminative models)

Z. Duan et al., “Speech Enhancement by Online Non-negative Spectrogram Decomposition in Non-stationary Noise Environments,” Interspeech 2012

Test-Time Model Adaptation

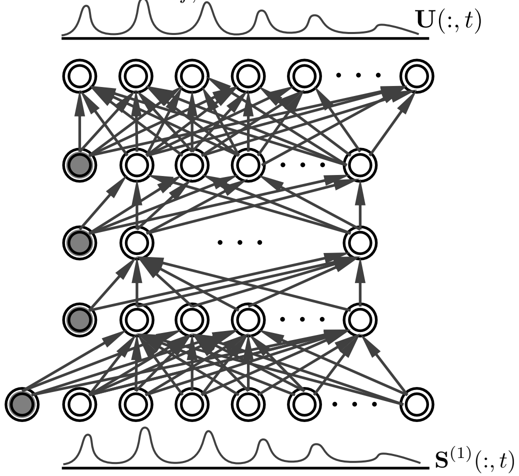
- Adaptive Denoising Autoencoders

$$\mathcal{E}_{DAE} = \frac{1}{2}\sum_{f,t}\left(\mathbf{Y}(f,t) - \mathbf{S}^{(1)}(f,t)\right)^2.$$



A DAE(not good enough)

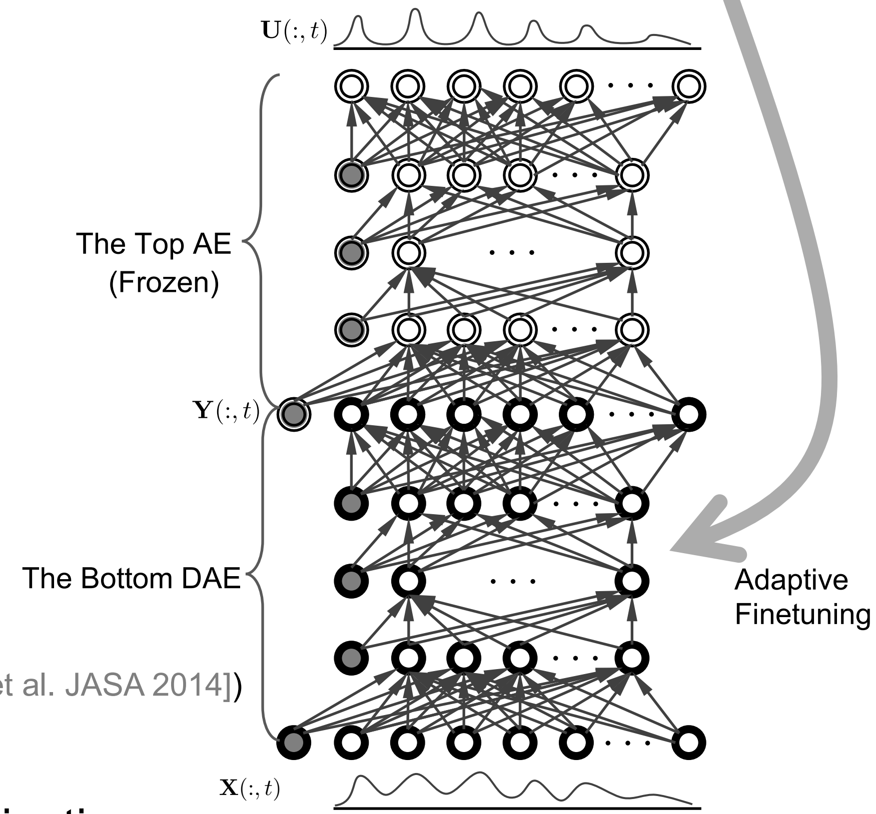
$$\mathcal{E}_{AE} = \frac{1}{2}\sum_{f,t}\left(\mathbf{U}(f,t) - \mathbf{S}^{(1)}(f,t)\right)^2.$$



An AE (purity checker)

Anything else?
(e.g., NMF [Williamson et al. JASA 2014])

$$\mathcal{E}_{AE} = \frac{1}{2}\sum_{f,t}\left(\mathbf{U}(f,t) - \mathbf{Y}(f,t)\right)^2$$



The Adaptive DAE

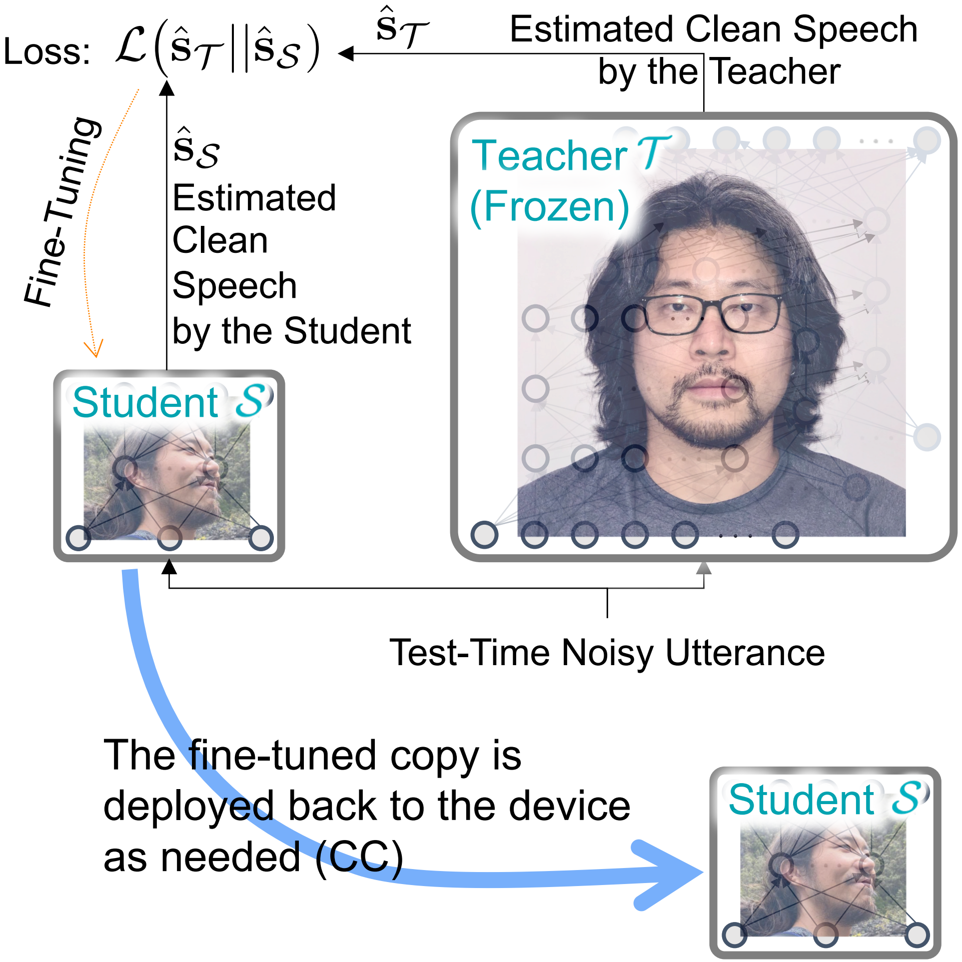
- Proves the concept, but we don't always need this sophistication
- Maybe not the best way to harmonize the two networks

M. Kim and P. Smaragdis, "Adaptive Denoising Autoencoders: A Fine-tuning Scheme to Learn from Test Mixtures," LVA/ICA 2015

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
- Test-time adaptation
  - Distill teacher's outputs as pseudo-targets to 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

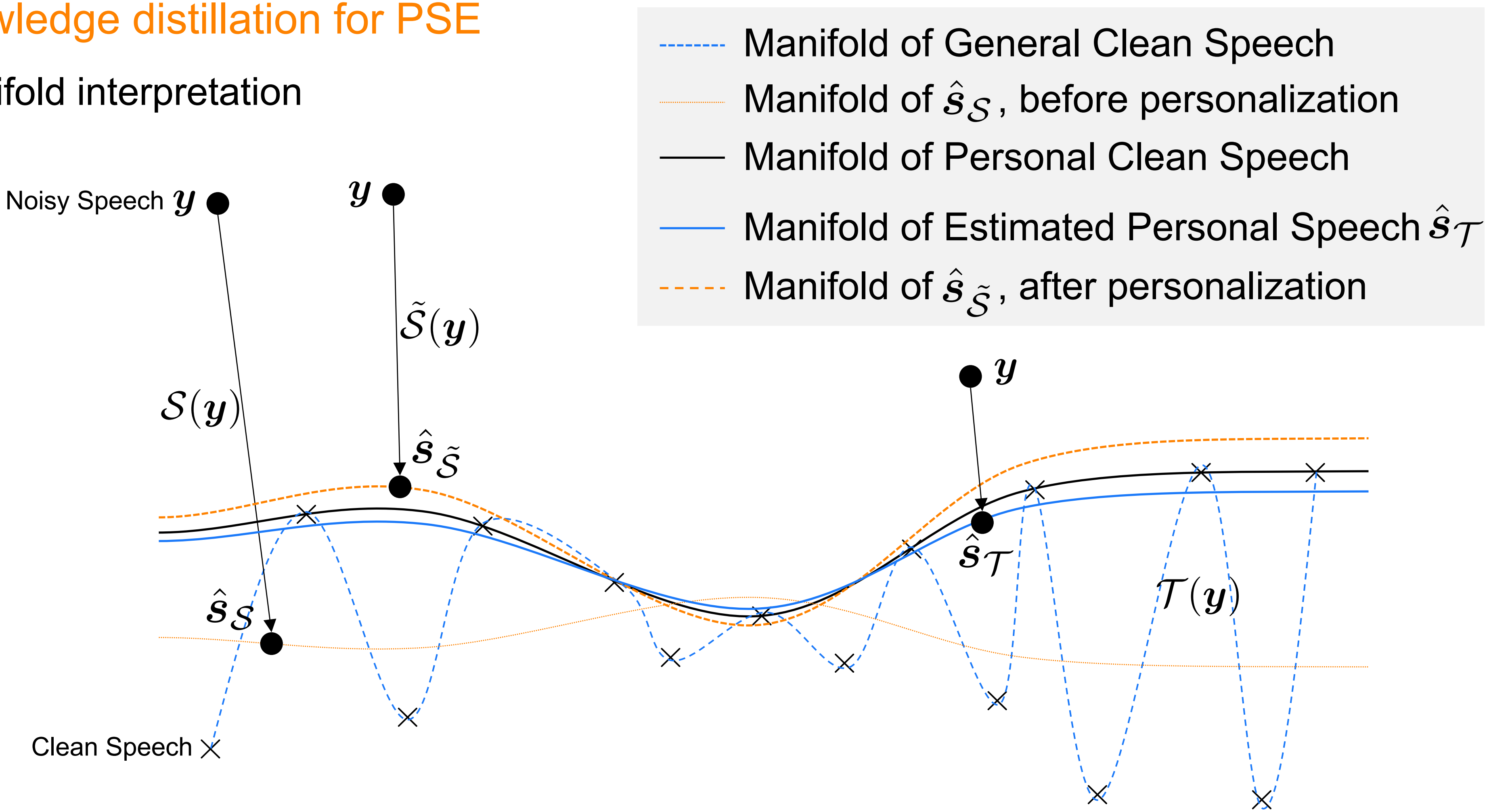


G. Hinton et al., "Distilling the Knowledge in a Neural Network," arXiv:1503.02531
S. Kim and M. Kim, "Test-Time Adaptation Toward Personalized Speech Enhancement: Zero-Shot Learning With Knowledge Distillation," WASPAA 2021

Test-Time Model Adaptation

- Knowledge distillation for PSE

- Manifold interpretation

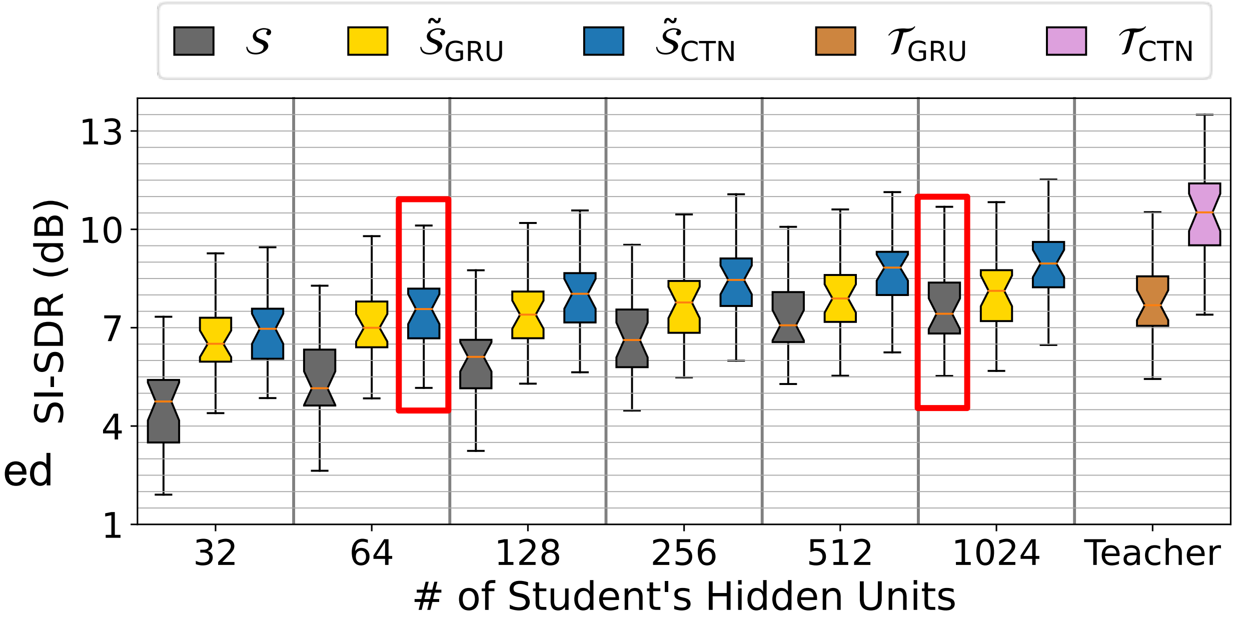


S. Kim and M. Kim, "Test-Time Adaptation Toward Personalized Speech Enhancement: Zero-Shot Learning With Knowledge Distillation," WASPAA 2021

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



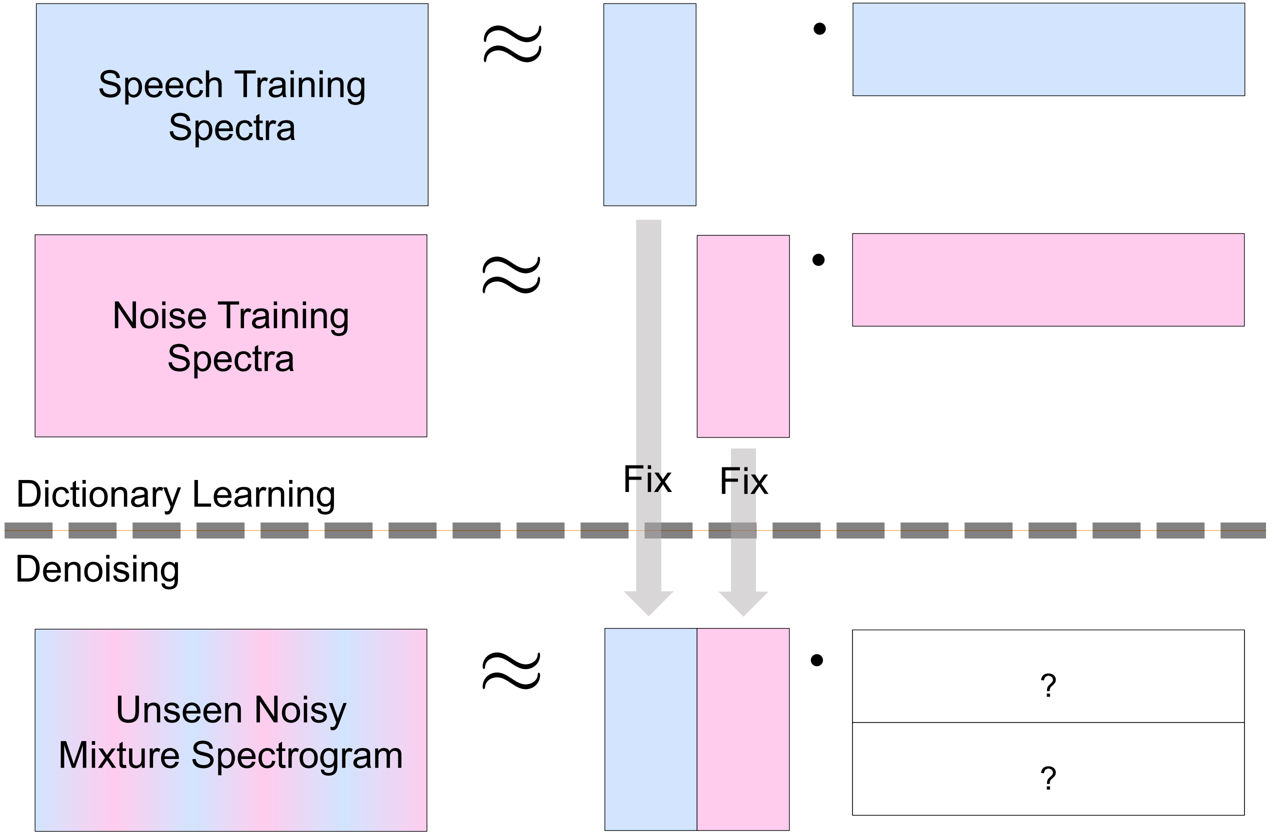
- PSE consistently outperforms all pre-trained student models
  - More improvement on smaller architectures
- $\tilde{S}_{CTN}$ always outperforms their corresponding $\tilde{S}_{GRU}$
- Lossless network compression
  - 2 x 64 $\tilde{S}_{CTN}$ vs. 2 x 1024 S
    - ~66x lower MACs and parameters

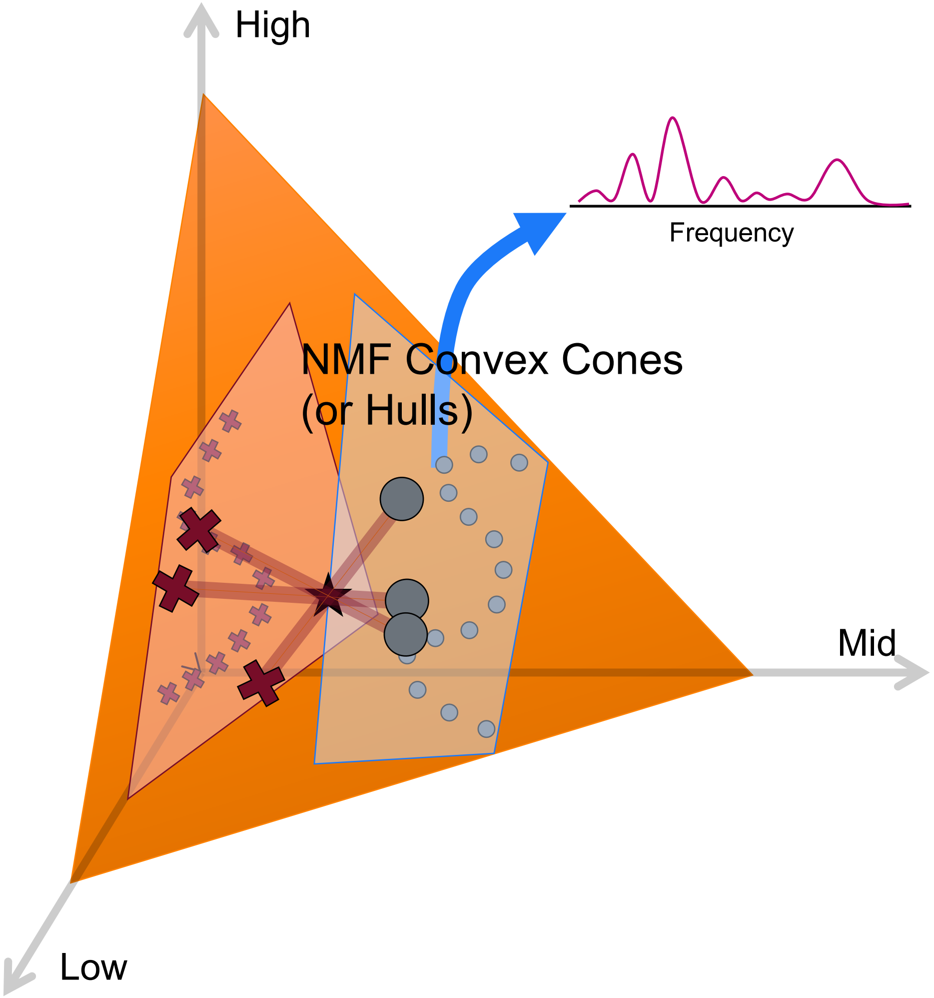
S. Kim and M. Kim, "Test-Time Adaptation Toward Personalized Speech Enhancement: Zero-Shot Learning With Knowledge Distillation," WASPAA 2021

Speaker-Specific Dictionaries for PSE

- The universal speech model

- Traditional nonnegative matrix factorization (NMF) for SE



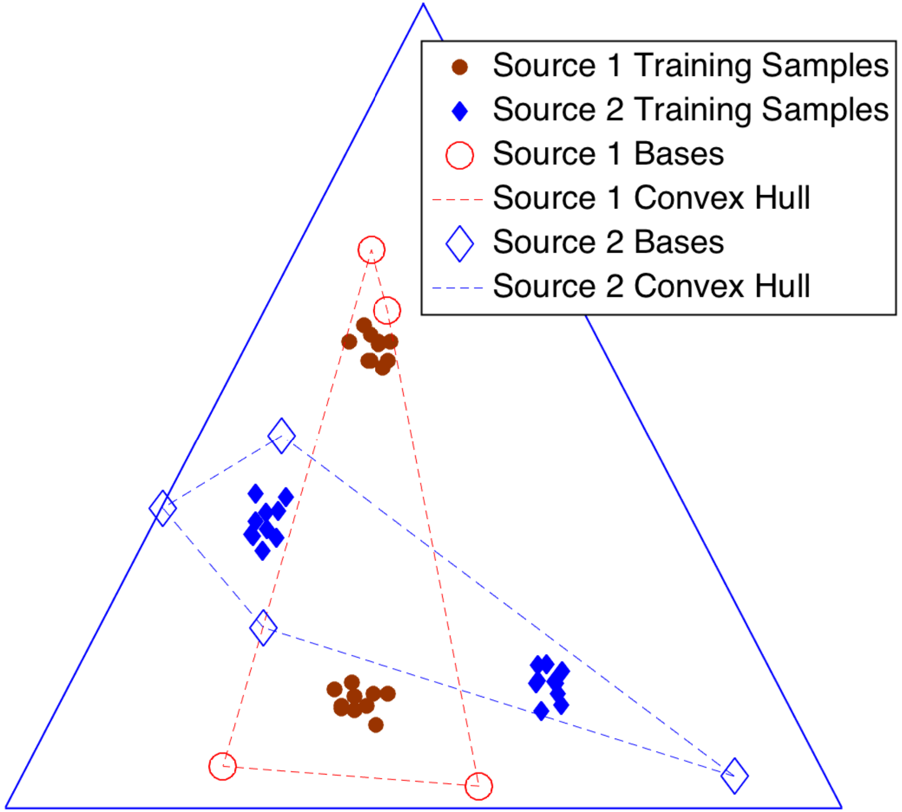


P. Smaragdis et al. “A sparse non-parametric approach for single channel separation of known sounds,” NIPS 2009
M. Kim and P. Smaragdis, “Manifold Preserving Hierarchical Topic Models for Quantization and Approximation,” ICML 2013

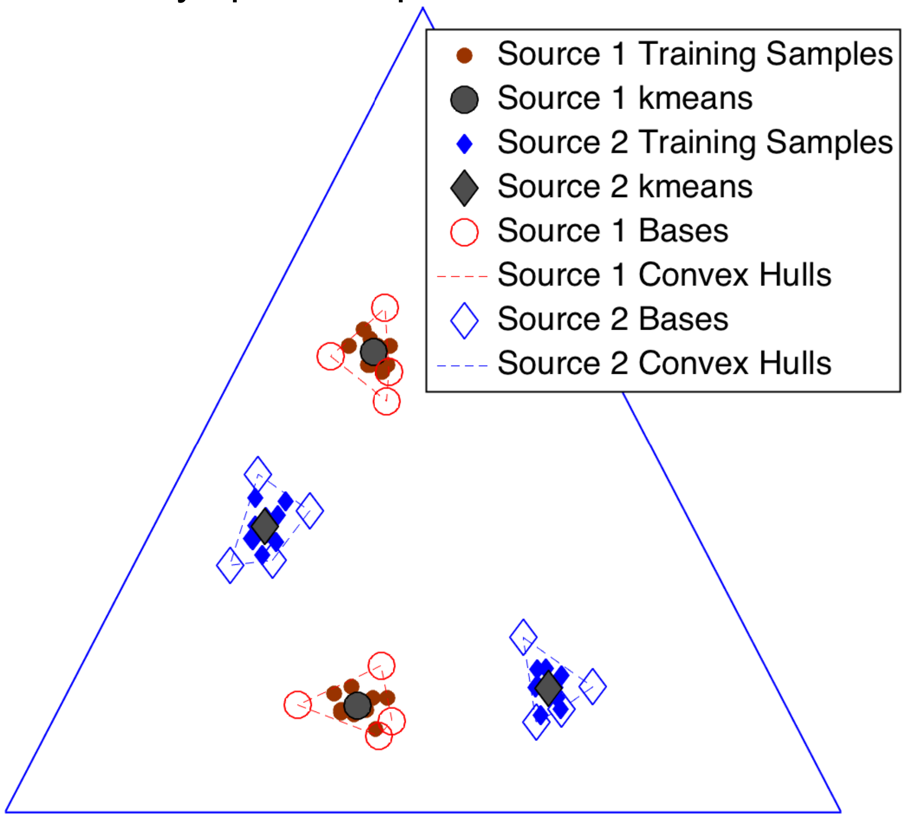
Speaker-Specific Dictionaries for PSE

- The universal speech model

- In practice, NMF dictionaries define too large subspaces



- Tightly defined dictionaries
  - Preferably speaker-specific

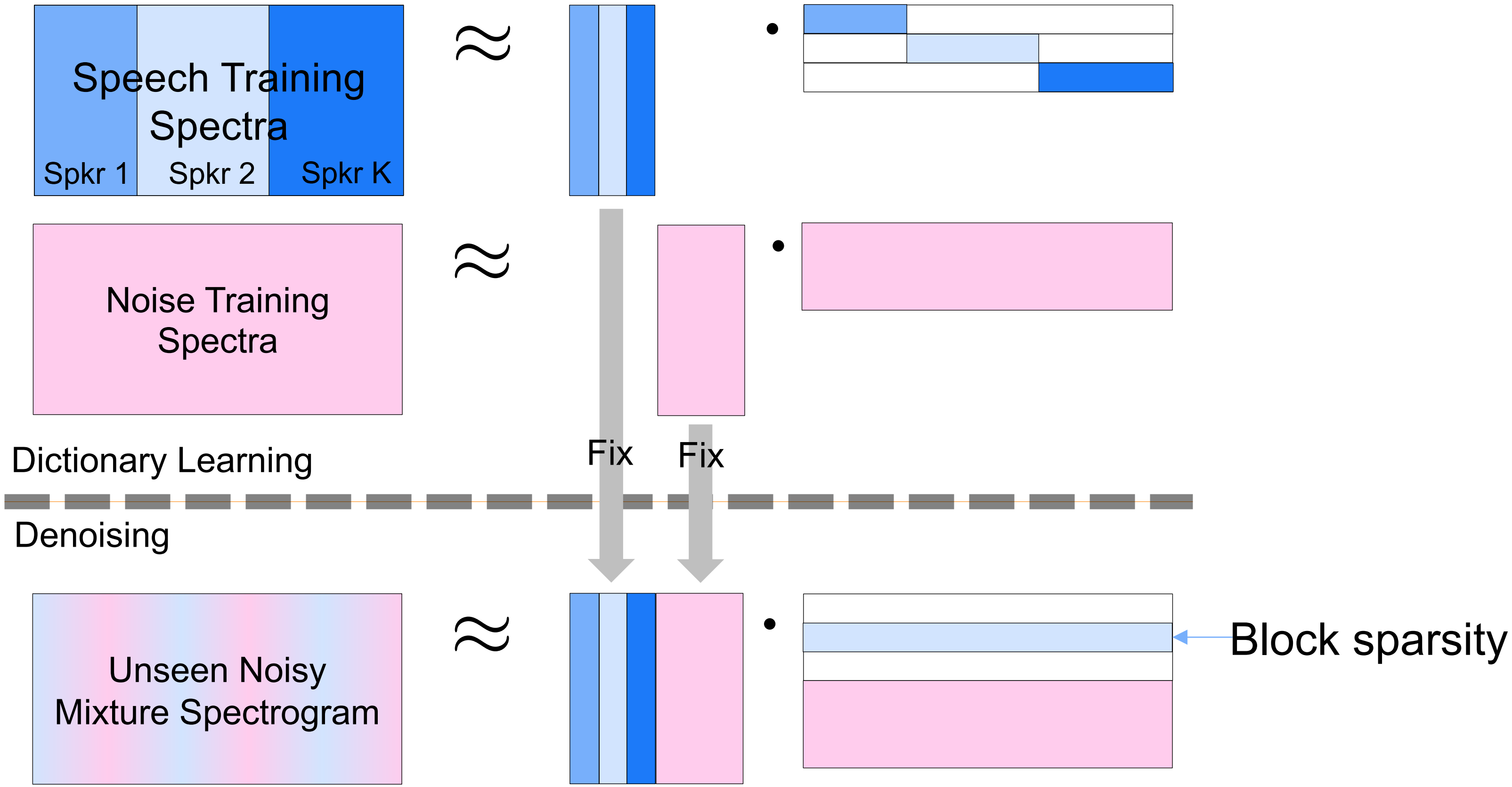


M. Kim and P. Smaragdis, “Mixtures of Local Dictionaries for Unsupervised Speech Enhancement,” IEEE SPL, 2015

Speaker-Specific Dictionaries for PSE

- The universal speech model

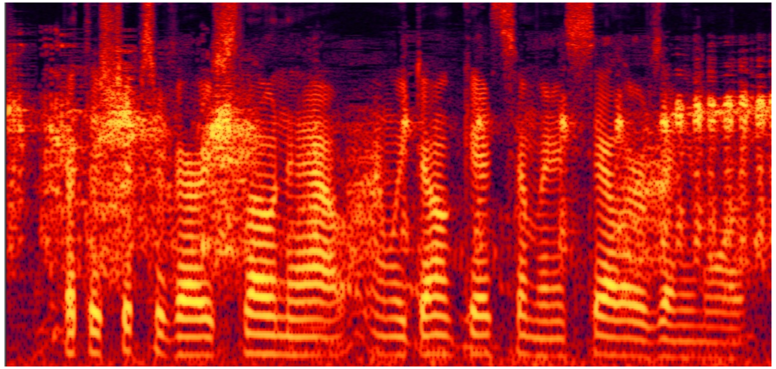
- In the USM, block sparsity ensures few speaker-specific dictionaries are used for denoising

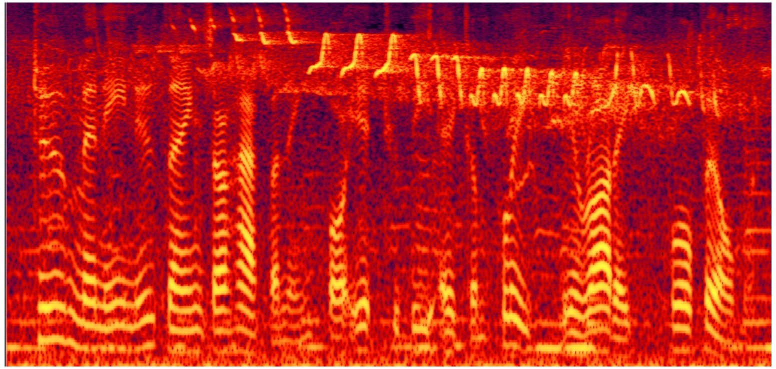


D. L. Sun et al., "Universal speech models for speaker independent single channel source separation," ICASSP 2013

Speaker-Specific Dictionaries for PSE

- The universal speech model



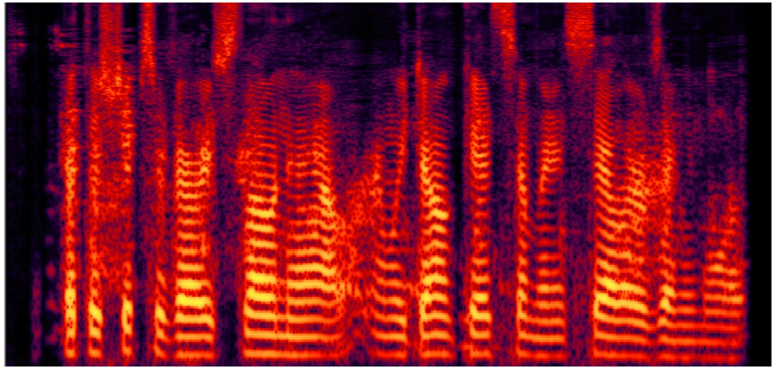


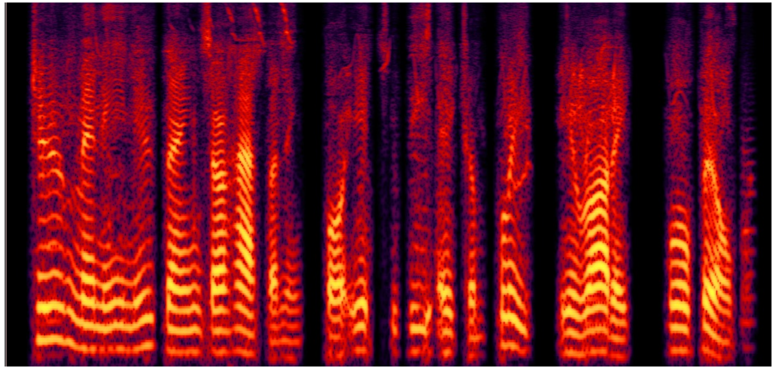
Mixture (frogs)

Mixture (birds)

Separation

Separation





M. Kim and P. Smaragdis, “Mixtures of Local Dictionaries for Unsupervised Speech Enhancement,” IEEE SPL, 2015

Test-Time Model Selection

- Collaborative deep learning

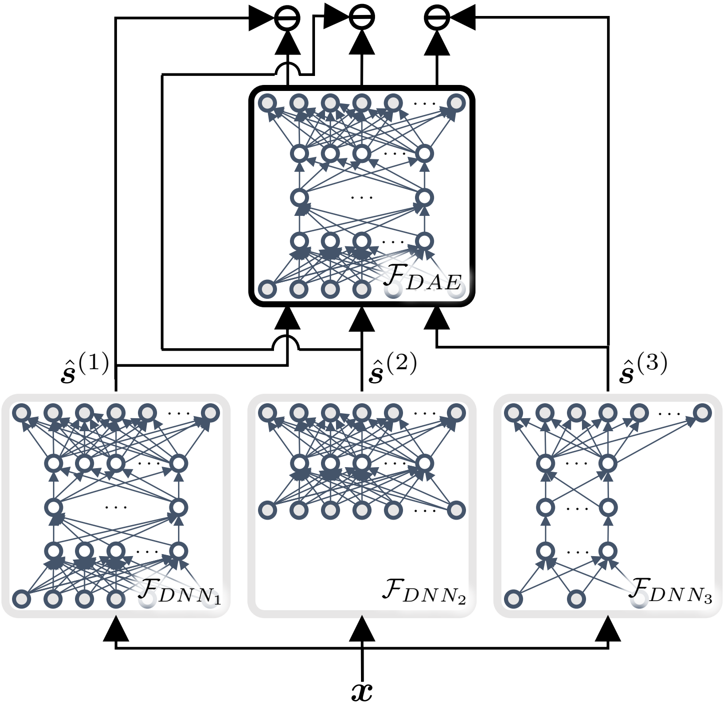
Noise Types	Mixture (Input)	Results from the Best Specialist	Results from the Worst Specialist
Bird Singing			
Typing			
Motorcycle			

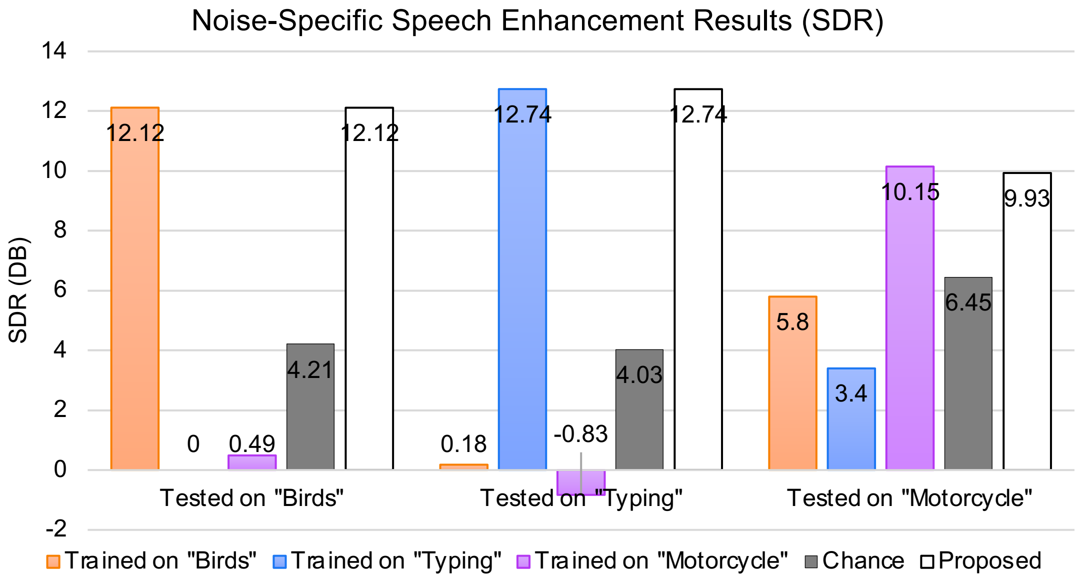
M. Kim, “Collaborative Deep Learning for Speech Enhancement: A Run-Time Model Selection Method Using Autoencoders,” ICASSP 2017

Test-Time Model Selection

- Collaborative deep learning

$$k^* = \underset{k\in\{1,\cdots,K\}}{\arg\min} \; \mathcal{E}(\hat{\boldsymbol{s}}_k || \hat{\hat{\boldsymbol{s}}}_k)$$





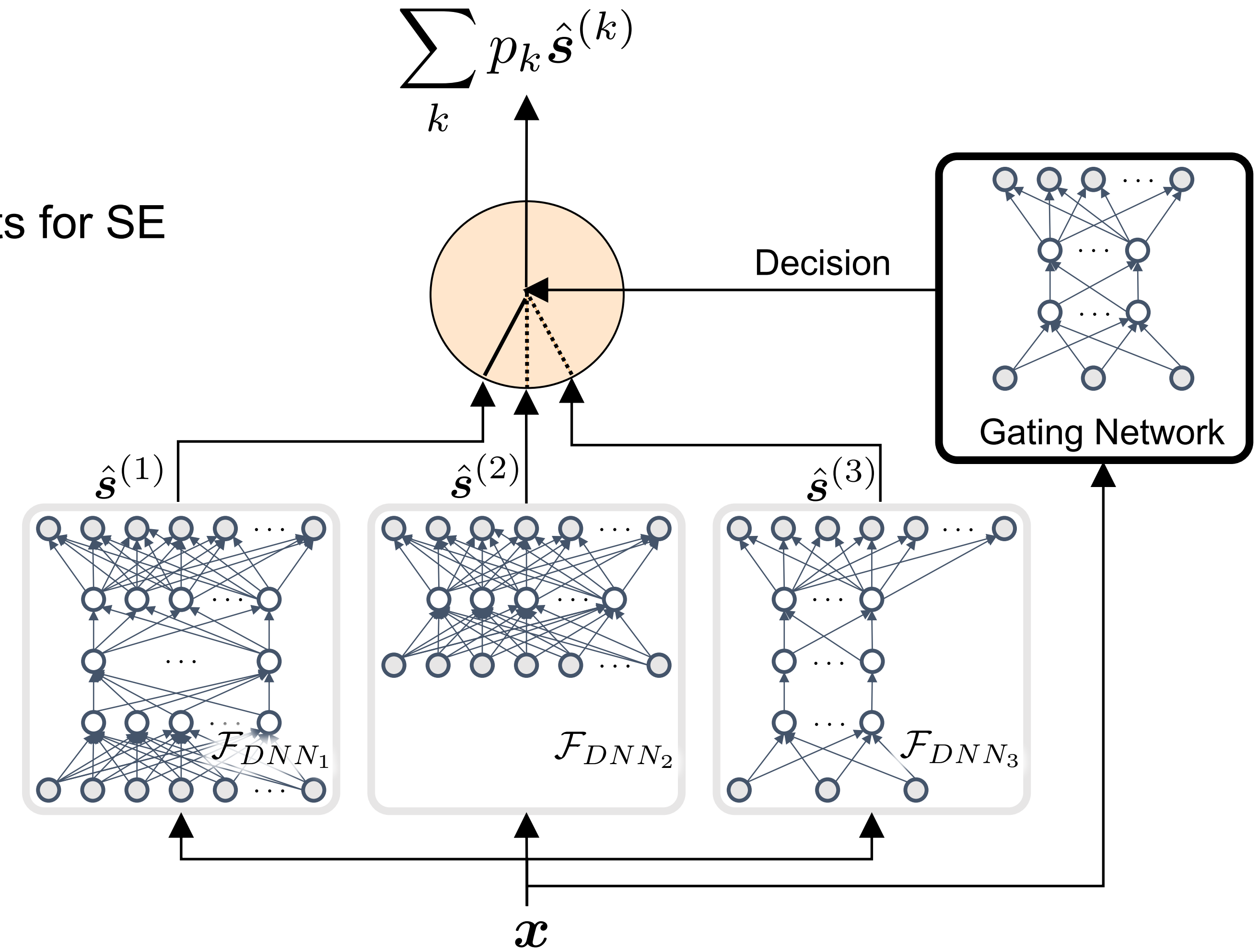
- QualityNet replaces DAE with a PESQ estimator [Zezario et al., Interspeech 2019]
- Expensive due to the potentially many candidate specialists
- The tradeoff between model complexity and performance?

M. Kim, "Collaborative Deep Learning for Speech Enhancement: A Run-Time Model Selection Method Using Autoencoders," ICASSP 2017

Test-Time Model Selection

- Mixture of local experts

- ○ Mixture of local experts
  - □ A general-purpose ensemble model
- ○ Deep recurrent mixture of local experts for SE
  [Chazan et al., WASPAA 2017]
- ○ p_k matters
  - □ Soft p_k values: an ensemble model
  - □ Could improve the performance
  - □ No structural gain from
- ○ Hard decision?
  - □ From convex combination
    to **model selection**



Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. Neural Computation, 3(1), 79-87.

Test-Time Model Selection

- Sparse Ensemble of Specialists

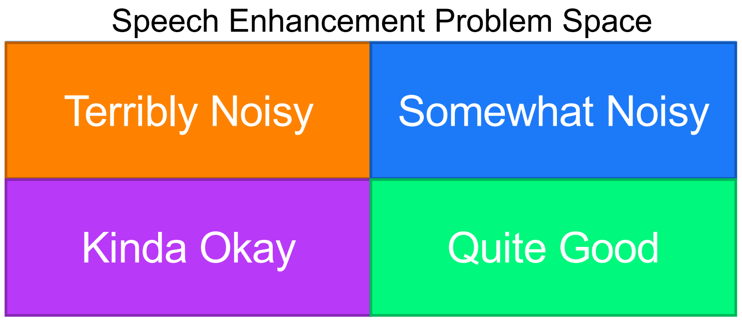
- The sparse mixture of local experts
  - Predefines small specialists
    - Based on, speaker identity, SNR levels, gender, phonemes, etc.
  - During the test time, selects the best specialists
- Fine-tuning for further adjustment
  - 1st stage: pre-train specialists based on the pre-defined subproblems
  - 2nd stage: pre-train gating network
  - 3rd stage: finetune all modules
    - Annealing $[p_1, p_2, \cdots, p_K] = \text{softmax}(\gamma \boldsymbol{z})$

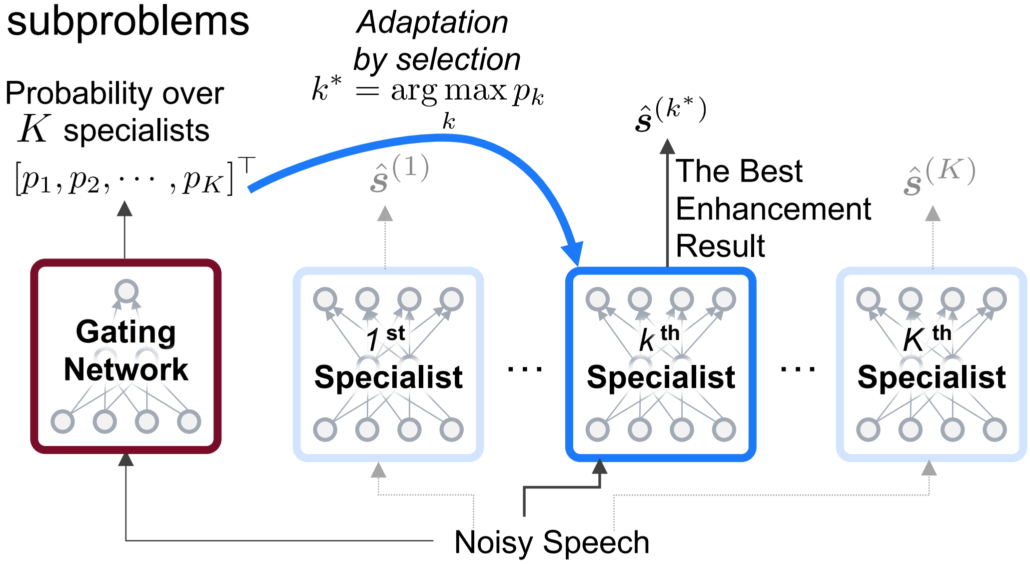
$$\hat{\boldsymbol{s}}^{(k^*)} = \lim_{\gamma \to \infty} \sum_k p_k \hat{\boldsymbol{s}}^{(k)}$$

- Complexity

$\mathcal{O}(S + C)$ $\quad$ $\underbrace{\mathcal{O}(KS + C)}_{\text{CDL}}$ vs $\mathcal{O}(G)$?

Complexity of a specialist (S) / Complexity of the gating module (C) / Complexity of a generalist (G)



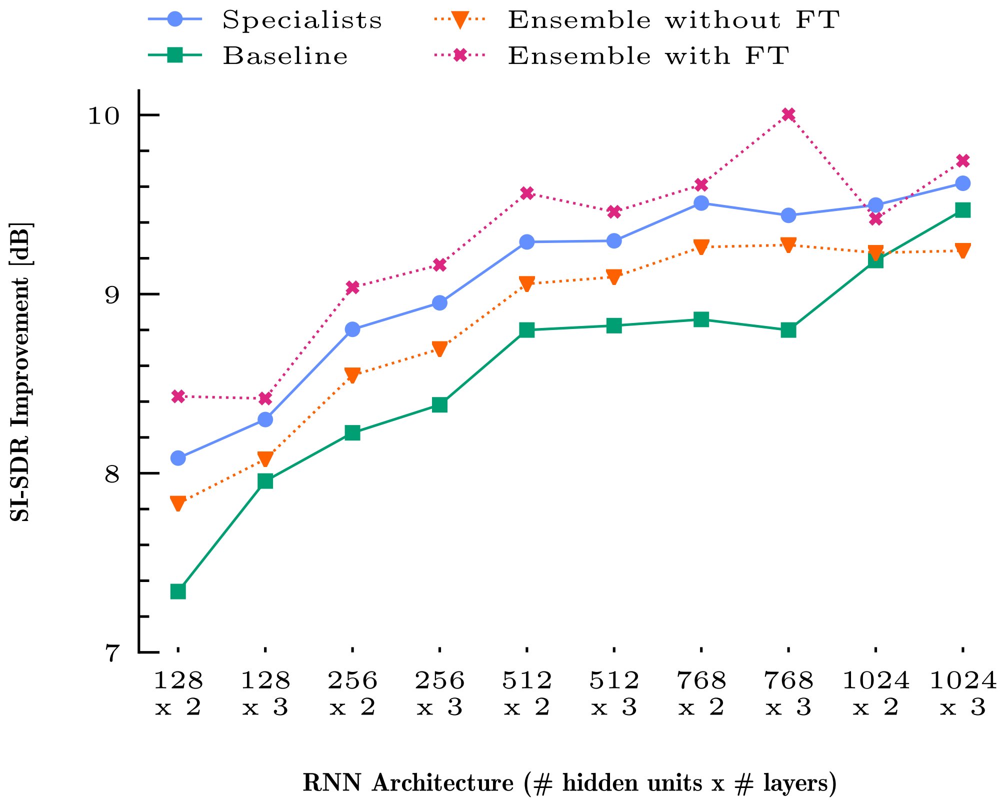


A. Sivaraman and M. Kim, "Sparse Mixture of Local Experts for Efficient Speech Enhancement," Interspeech 2020

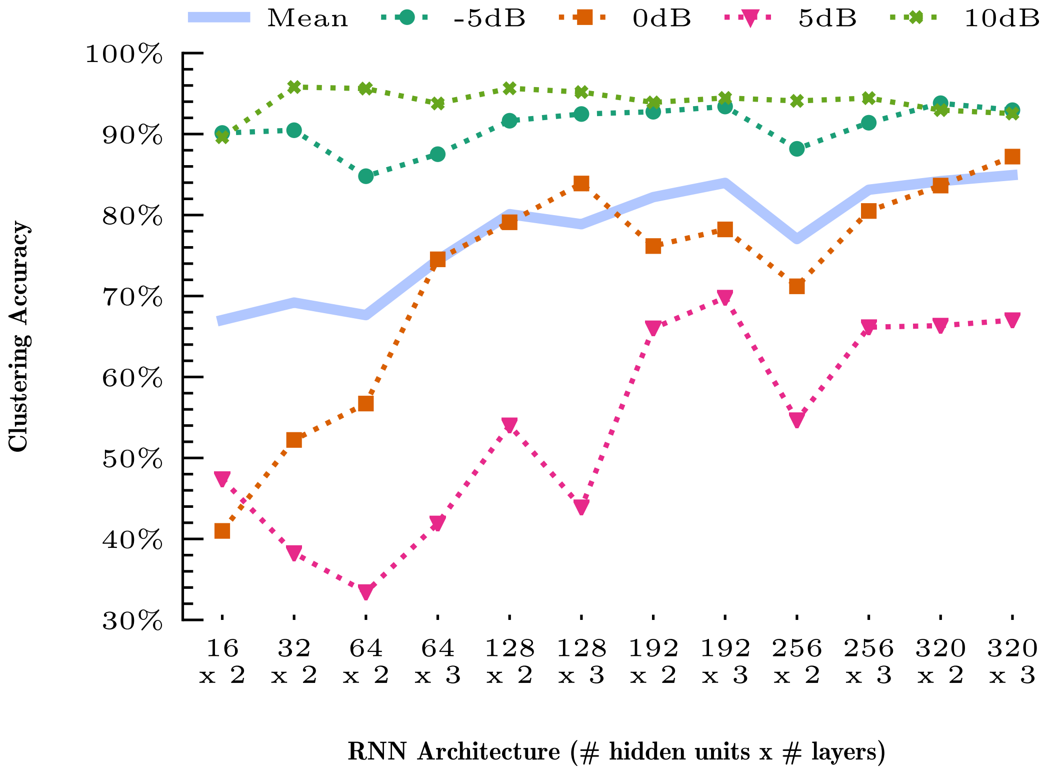
Test-Time Model Selection

- Sparse Ensemble of Specialists

- Finetuning surpasses the oracle



- Gating is not that complex

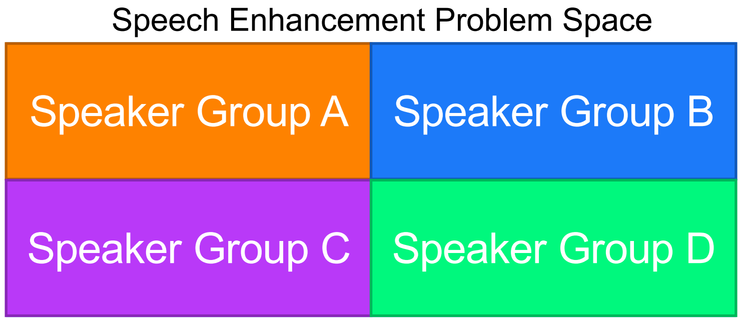


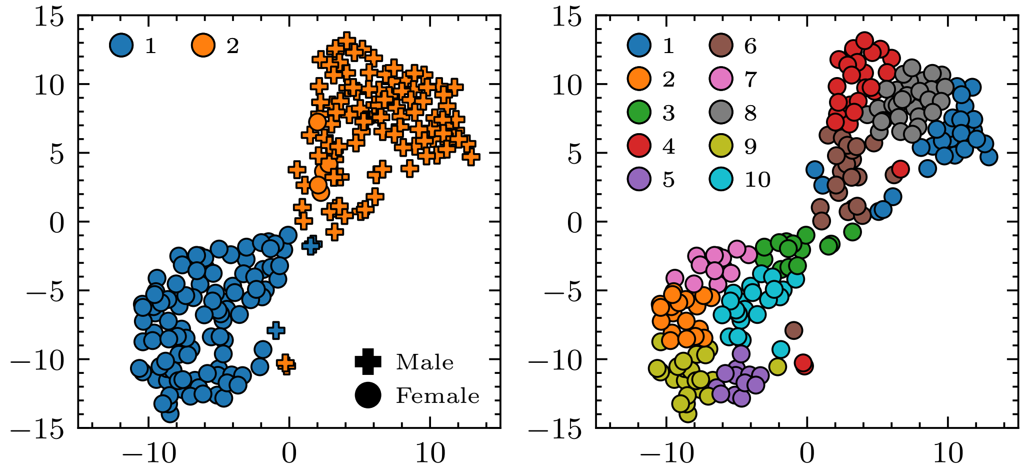
A. Sivaraman and M. Kim, "Sparse Mixture of Local Experts for Efficient Speech Enhancement," Interspeech 2020

Test-Time Model Selection

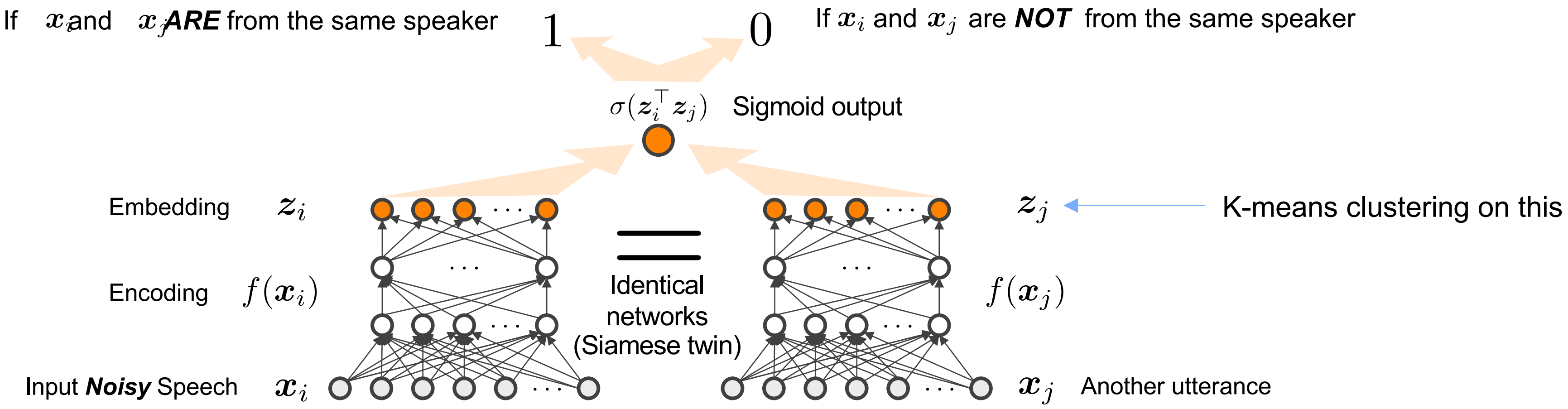
- Speaker-Specific Sparse Ensemble of Specialists

- Speaker-specific subproblem





- Noise-robust speaker embedding

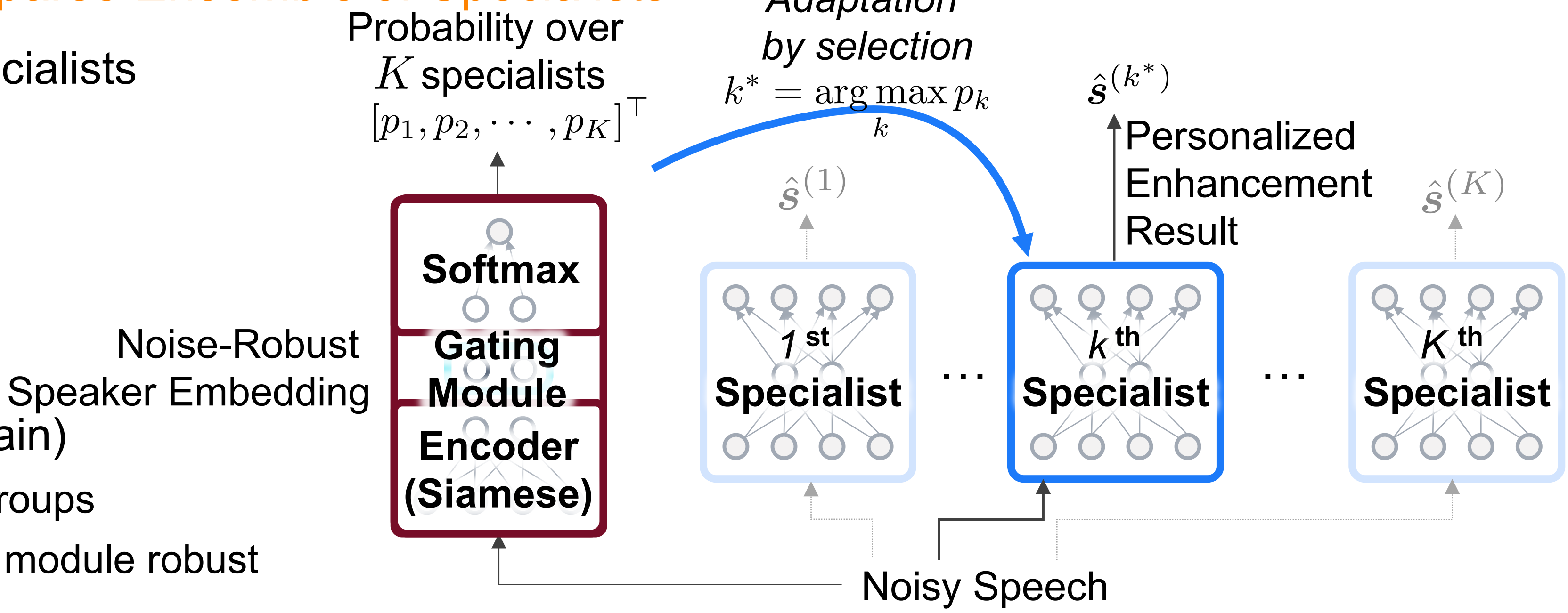


A. Sivaraman and M. Kim, “Zero-Shot Personalized Speech Enhancement Through Speaker-Informed Model Selection,” WASPAA 2021

Test-Time Model Selection

- Speaker-Specific Sparse Ensemble of Specialists

- ○ Speaker-specific specialists



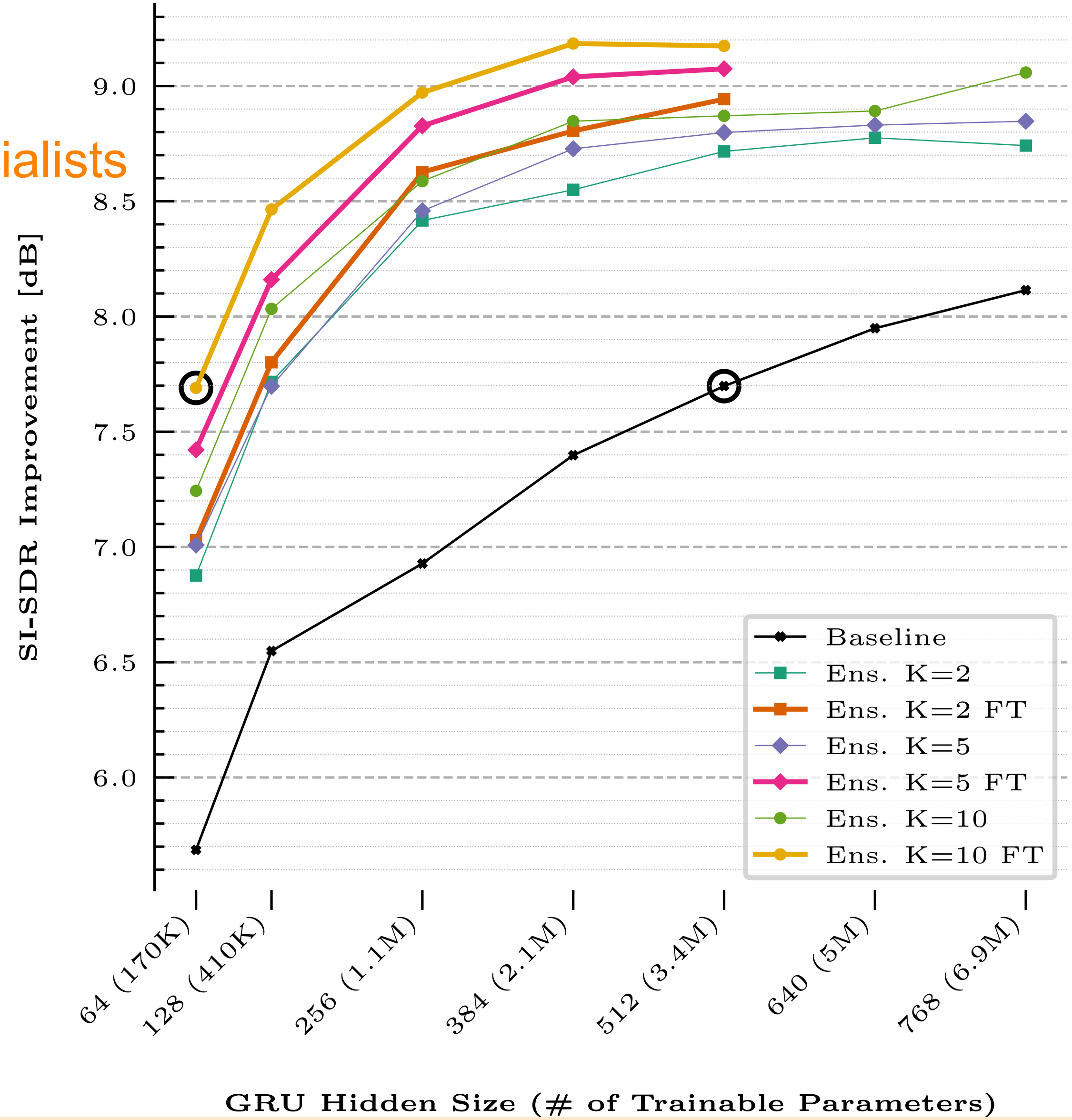
- ○ Finetuning helps (again)
  - □ Can refine speaker groups
  - □ Can make the gating module robust
- ○ Clustering on clean speech signals [Chazan et al., ICASSP 2021]
  - □ Unsupervised clustering on clean embedding vs. noise-robust embedding for speaker grouping
  - □ Reuse of the Siamese encoder
  - □ Finetuning via soft-to-hard quantization
  - □ Complexity analysis

A. Sivaraman and M. Kim, "Zero-Shot Personalized Speech Enhancement Through Speaker-Informed Model Selection," WASPAA 2021

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



A. Sivaraman and M. Kim, "Zero-Shot Personalized Speech Enhancement Through Speaker-Informed Model Selection," WASPAA 2021

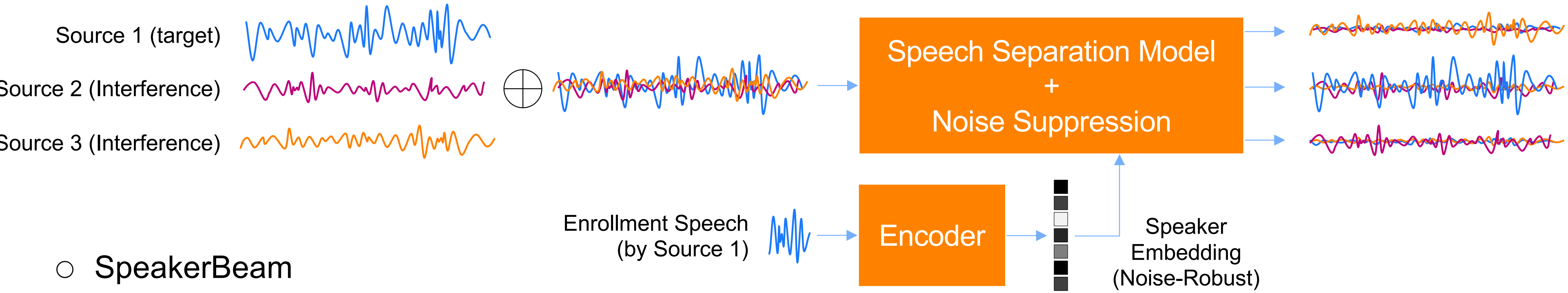
Few-Shot PSE

Target Speaker Extraction as PSE
Self-Supervised Learning
Data Purification
Contrastive Mixtures

Target Speaker Extraction

- Another view of PSE

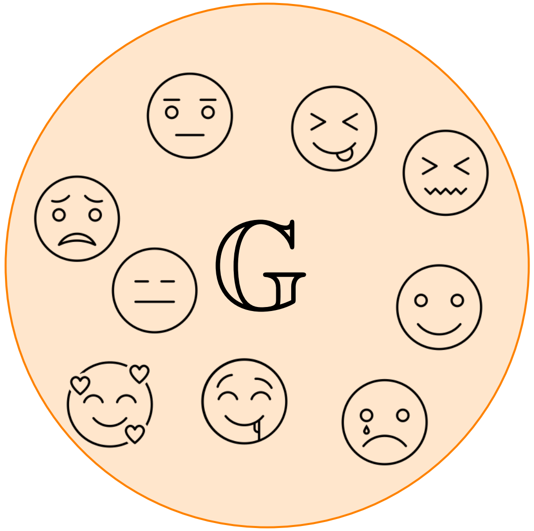
- From speech separation to target speaker extraction

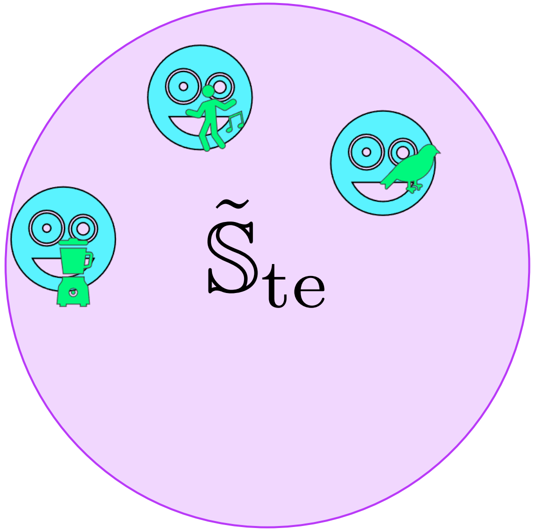


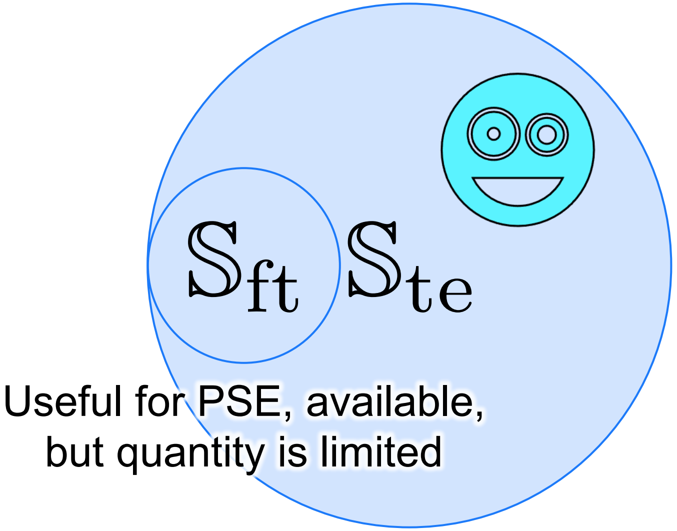
- SpeakerBeam
  - Multichannel [Žmolíková et al., Interspeech 2017; IEEE JSTSP 2019]
- VoiceFilter [Q. Wang et al., “VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking,” Interspeech 2019]
- Deep Noise Suppression Challenge
  [S. E. Eskimez et al., "Personalized speech enhancement: new models and comprehensive evaluation," ICASSP 2022]
  - https://www.microsoft.com/en-us/research/academic-program/deep-noise-suppression-challenge-icassp-2021/
  - https://www.microsoft.com/en-us/research/academic-program/deep-noise-suppression-challenge-icassp-2022/
- It’s a more challenging problem setup
  - Less consideration about the resource and data efficiency due to the SS nature

Self-Supervised Learning

- The dataset formulation for PSE







Anonymous Clean Utterances	Test-Time User's Noisy Utterances (***Premixture***)	Test-Time User's Clean Utterances
Available, but not personalized	More available, but not clean enough	Useful for PSE, but not available

We never know what $\mathbb{S}_{\text{te}}$ is

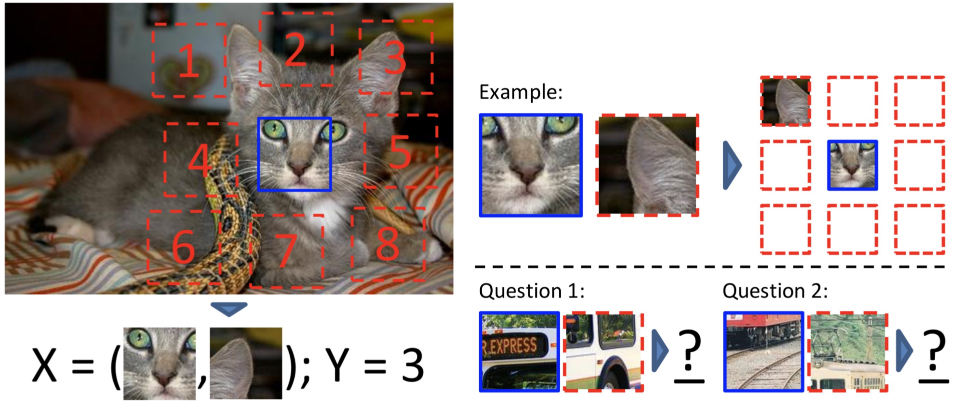
They are noisy, but specifically about the test environment

Let's make it more useful via self-supervised learning!

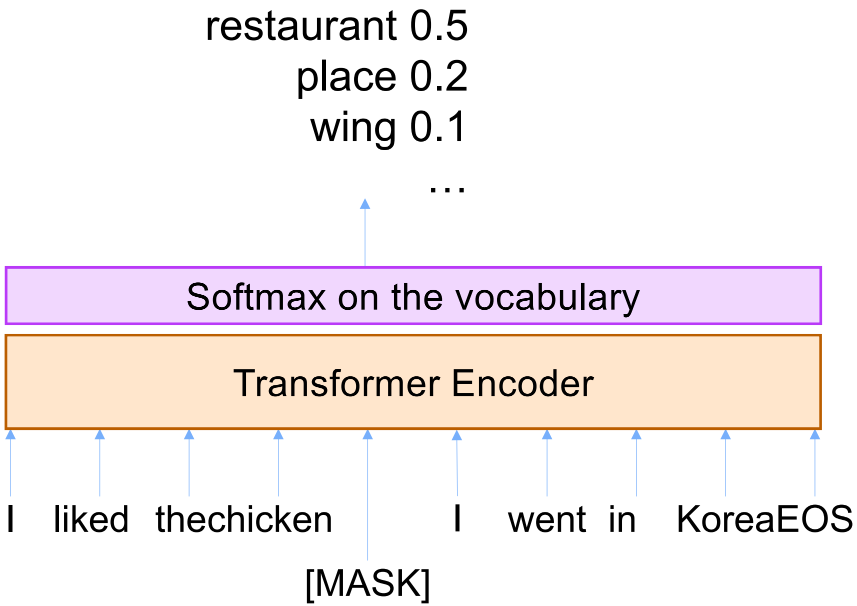
Self-Supervised Learning

- In CV and NLP

- What if we do have *some* clean speech from the test user?
  - But just a little
- Self-supervised feature learning
  - Learns discriminative features in an unsupervised way
    - Asking the network to solve jigsaw puzzle



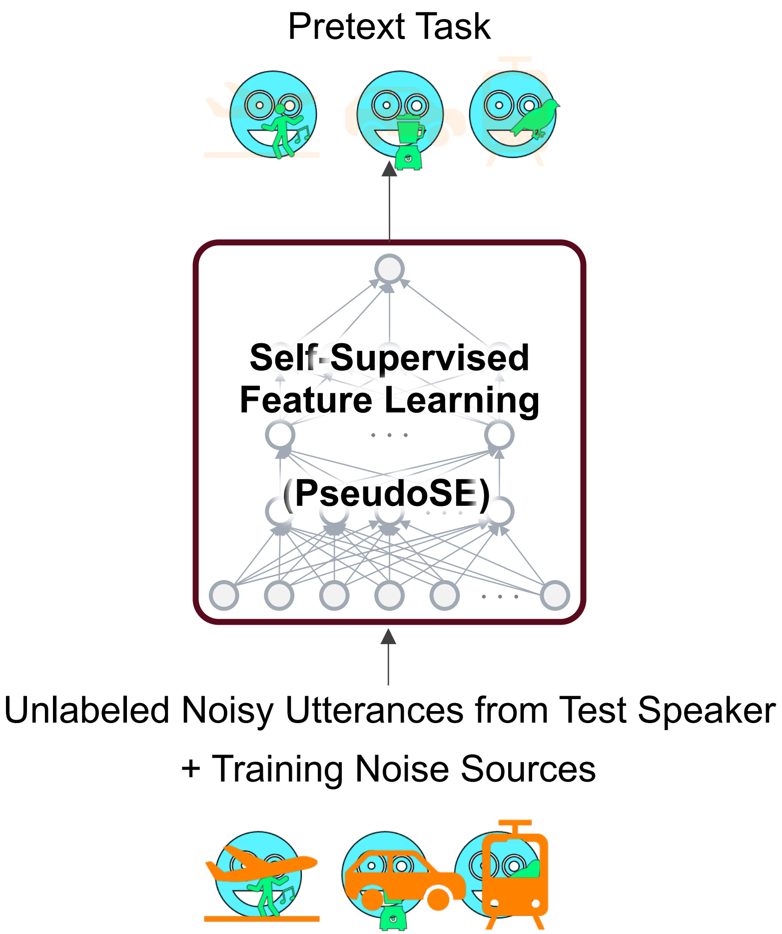
[Doersch et al., ICCV 2015]

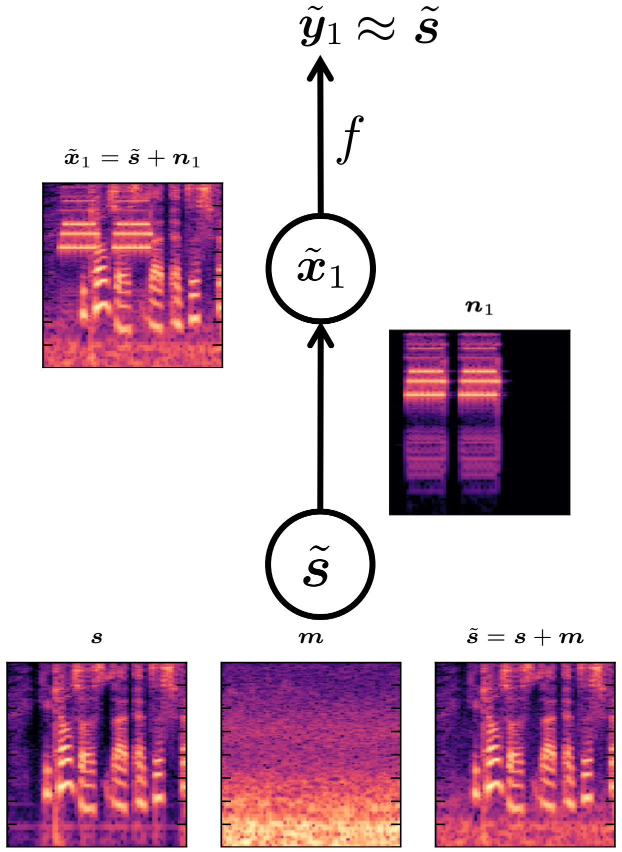


Carl Doersch, Abhinav Gupta, Alexei A. Efros, "Unsupervised Visual Representation Learning by Context Prediction," ICCV 2015
J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL-HLT 2019

Self-Supervised Learning

- Pseudo speech enhancement



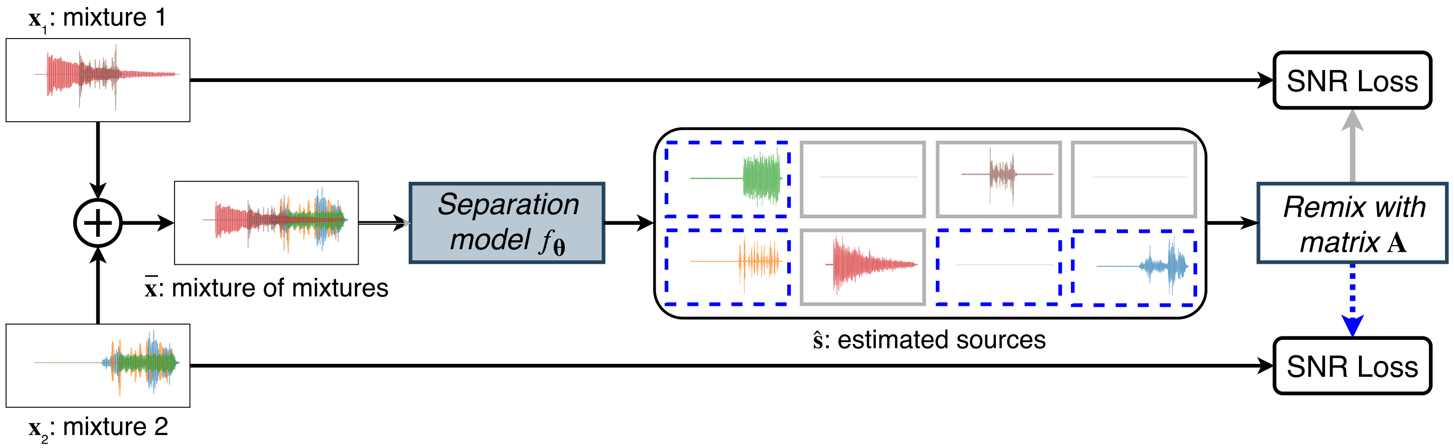


A. Sivaraman and M. Kim, "Self-Supervised Learning from Contrastive Mixtures for Personalized Speech Enhancement," NeurIPS 2020 SSL for SAP Workshop

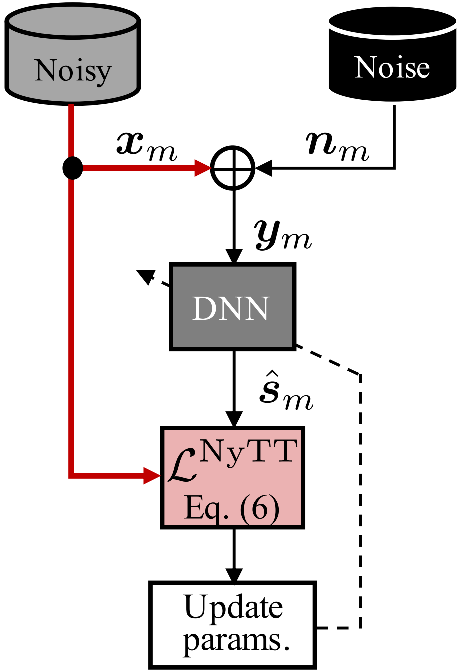
Self-Supervised Learning

- In the source separation domain

- Mixture Invariant Training
  [Wisdom et al., NeurIPS 2020]
  - Not tailored for SE



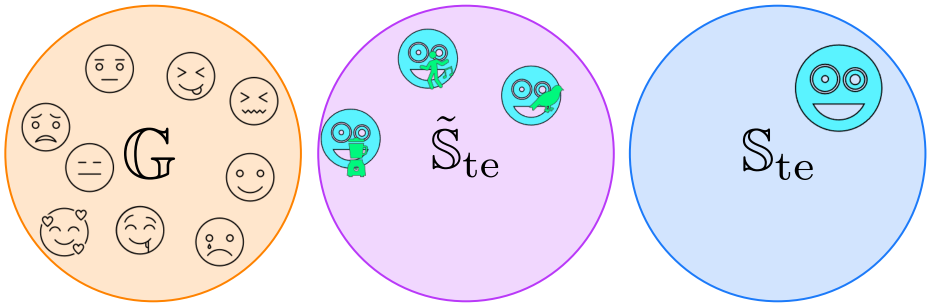
- Noisy-Target Training
  [Fujimura et al., EUSIPCO 2021]



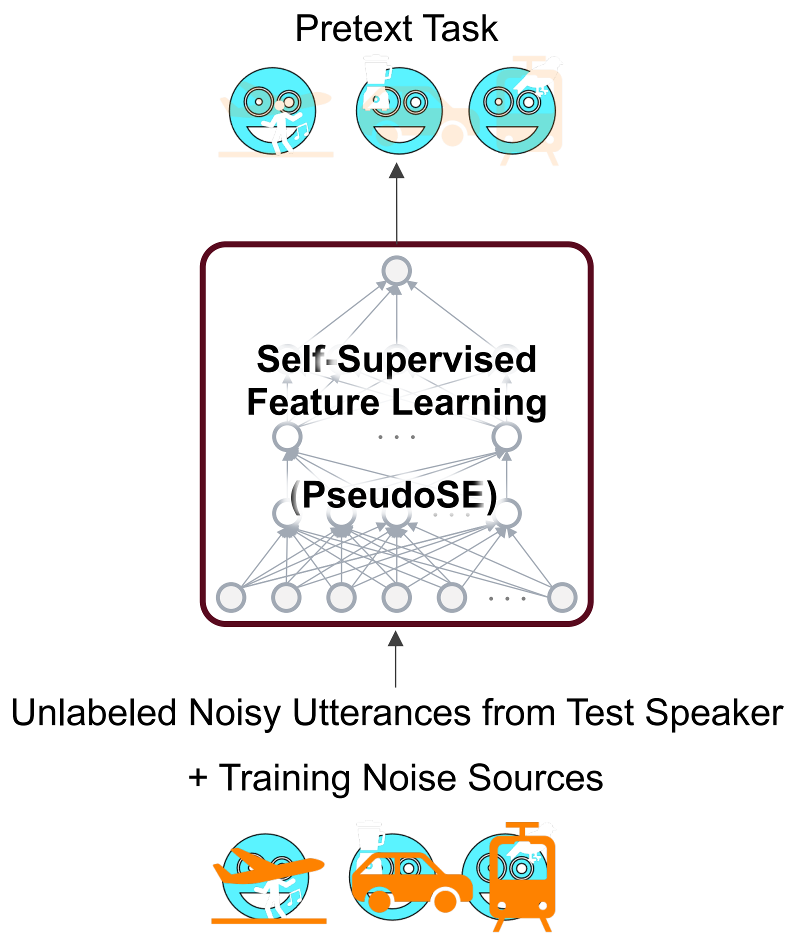
- Noisy2Noisy [Alamdari et al., Applied Acoustics 2021]
  - Requires two noisy signals that share the *same* speech source

Data Purification

- Pseudo SE is not good enough



- If $\tilde{\mathbb{S}}_{te}$ are clean enough, $\tilde{\mathbb{S}}_{te} \approx \mathbb{S}_{te}$
  - But they are not
- Data purification?
  - May work if premixture noise is sparse

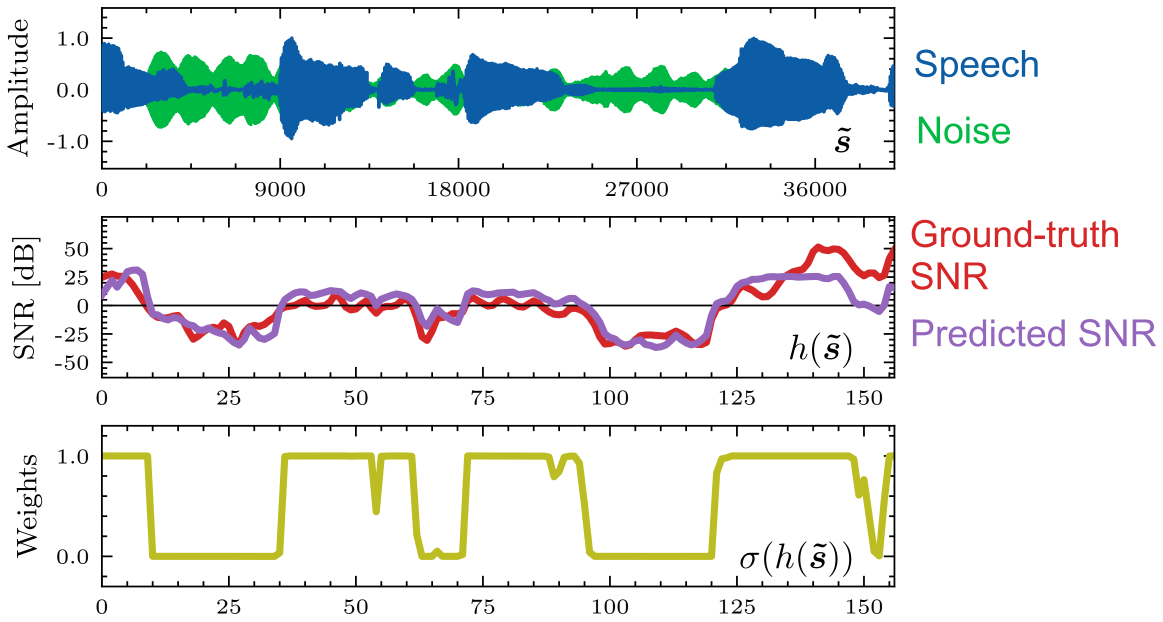


A. Sivaraman, S. Kim and M. Kim, "Personalized Speech Enhancement through Self-Supervised Data Augmentation and Purification," Interspeech 2021

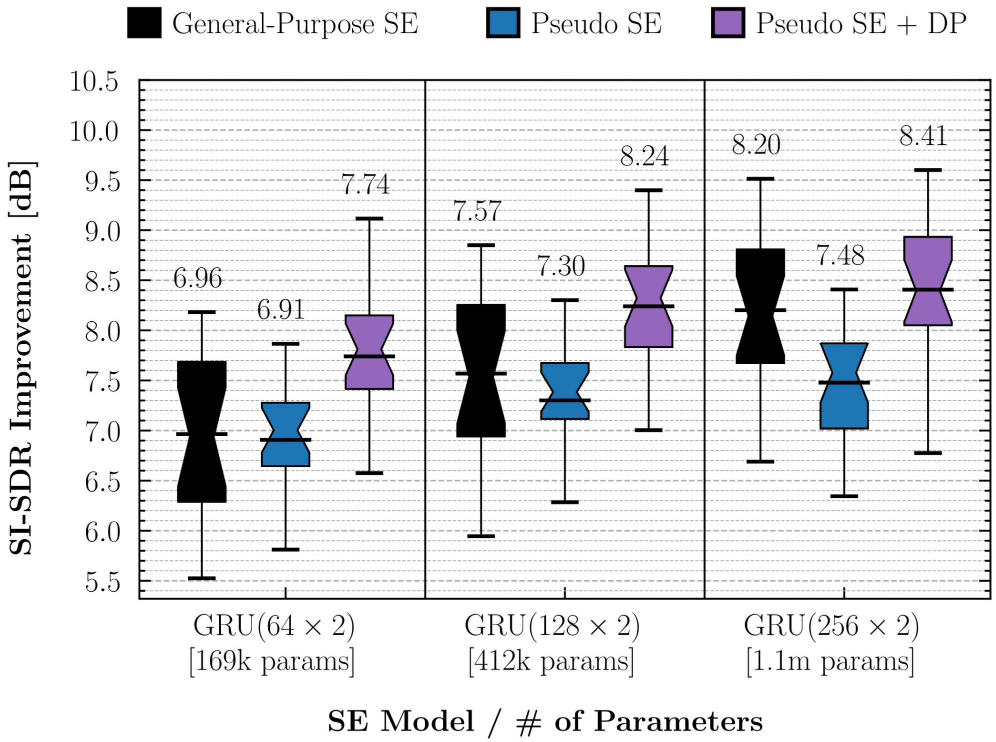
Data Purification

- For PSE?

- The algorithm (in English)
  - PseudoSE, but only on clean-ish target
  - Weighted loss
  - Then, we can pretend like $\tilde{\mathbb{S}}_{\text{te}} = \mathbb{S}_{\text{te}}$



- On various speaker-specific SE tasks

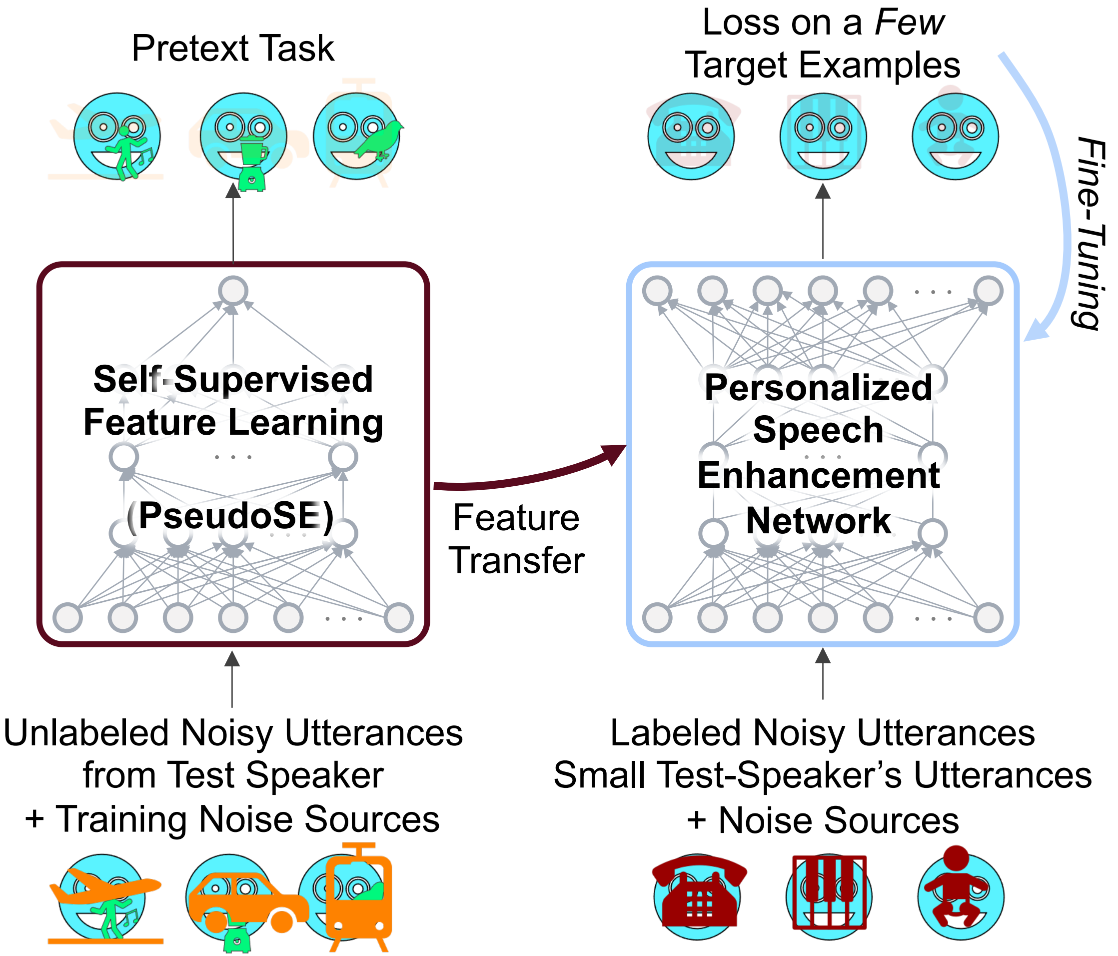


A. Sivaraman, S. Kim and M. Kim, “Personalized Speech Enhancement through Self-Supervised Data Augmentation and Purification,” Interspeech 2021

Contrastive Mixtures

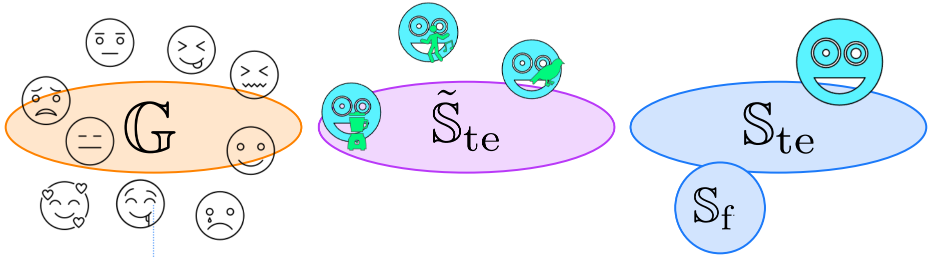
- What are we missing?

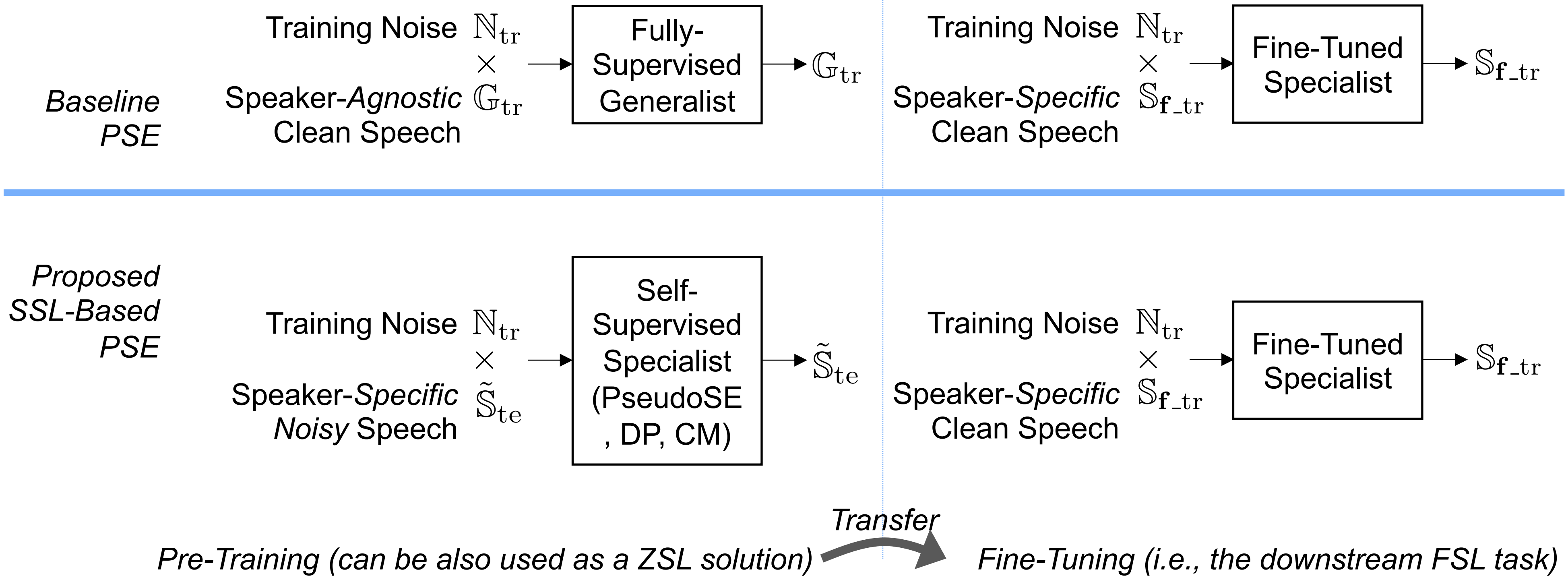
- Transfer learning!
  - Or fine-tuning
  - Or personalization
  - Or few-shot learning



Contrastive Mixtures

- SSL + FSL overview

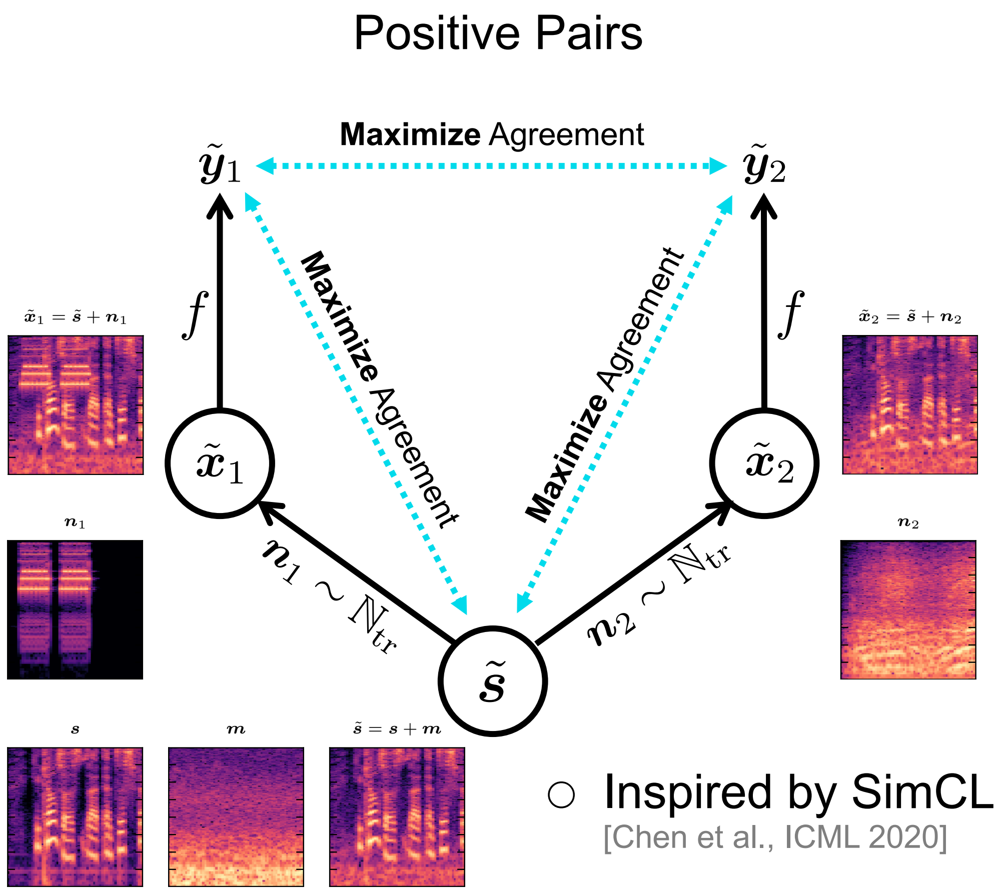




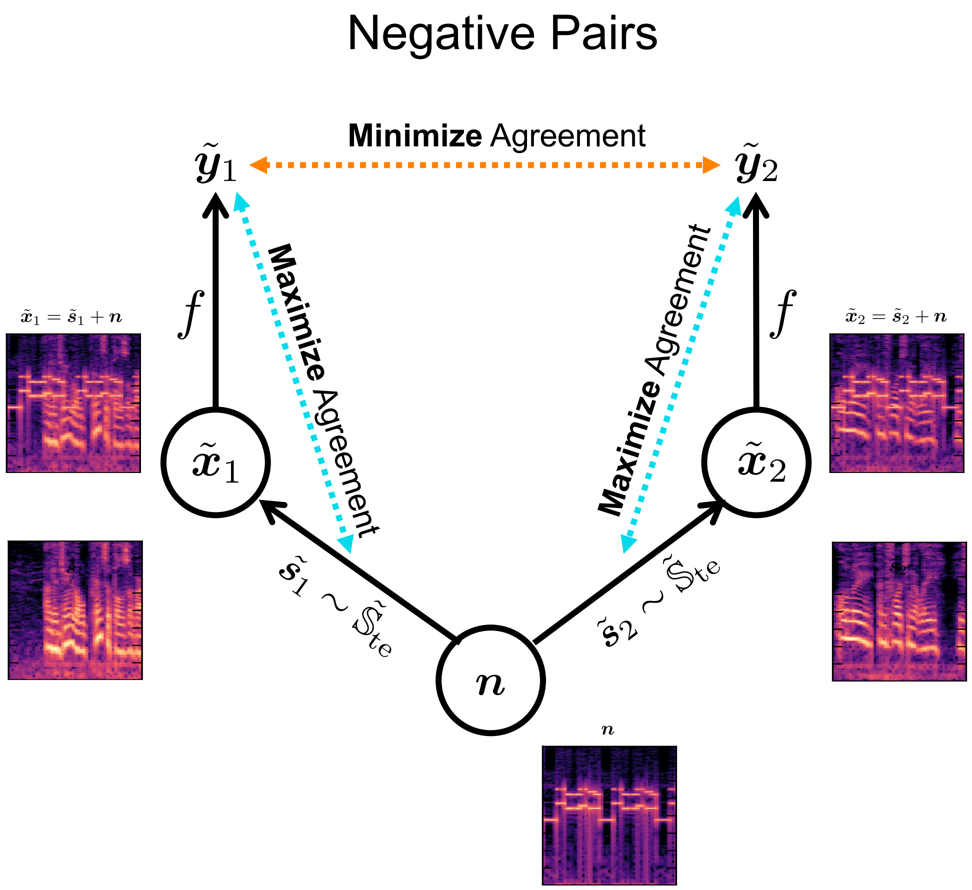
A. Sivaraman and M. Kim, "Efficient Personalized Speech Enhancement through Self-Supervised Learning," IEEE JSTSP 2022

Contrastive Mixtures

- Contrastive mixtures



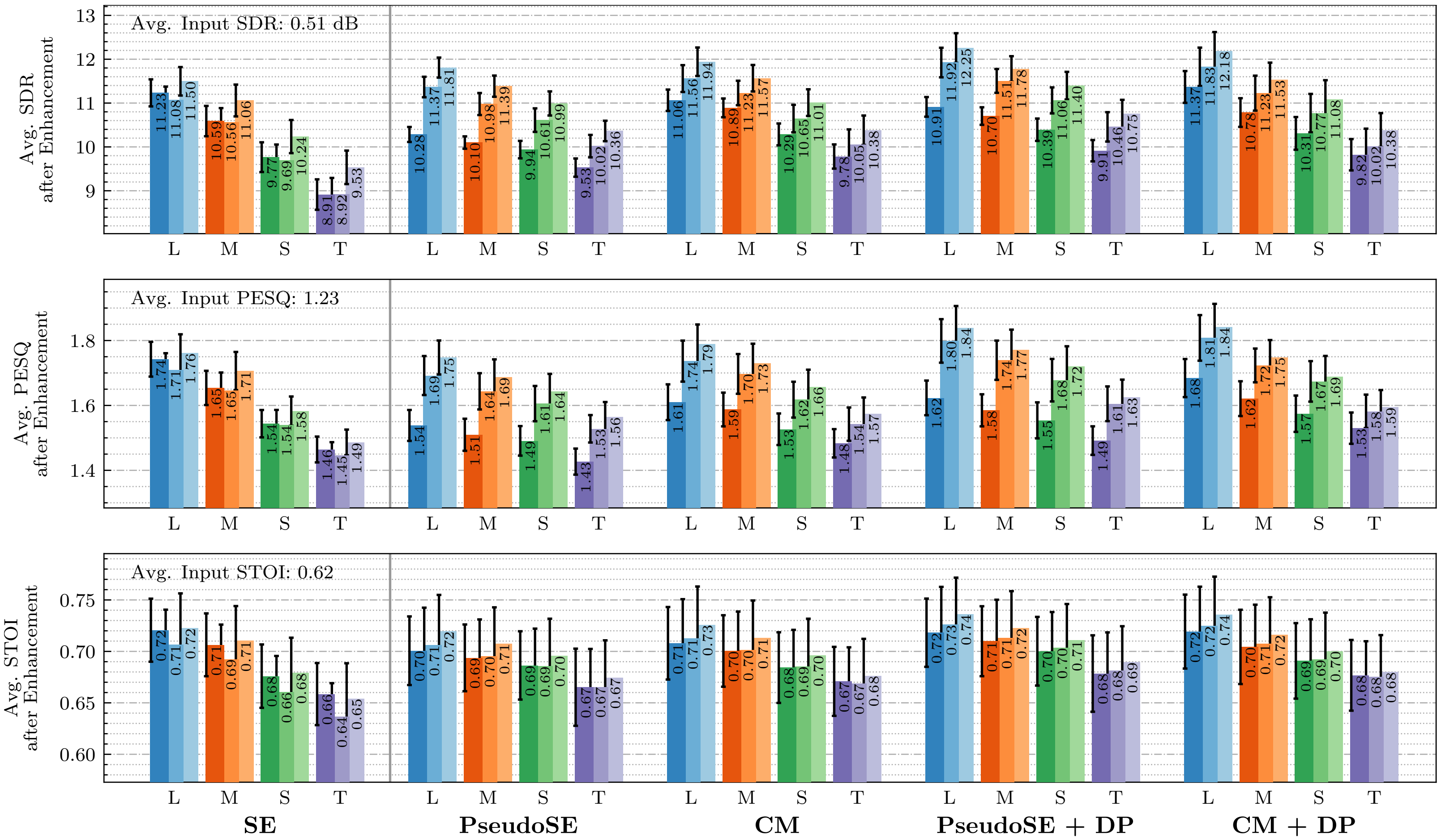
- Inspired by SimCLR
  [Chen et al., ICML 2020]

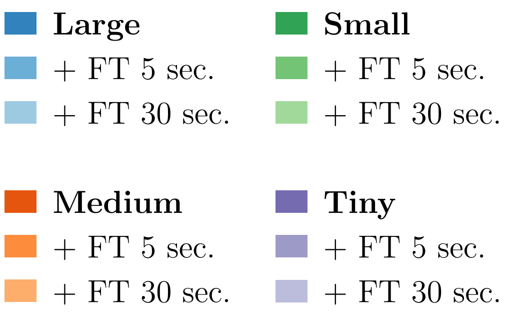


A. Sivaraman and M. Kim, "Efficient Personalized Speech Enhancement through Self-Supervised Learning," IEEE JSTSP 2022

Few-Shot Learning for PSE

- Results





Size	**Configuration**	**Params**	**MACs**
Large	$B_c = 64,\ H_c = 256$	1.0 M	8.4 G
Medium	$B_c = 32,\ H_c = 128$	437.8 k	3.5 G
Small	$B_c = 16,\ H_c = 64$	224.1 k	1.8 G
Tiny	$B_c = 8,\ H_c = 32$	138.8 k	1.1 G

A. Sivaraman and M. Kim, “Efficient Personalized Speech Enhancement through Self-Supervised Learning,” IEEE JSTSP 2022

Conclusion

- Personalization is meaningful
  - \+ Improves performance
  - \+ Reduces model complexity
  - \+ Reduces model’s bias (caused by data imbalance)
  - \- Difficult to acquire personal labeled data
  - \- Can breach the privacy
- Zero-shot learning
  - Doesn’t require clean speech target
  - Tricky to train due to the lack of data, but there are ways
    - Run-time model adaptation via knowledge distillation
    - Sub-grouping the problem into smaller sub-problems
- Few-shot learning
  - SSL helps few-shot learning
    - Pseudo SE (noisy target training)
    - Data purification
    - Contrastive mixtures

Thank You!

Minje Kim

Email: minje@Indiana.edu

https://minjekim.com/research-projects/pse/

(slides, source codes, and demos are available)

References

- 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.
- S. Han et al., “Learning both Weights and Connections for Efficient Neural Networks,” NIPS 2015.
- A. Howard et al., “Searching for MobileNetV3” ICCV 2019
- E. Strubell et al., “Energy and Policy Considerations for Deep Learning in NLP,” arXiv:1906.02243
- M. Kim and P. Smaragdis, "Bitwise Neural Networks for Efficient Single-Channel Source Separation," ICASSP 2018
- S. Kim, M. Maity, and M. Kim, “Incremental Binarization On Recurrent Neural Networks for Single-Channel Source Separation,” ICASSP 2019
- K. Tan and D. Wang, "Compressing Deep Neural Networks for Efficient Speech Enhancement," ICASSP, 2021
- Y Luo, C Han, N Mesgarani, “Ultra-lightweight speech separation via group communication,” ICASSP, 2021
- J. Buolamwini and T. Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification," Conference on fairness, accountability and transparency. 2018.
- Z. Duan et al., “Speech Enhancement by Online Non-negative Spectrogram Decomposition in Non-stationary Noise Environments,” Interspeech 2012
- M. Kim and P. Smaragdis, “Adaptive Denoising Autoencoders: A Fine-tuning Scheme to Learn from Test Mixtures,” LVA/ICA 2015
- D. Williamson et al., “Reconstruction techniques for improving the perceptual quality of binary masked speech,” JASA 2014
- G. Hinton et al., “Distilling the Knowledge in a Neural Network,” arXiv:1503.02531

References

- S. Kim and M. Kim, “Test-Time Adaptation Toward Personalized Speech Enhancement: Zero-Shot Learning With Knowledge Distillation,” WASPAA 2021
- P. Smaragdis et al. “A sparse non-parametric approach for single channel separation of known sounds,” NIPS 2009
- M. Kim and P. Smaragdis, “Manifold Preserving Hierarchical Topic Models for Quantization and Approximation,” ICML 2013
- M. Kim and P. Smaragdis, “Mixtures of Local Dictionaries for Unsupervised Speech Enhancement,” IEEE SPL, 2015
- D. L. Sun et al., “Universal speech models for speaker independent single channel source separation,” ICASSP 2013
- M. Kim, “Collaborative Deep Learning for Speech Enhancement: A Run-Time Model Selection Method Using Autoencoders,” ICASSP 2017
- R. Zezario et al., “Specialized Speech Enhancement Model Selection Based on Learned Non-Intrusive Quality Assessment Metric,” Interspeech 2019
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. Neural Computation, 3(1), 79-87.
- S. E. Chazan, J. Goldberger and S. Gannot, "Deep recurrent mixture of experts for speech enhancement," WASPAA, 2017
- A. Sivaraman and M. Kim, “Sparse Mixture of Local Experts for Efficient Speech Enhancement,” Interspeech 2020
- A. Sivaraman and M. Kim, “Zero-Shot Personalized Speech Enhancement Through Speaker-Informed Model Selection,” WASPAA 2021
- S. E. Chazan, J. Goldberger and S. Gannot, "Speech Enhancement with Mixture of Deep Experts with Clean Clustering Pre-Training," ICASSP 2021
- Žmolíková, K. et al., “Speaker-Aware Neural Network Based Beamformer for Speaker Extraction in Speech Mixtures,” Interspeech 2017
- K. Žmolíková et al., "SpeakerBeam: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures," IEEE JSTSP 2019

References

- S. E. Eskimez et al., "Personalized speech enhancement: new models and comprehensive evaluation," ICASSP 2022
- Carl Doersch, Abhinav Gupta, Alexei A. Efros, "Unsupervised Visual Representation Learning by Context Prediction," ICCV 2015
- J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL-HLT 2019
- A. Sivaraman and M. Kim, "Self-Supervised Learning from Contrastive Mixtures for Personalized Speech Enhancement," NeurIPS 2020 SSL for SAP Workshop
- S. Wisdom et al., "Unsupervised sound separation using mixture invariant training," NeurIPS 2020
- T. Fujimura et al., "Noisy-target Training: A Training Strategy for DNN-based Speech Enhancement without Clean Speech," EUSIPCO 2021
- Alamdari et al., "Improving deep speech denoising by Noisy2Noisy signal mapping," Applied Acoustics 2021
- A. Sivaraman, S. Kim and M. Kim, "Personalized Speech Enhancement through Self-Supervised Data Augmentation and Purification," Interspeech 2021
- A. Sivaraman and M. Kim, "Efficient Personalized Speech Enhancement through Self-Supervised Learning," IEEE Journal of Selected Topics in Signal Processing, 2022
- Yi Li et al., "Self-Supervised Learning based Monaural Speech Enhancement with Complex-Cycle-Consistent," IEEE JSTSP, 2022
- K.-H. Hung et al., "Boosting Self-Supervised Embeddings for Speech Enhancement," IEEE JSTSP, 2022
- E. Tzinis, "RemixIT: Continual self-training of speech enhancement models via bootstrapped remixing," IEEE JSTSP, 2022
- Yi Li et al., "Feature Learning and Ensemble Pre-Tasks Based Self-Supervised Speech Denoising and Dereverberation," IEEE JSTSP, 2022