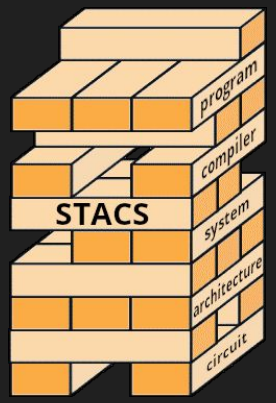
The Spatial Effect of the Pinna for Neuromorphic Speech Denoising

Neuro-Inspired Computational Elements
March 2025

Ranganath (Bujji) Selagamsetty, Joshua San Miguel, Mikko Lipasti



Computer Sciences
SCHOOL OF COMPUTER, DATA & INFORMATION SCIENCES
UNIVERSITY OF WISCONSIN–MADISON

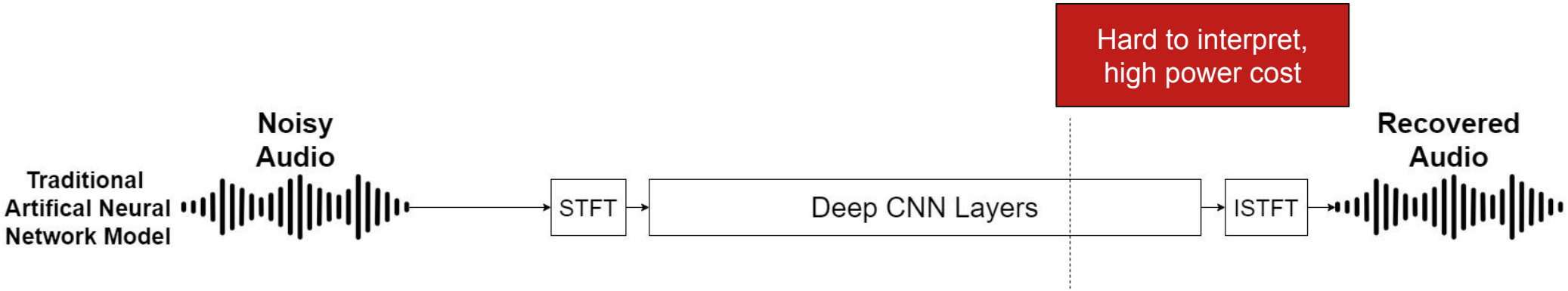
Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Outline

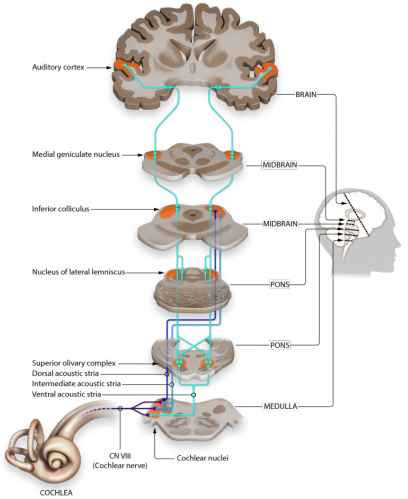
- Motivation
- Methodology
- Evaluation
- Conclusion & Future directions

Motivation: Biological Inspiration

Motivation: Biological Inspiration



Motivation: Biological Inspiration



Hard to interpret, high power cost

Noisy Audio

Traditional Artifical Neural Network Model

STFT

Deep CNN Layers

ISTFT

Recovered Audio

Biologial Plausibility

Auditory Pathway of the Human Brain

CN VIII

CN

SOC

IC

MGN

Speech Reception Pathway

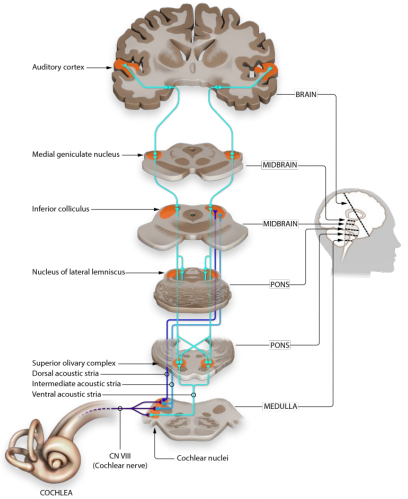
Speech Production Pathway

Low power consumption, high performance

https://pressbooks.umn.edu/sensationandperception/chapter/auditory-pathways-to-the-brain-draft/

Motivation: Biological Inspiration

- Shallow SNNs get us part way there



Hard to interpret, high power cost

Noisy Audio

Recovered Audio

Traditional Artifical Neural Network Model

STFT

Deep CNN Layers

ISTFT

Spiking Neural Network Model

STFT

Shallow SNN Layers

ISTFT

Biologial Plausibility

Auditory Pathway of the Human Brain

CN VIII

CN

SOC

IC

MGN

Speech Reception Pathway

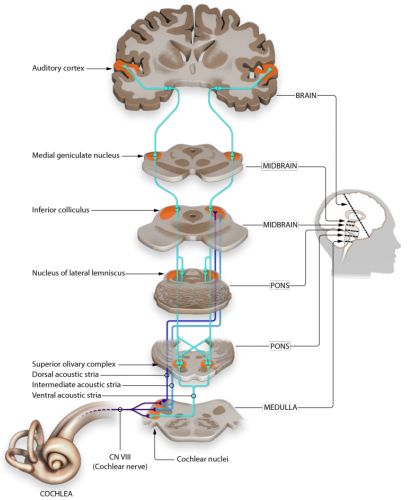
Speech Production Pathway

Low power consumption, high performance

https://pressbooks.umn.edu/sensationandperception/chapter/auditory-pathways-to-the-brain-draft/

Motivation: Biological Inspiration

- Shallow SNNs get us part way there



Hard to interpret, high power cost

Noisy Audio

Recovered Audio

Traditional Artifical Neural Network Model

STFT

Deep CNN Layers

ISTFT

Spiking Neural Network Model

STFT

Shallow SNN Layers

ISTFT

Our work

Pinna Enhanced Spiking Neural Network

HRIR

STFT

Shallow SNN Layers

ISTFT

Auditory Pathway of the Human Brain

CN VIII

CN

SOC

IC

MGN

Biologial Plausibility

Speech Reception Pathway

Speech Production Pathway

Low power consumption, high performance

https://pressbooks.umn.edu/sensationandperception/chapter/auditory-pathways-to-the-brain-draft/

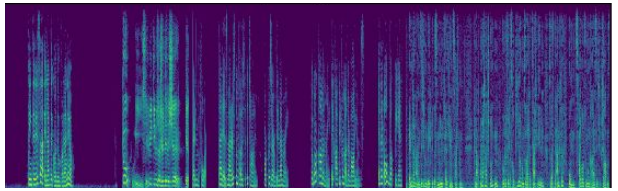
Methodology

Methodology: Datasets

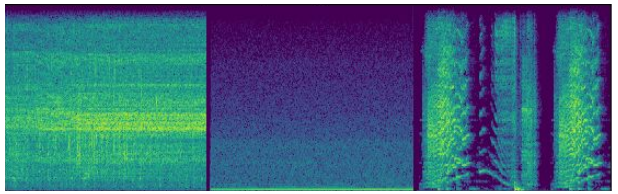
- CIPIC HRTF Database for HRIRs (45 subjects, 1250 orientations)
- Noisy sample synthesized from Microsoft DNS corpus
  - Speech recordings from speakers from multiple languages
  - Noise recording include both with biological origin and abiogenic
  - 30 second clip
- Consists of 500 hours (60K samples) in training and validation sets

Methodology: Data Augmentation with Pinna Cues

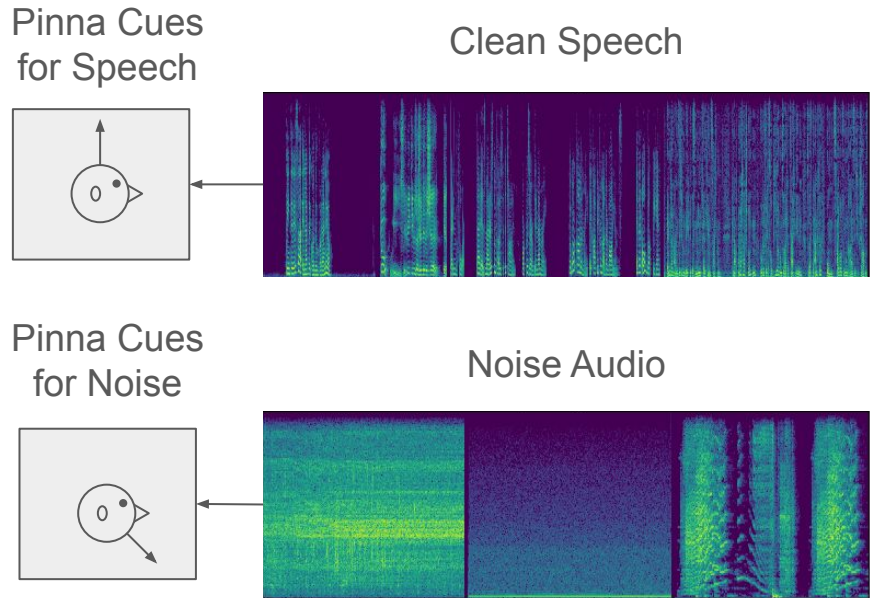
Clean Speech



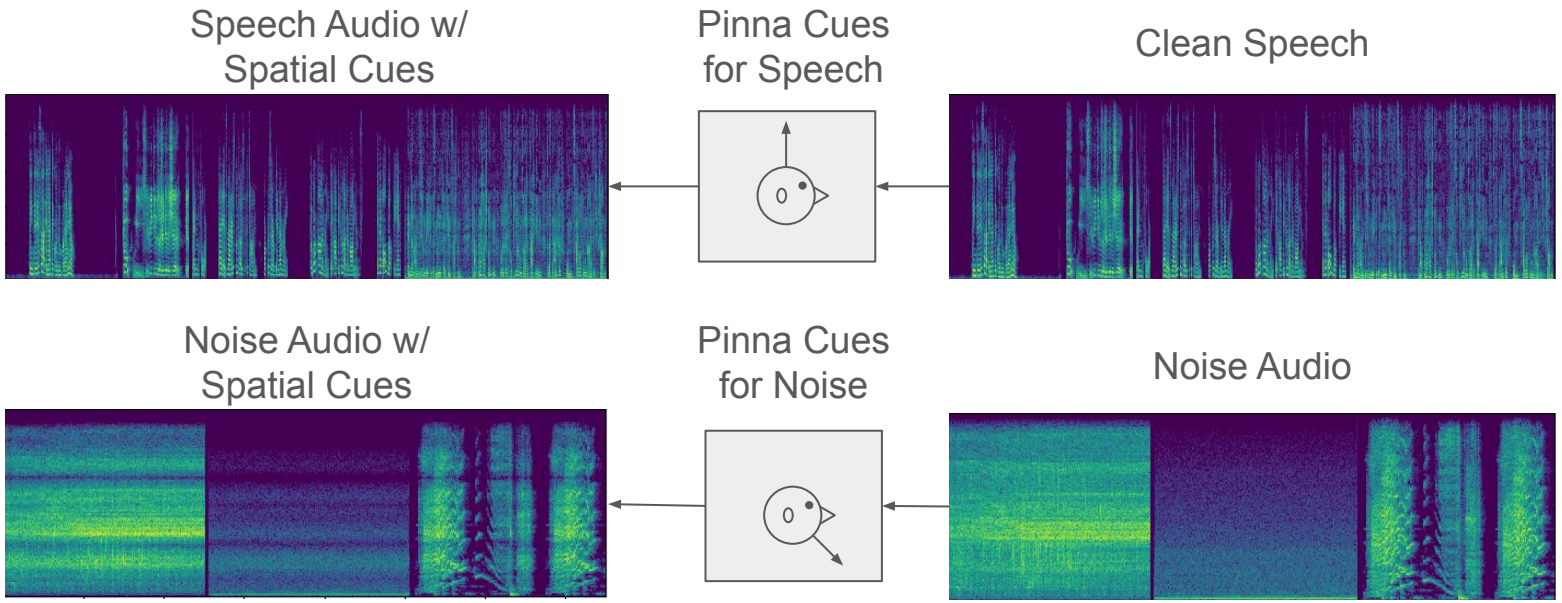
Noise Audio



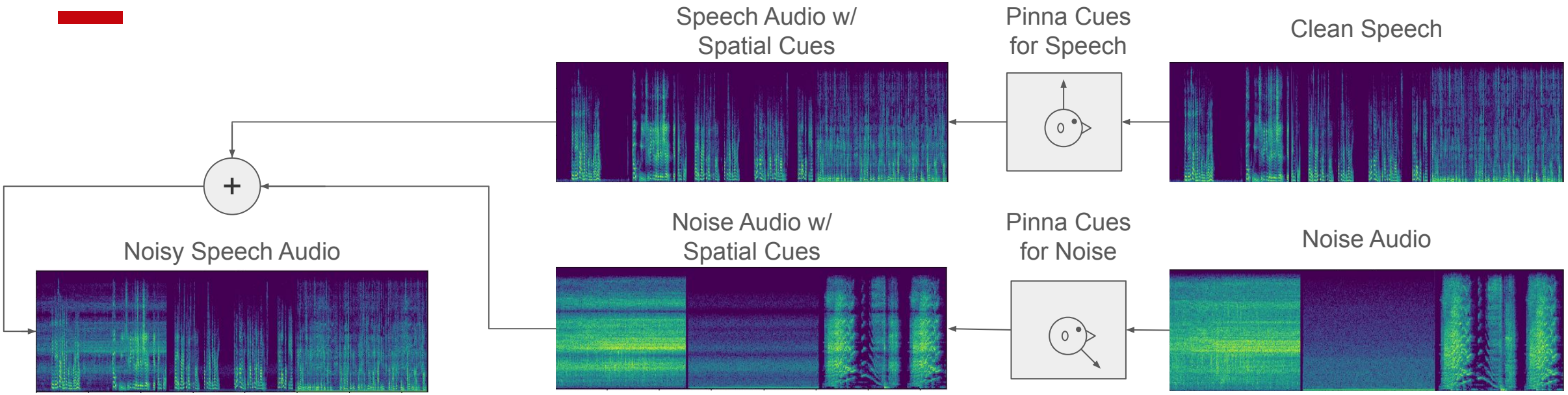
Methodology: Data Augmentation with Pinna Cues



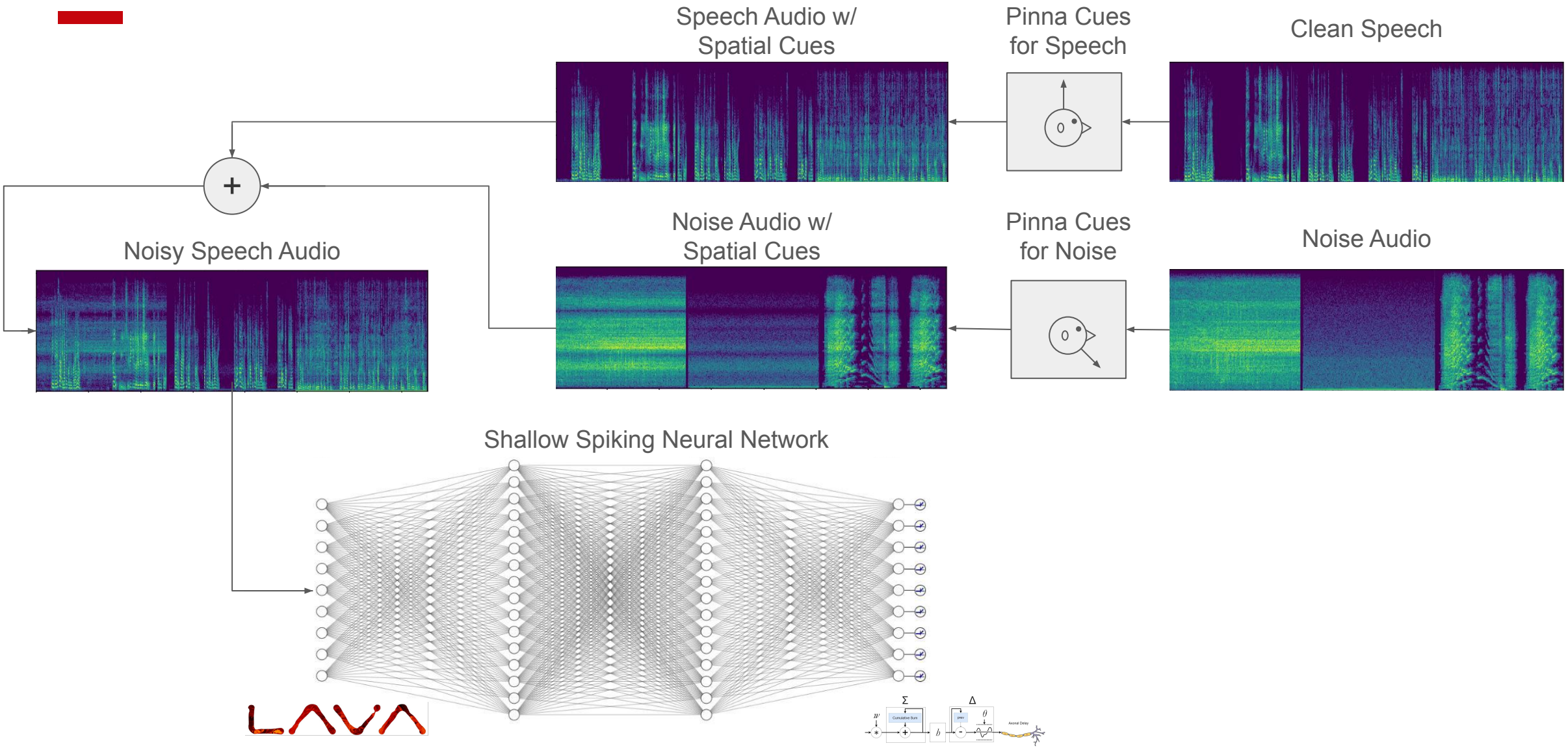
Methodology: Data Augmentation with Pinna Cues



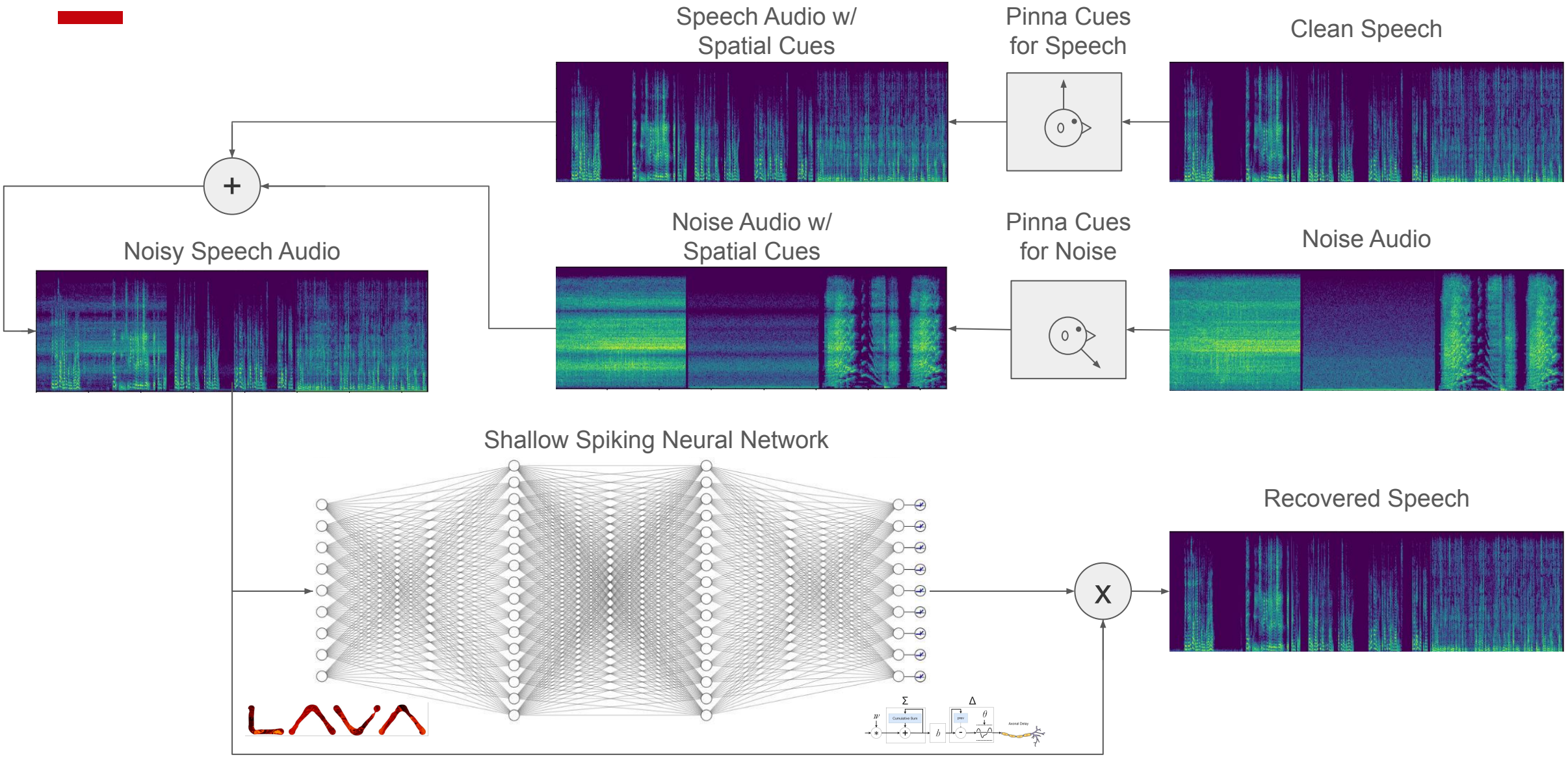
Methodology: Data Augmentation with Pinna Cues



Methodology: Data Augmentation with Pinna Cues

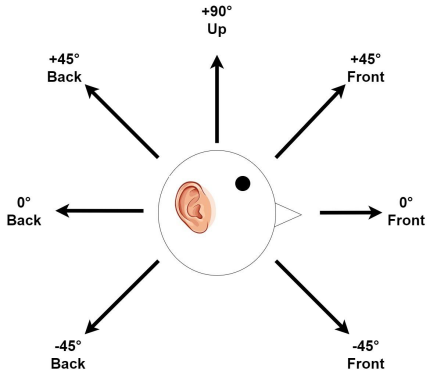


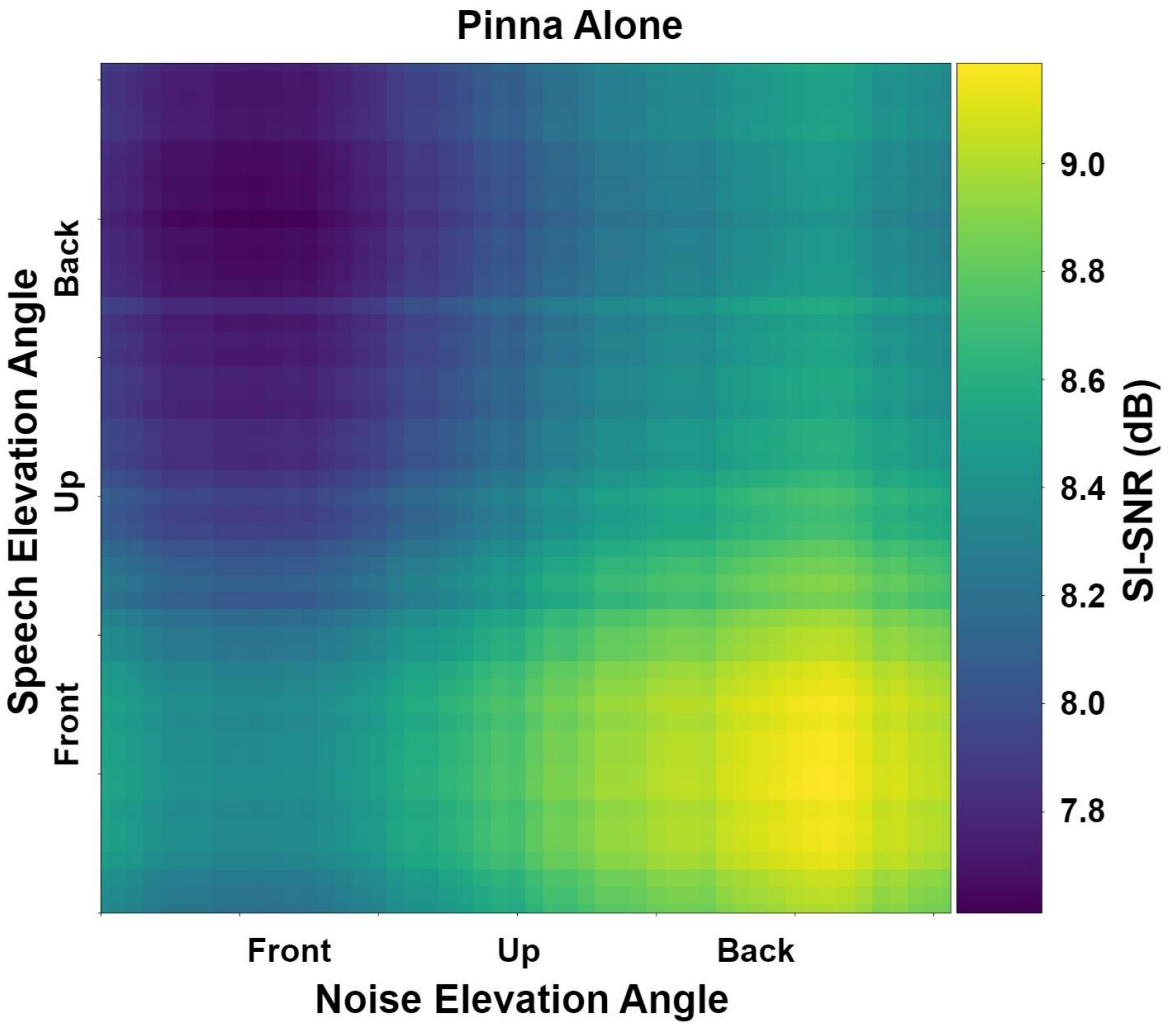
Methodology: Data Augmentation with Pinna Cues



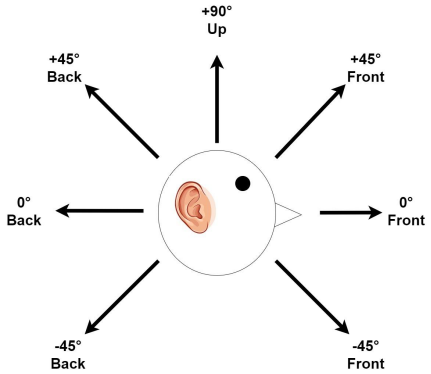
Evaluation

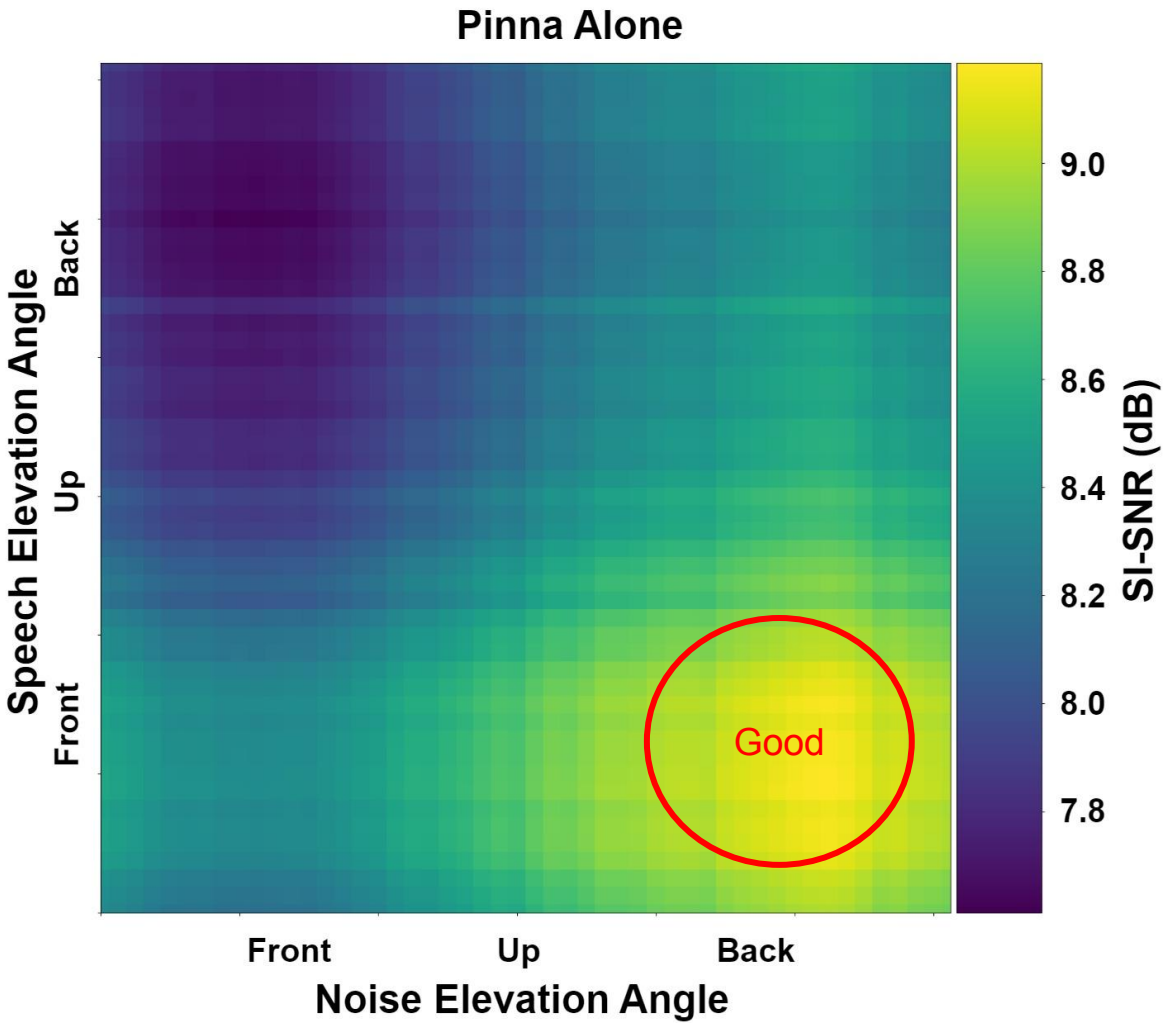
Evaluation: Fixed Pinna Results



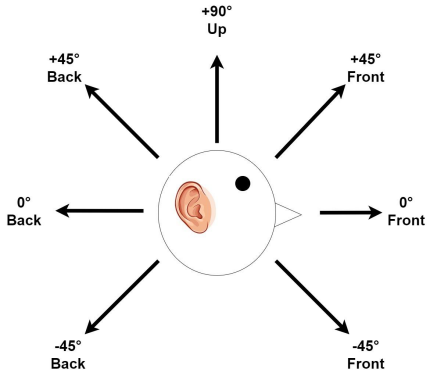


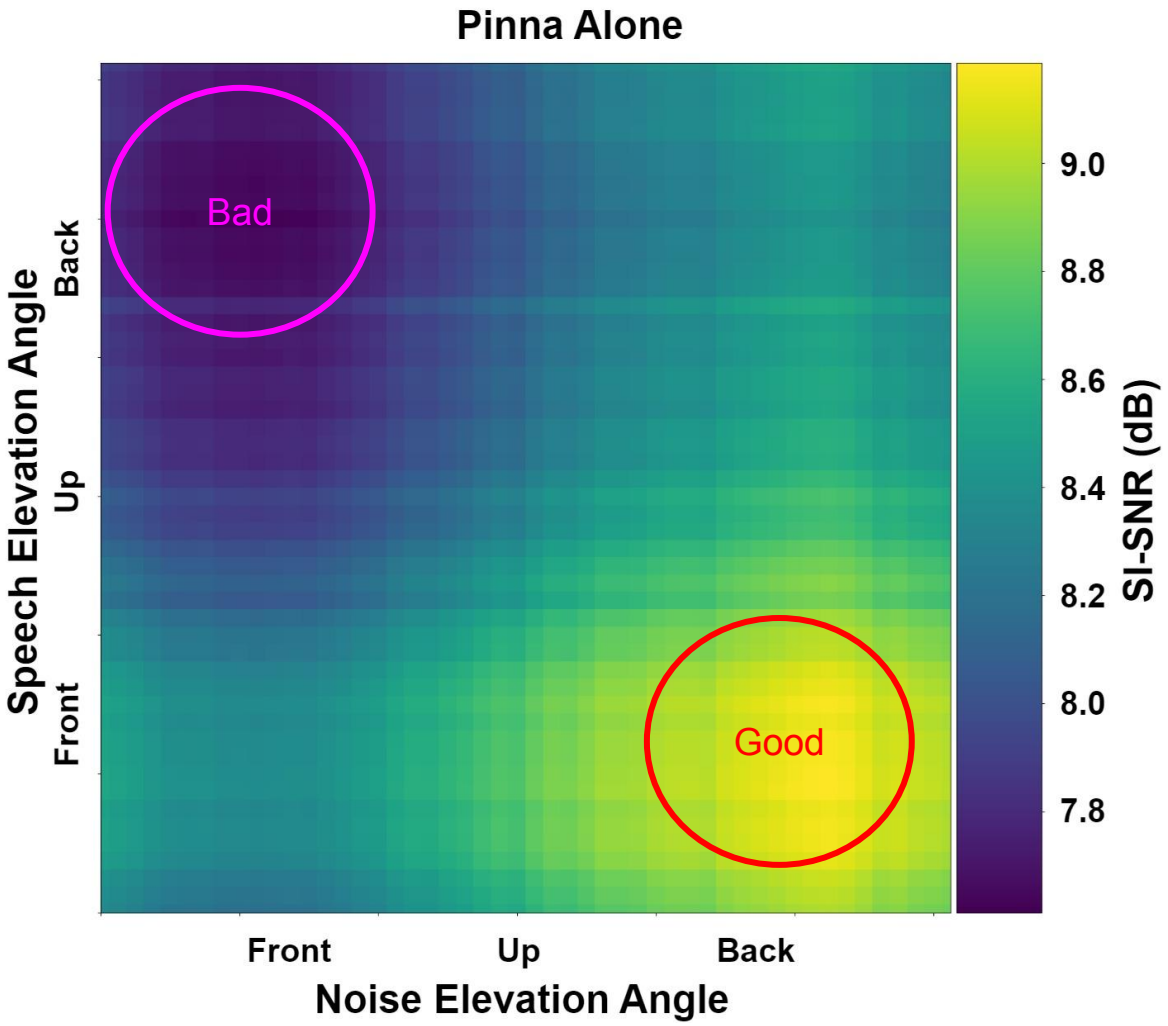
Evaluation: Fixed Pinna Results



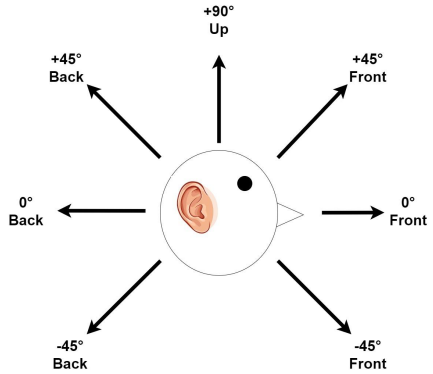


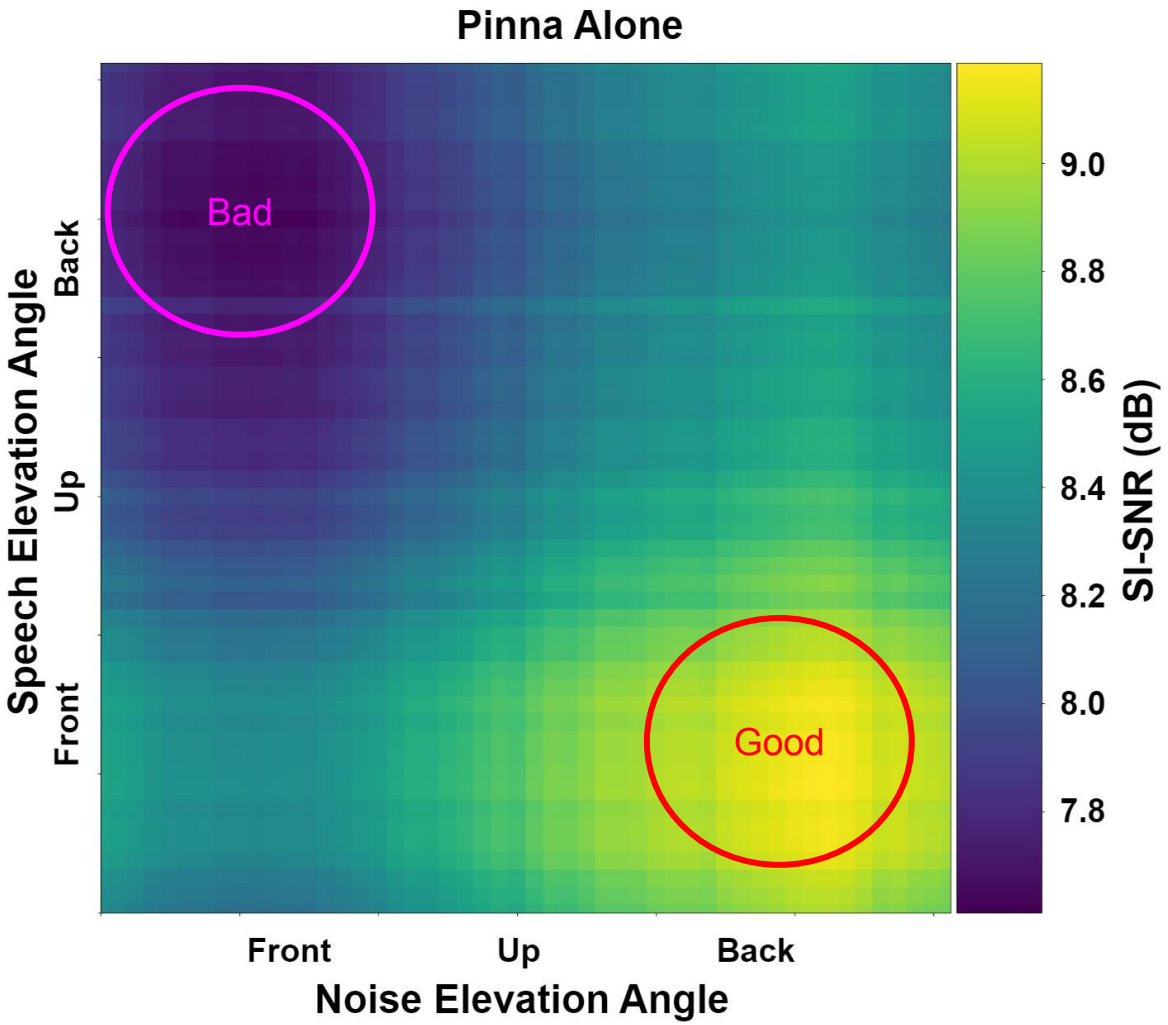
Evaluation: Fixed Pinna Results

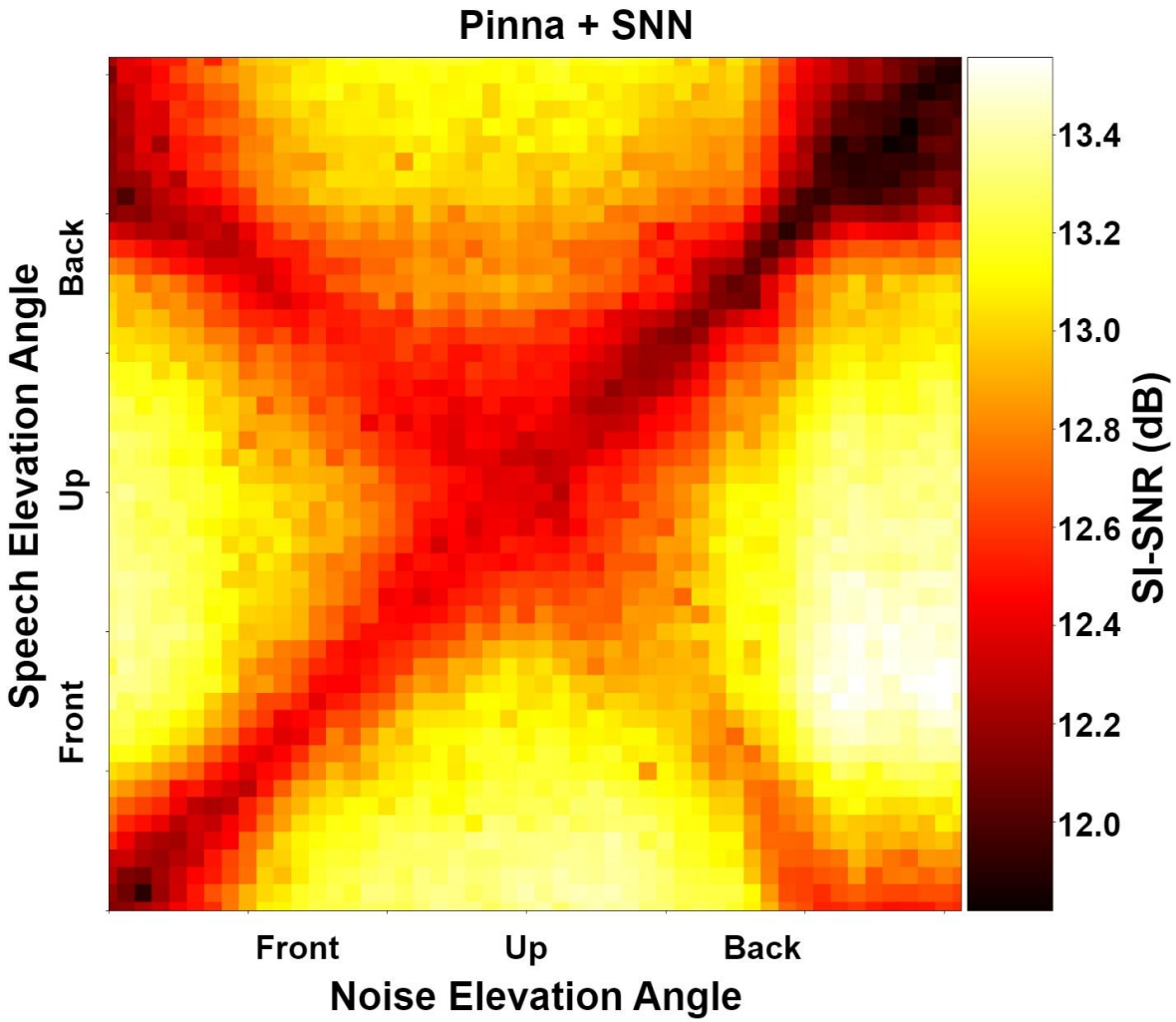




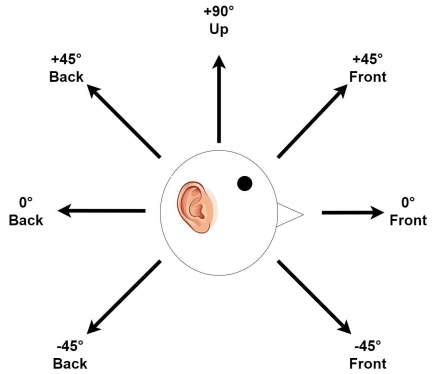
Evaluation: Fixed Pinna Results

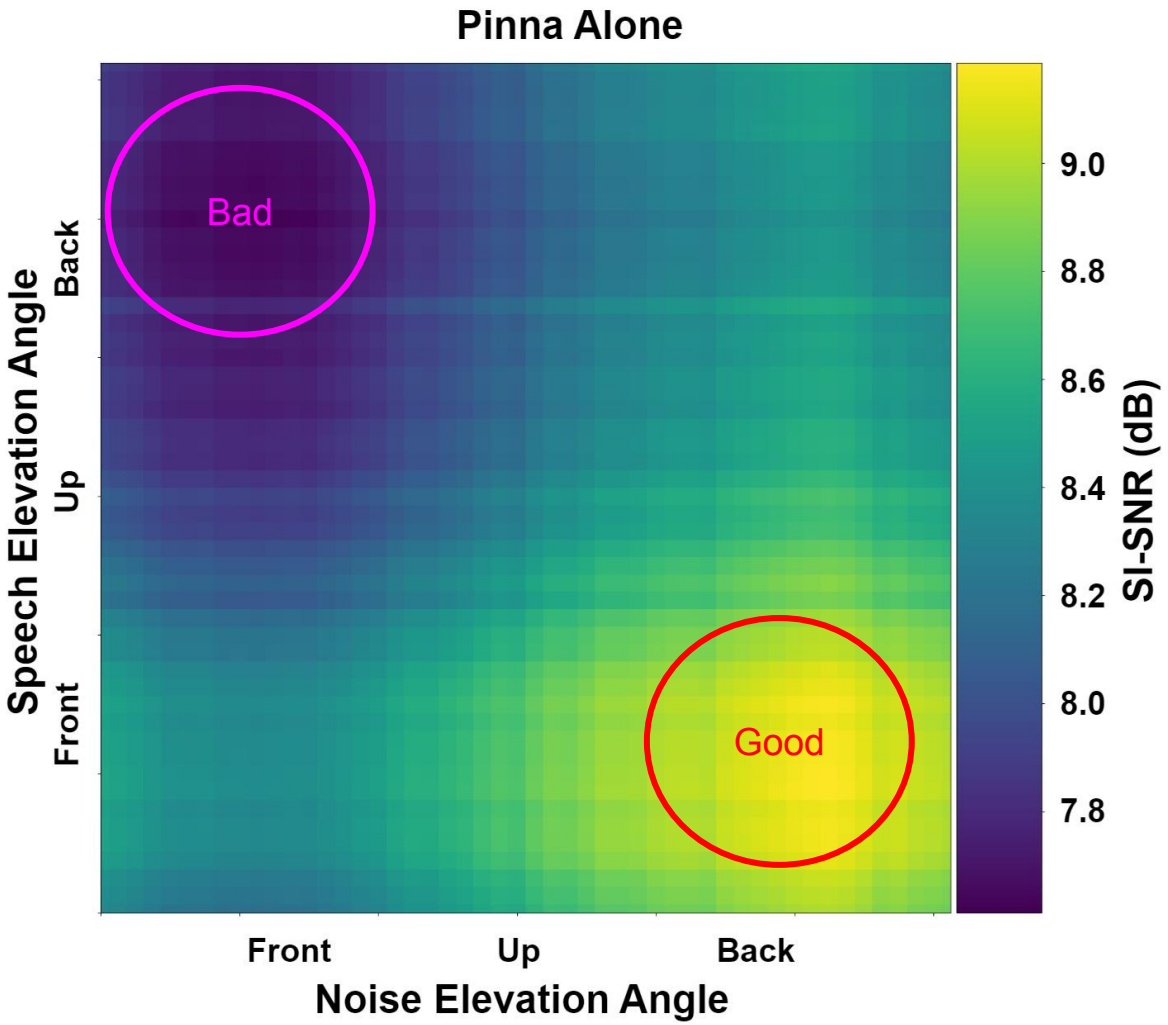


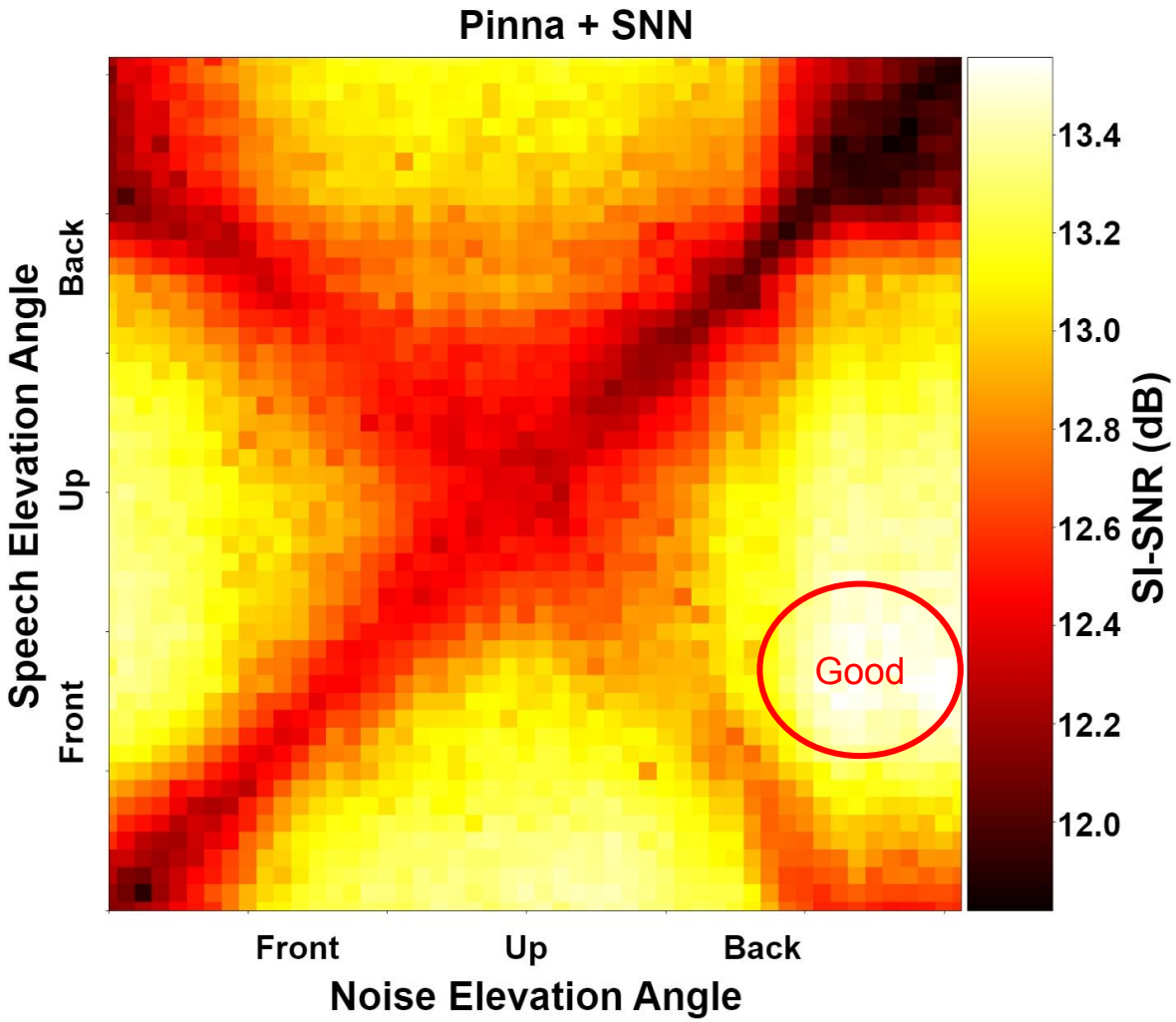




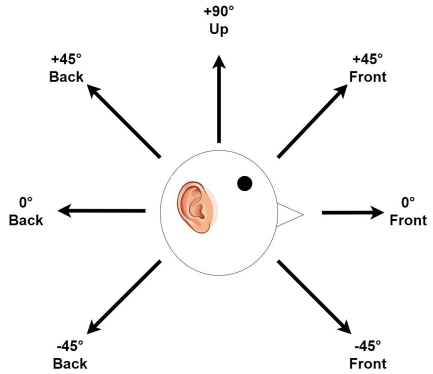
Evaluation: Fixed Pinna Results

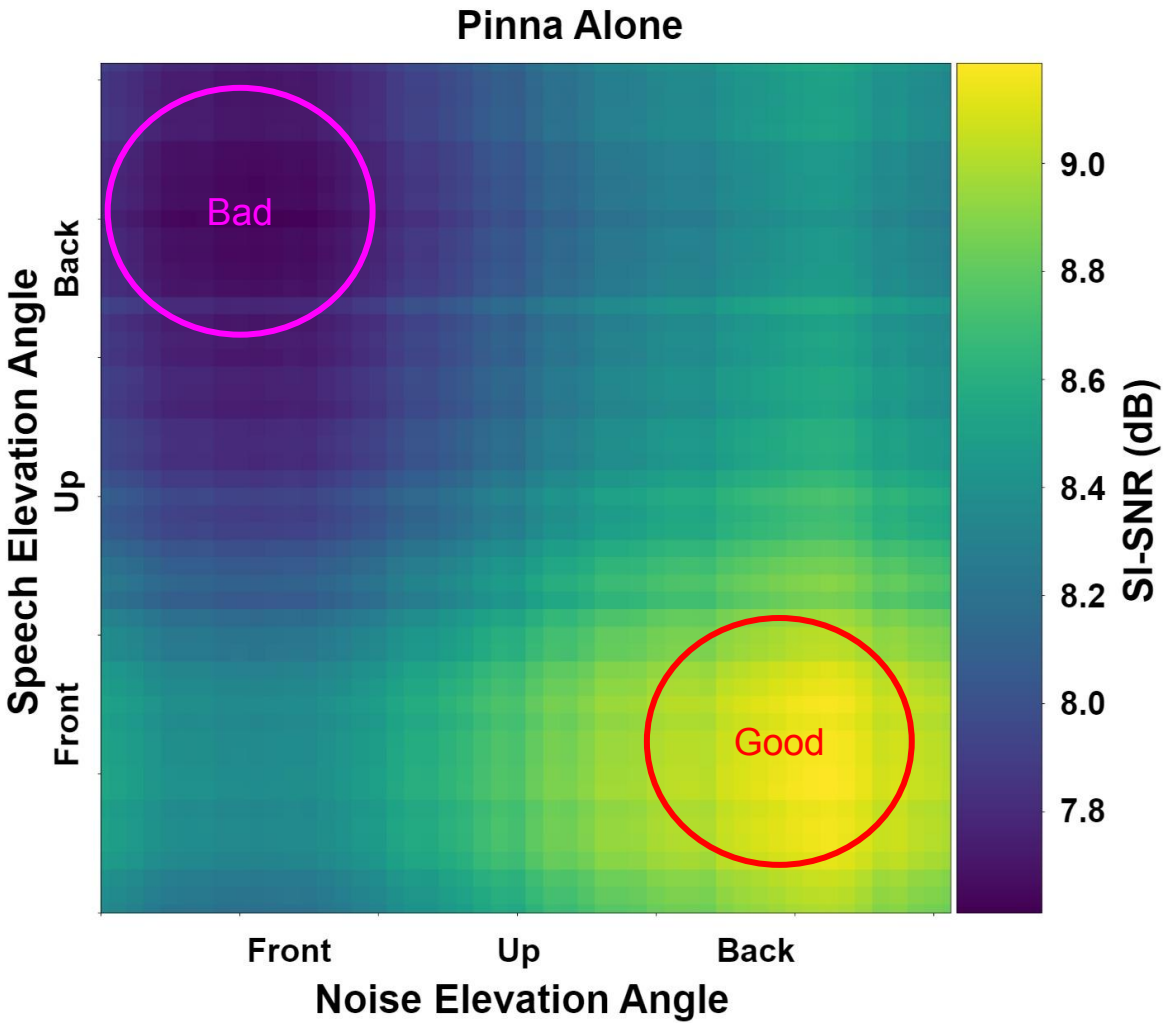


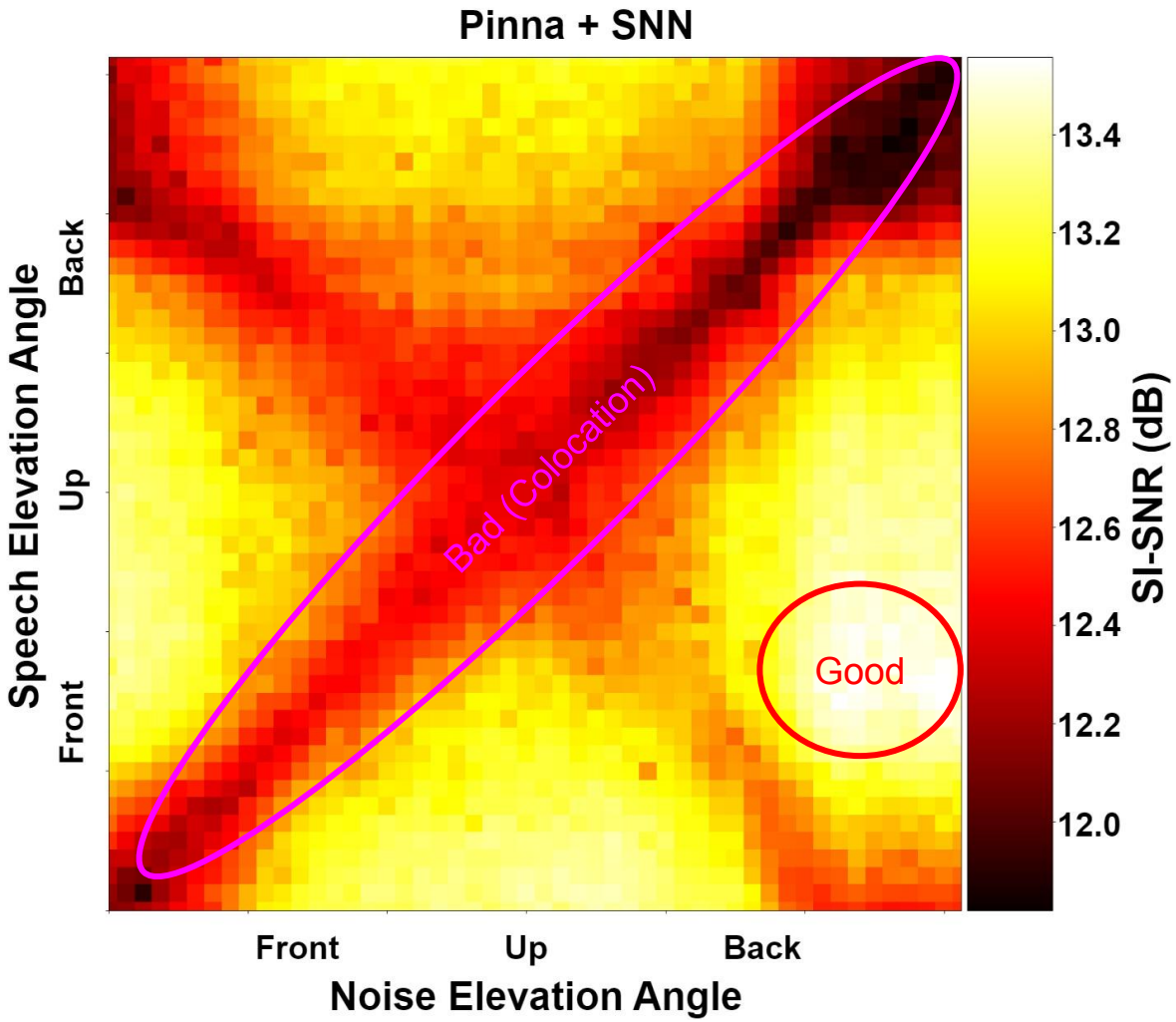




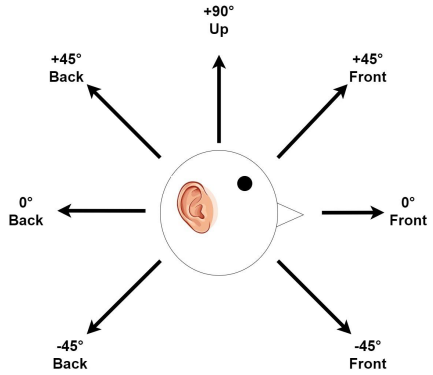
Evaluation: Fixed Pinna Results



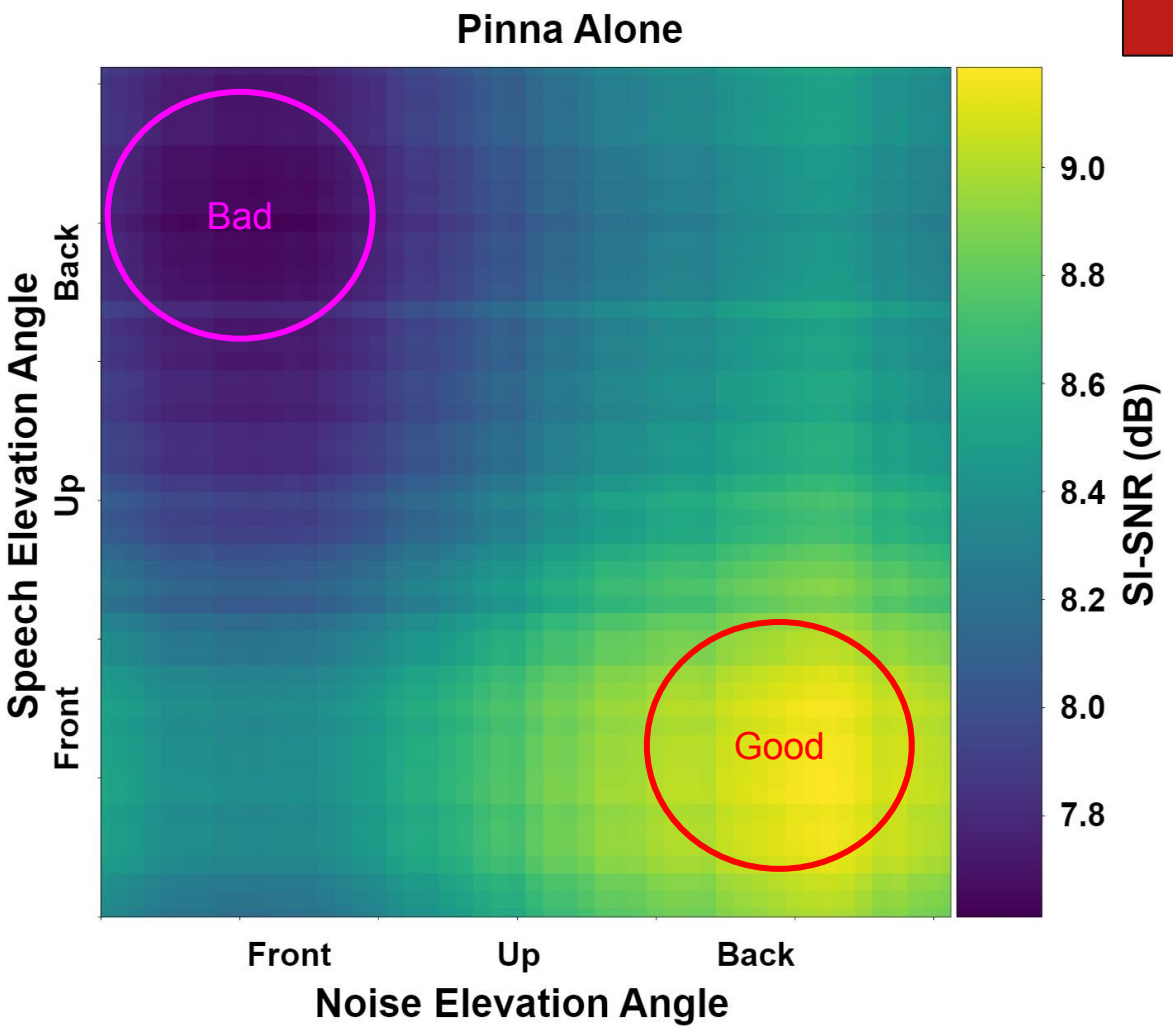


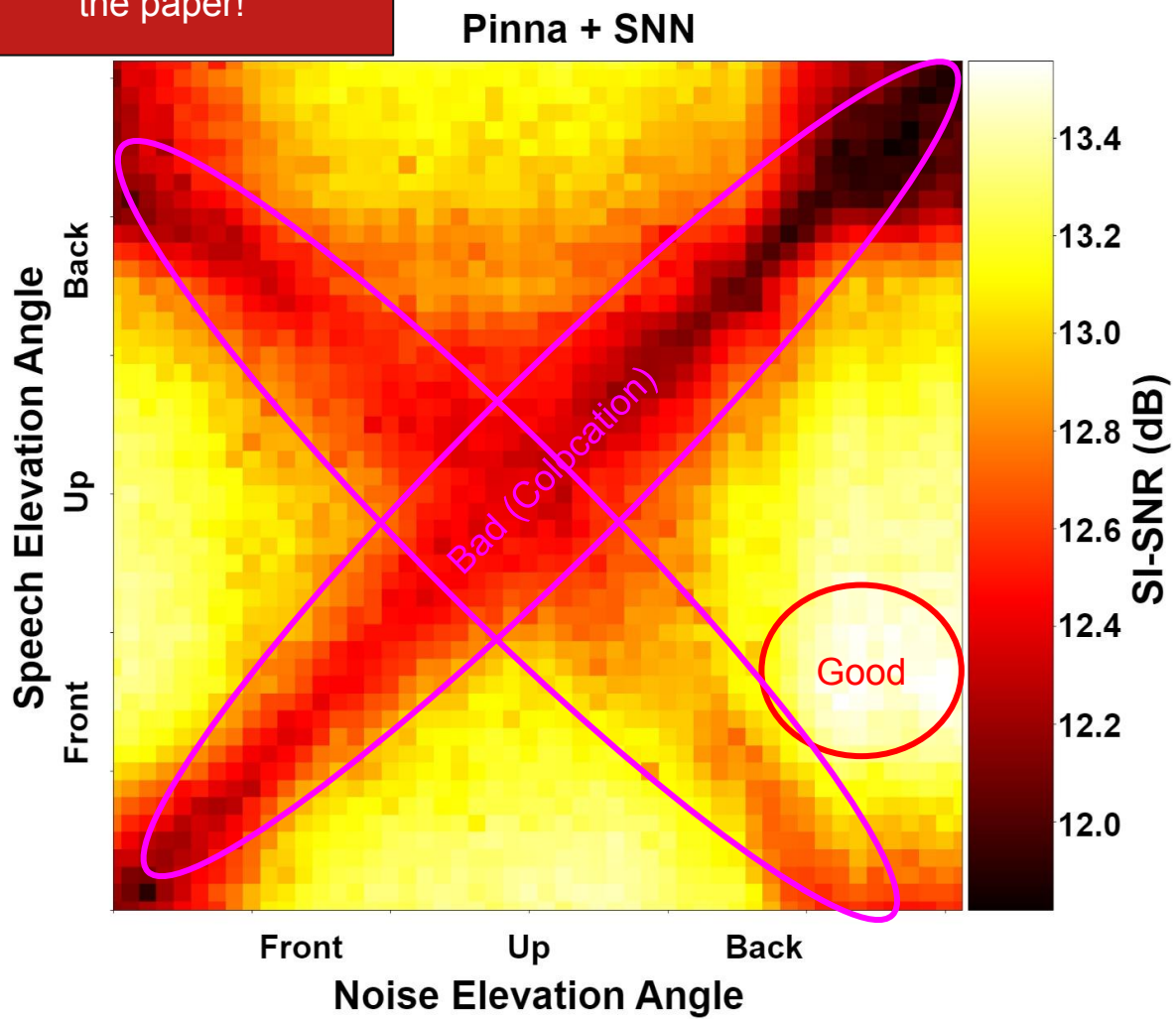


Evaluation: Fixed Pinna Results



More details about this in the paper!



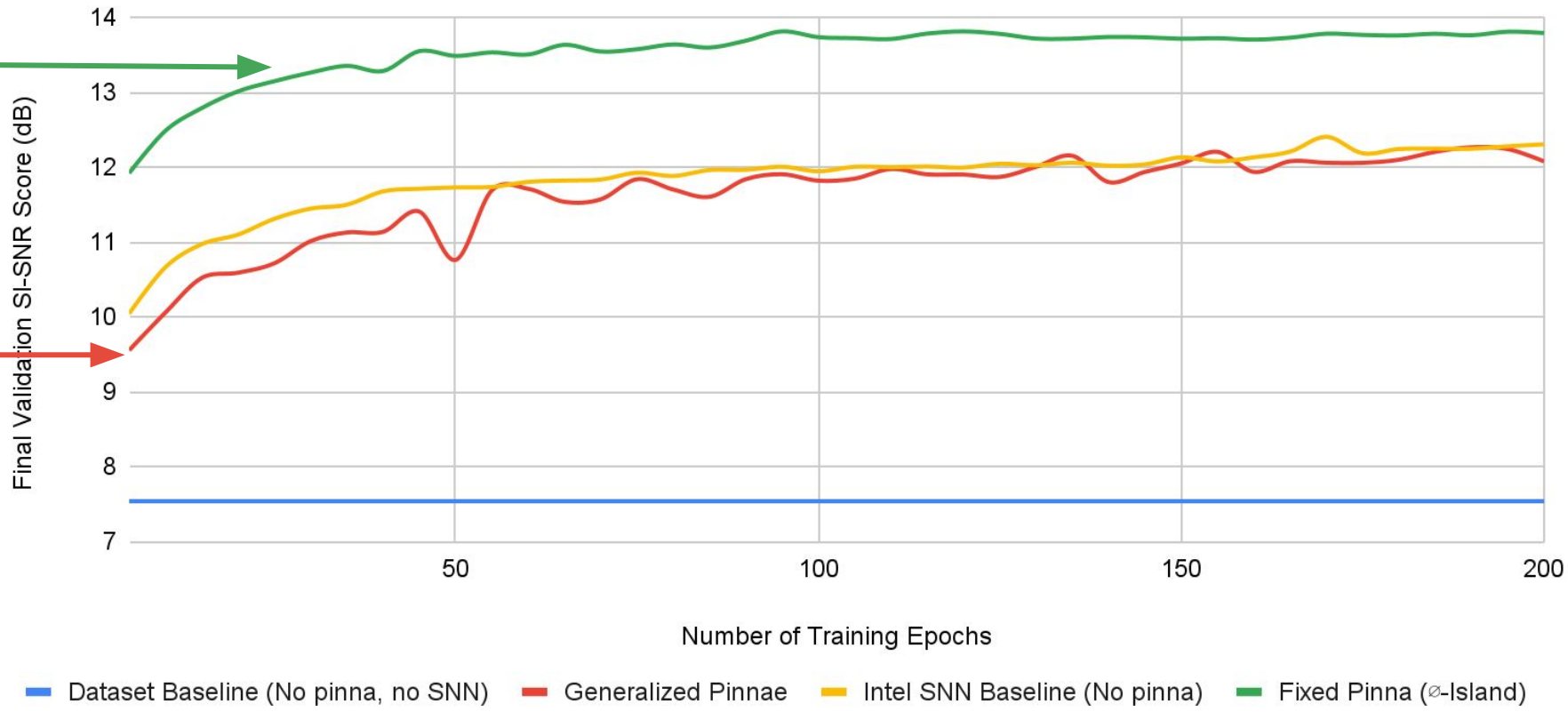


Evaluation: Towards a Generalized Pinna Model

- Desired feature: high denoising performance from arbitrary orientations
- Problem: broad range of orientations confuses the network

Want this perf

At this diversity

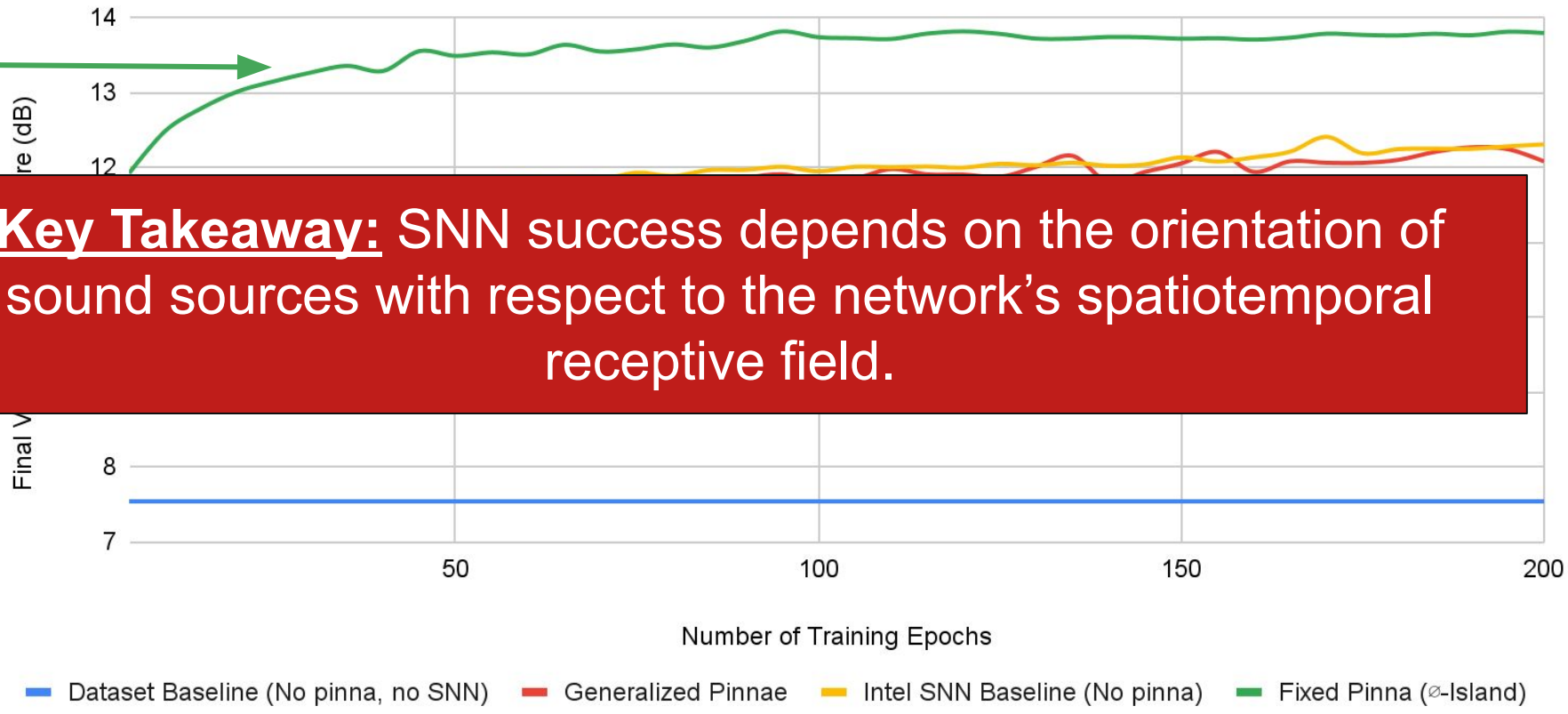


Evaluation: Towards a Generalized Pinna Model

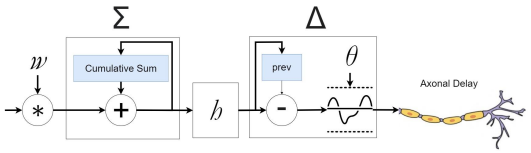
- Desired feature: high denoising performance from arbitrary orientations
- Problem: broad range of orientations confuses the network

Want this perf

At this diversit



Key Takeaway: SNN success depends on the orientation of sound sources with respect to the network's spatiotemporal receptive field.

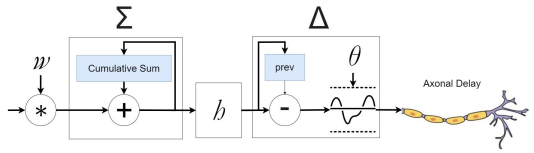


Evaluation: What helped the model learn?

- Average spoken word length: ~400 msec [1]
- Average spoken phoneme length: ~80 msec [2]
- Each FFT frame to the network represents 8 msec window

Trial	Set of Orientation Pairs {Speech}	Final Validation SI-SNR (dB)	Mean Weight for L1 Axonal Delay	Resultant Context Window (msec)
1	{∅-Island}	13.79	3.25	26
4	{NE_V, SE_V}	14.93	4.30	34.4
8	{NE_V, NW_V, SE_V, SW_V, NE_D, NW_D, SE_D, SW_D}	12.08	2.90	23.2

[1] Tian, Ye & Ferguson, et al. (2016). Processing negation without context – why and when we represent the positive argument. Language, Cognition and Neuroscience. 31. 1-16. 10.1080/23273798.2016.1140214.
[2] Ma, Guodong & Hu, et al. (2021). Leveraging Phone Mask Training for Phonetic-Reduction-Robust E2E Uyghur Speech Recognition. 10.21437/Interspeech.2021-964.



Evaluation: What helped the model learn?

- Average spoken word length: ~400 msec [1]
- Average spoken phoneme length: ~80 msec [2]
- Each FFT frame to the network represents 8 msec window

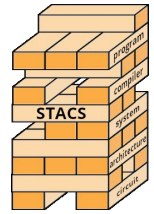
Key Takeaway: Pinna-enhanced SNNs only need sub-phoneme features for successful speech denoising.

Trial				tant Context dow (msec)
1				26
4	{NE_V, SE_V}	14.93	4.30	34.4
8	{NE_V, NW_V, SE_V, SW_V, NE_D, NW_D, SE_D, SW_D}	12.08	2.90	23.2

[1] Tian, Ye & Ferguson, et al. (2016). Processing negation without context – why and when we represent the positive argument. Language, Cognition and Neuroscience. 31. 1-16. 10.1080/23273798.2016.1140214.
[2] Ma, Guodong & Hu, et al. (2021). Leveraging Phone Mask Training for Phonetic-Reduction-Robust E2E Uyghur Speech Recognition. 10.21437/Interspeech.2021-964.

Conclusion

- Shallow SNN pipelines with improved biological fidelity are:
  - Efficient, both in model size and dataflow
  - Performant, able to achieve SOTA capabilities
  - Interpretable, may pose an alternative approach to mimicking and understanding the workings of the brain
- Entire workflow is open source and cloud ready via github + Docker
- Future: binaural audio, pinna shape predictor, pitch/SSL/foundation model, GPU training optimizations, edge compute



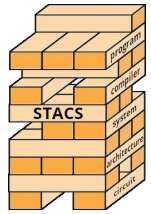
Conclusion

- Shallow SNN pipelines with improved biological fidelity are:
  - Efficient, both in model size and dataflow
  - Performant, able to achieve SOTA capabilities
  - In[illegible]nding the w[illegible]

Goal: Develop spatial audio tasks as killer applications for neuromorphic computing

- Entire workflow is open source and cloud ready via github + Docker
- Future: binaural audio, pinna shape predictor, pitch/SSL/foundation model, GPU training optimizations, edge compute

CHTC




WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON