CS441 Applied ML
ML for Audio Processing

Minje Kim, Ph.D.
Associate Professor
Dept. of Computer Science
https://minjekim.com
minje@illinois.edu

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

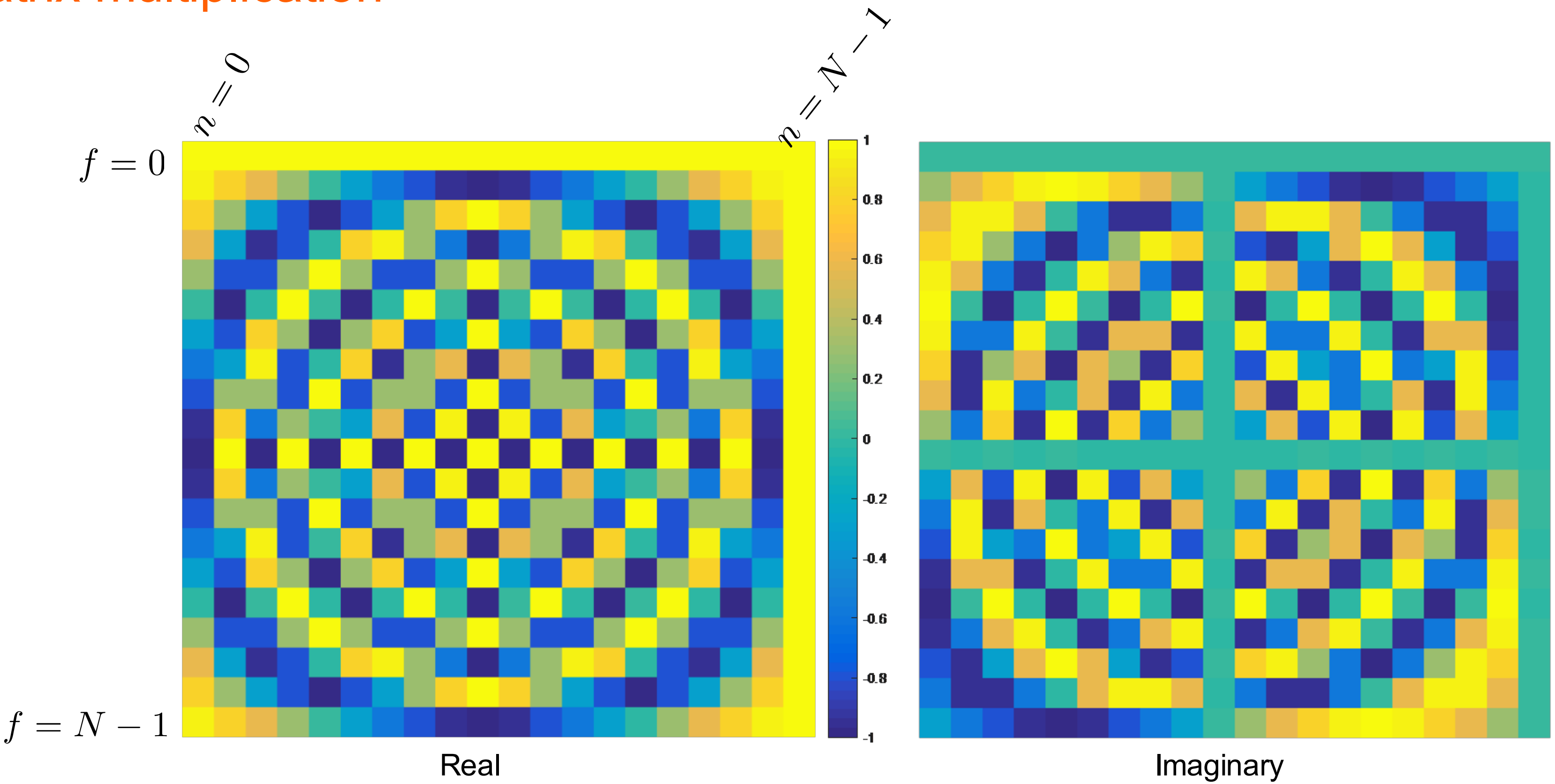
COMPUTER SCIENCE

GRAINGER ENGINEERING

Preliminaries

Discrete Fourier Transform

- As a matrix multiplication



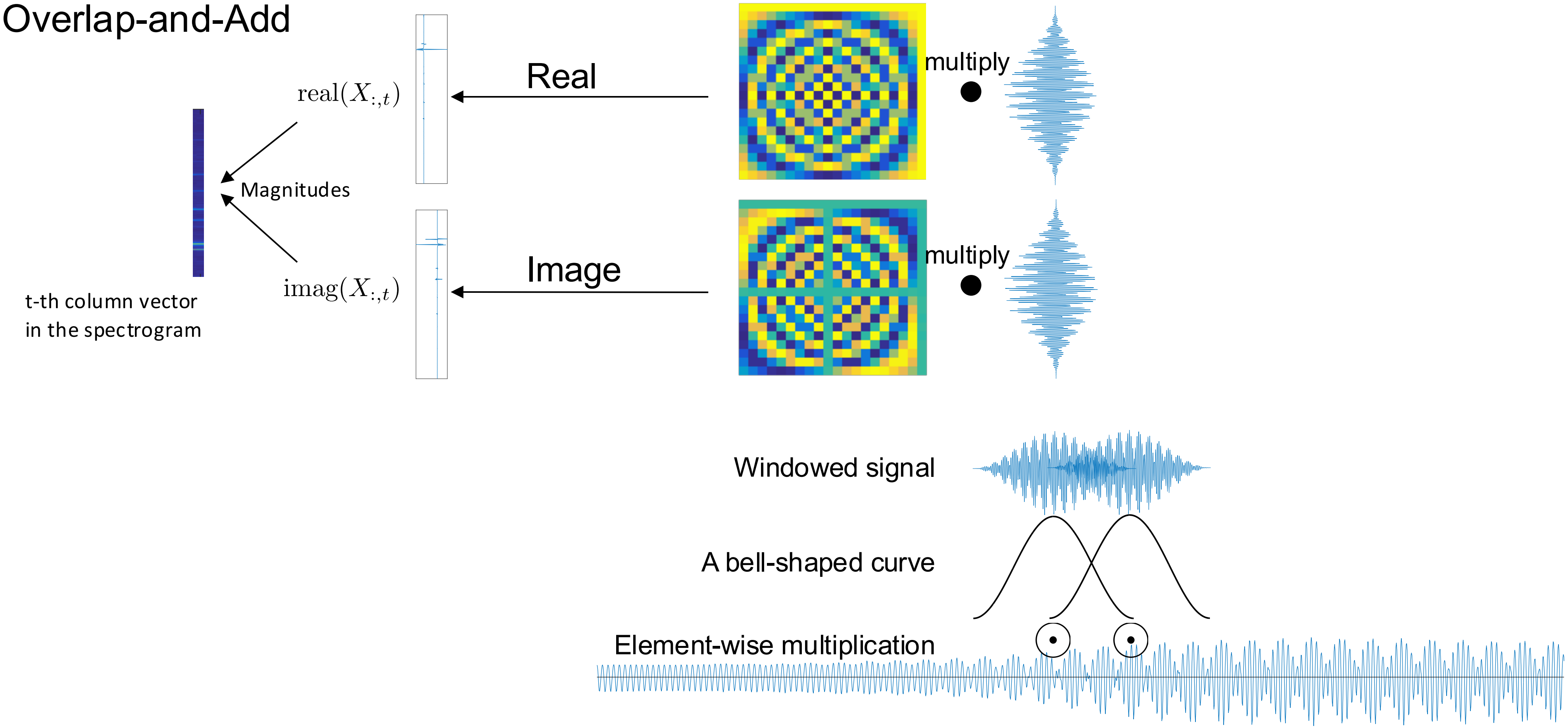
$$X_r(f) = \sum_n x[n]\cos(2\pi f \frac{n}{N})$$

$$X_i(f) = \sum_n x[n]\sin(2\pi f \frac{n}{N})$$

Short-Time Fourier Transform

- Windowing and overlap-and-add

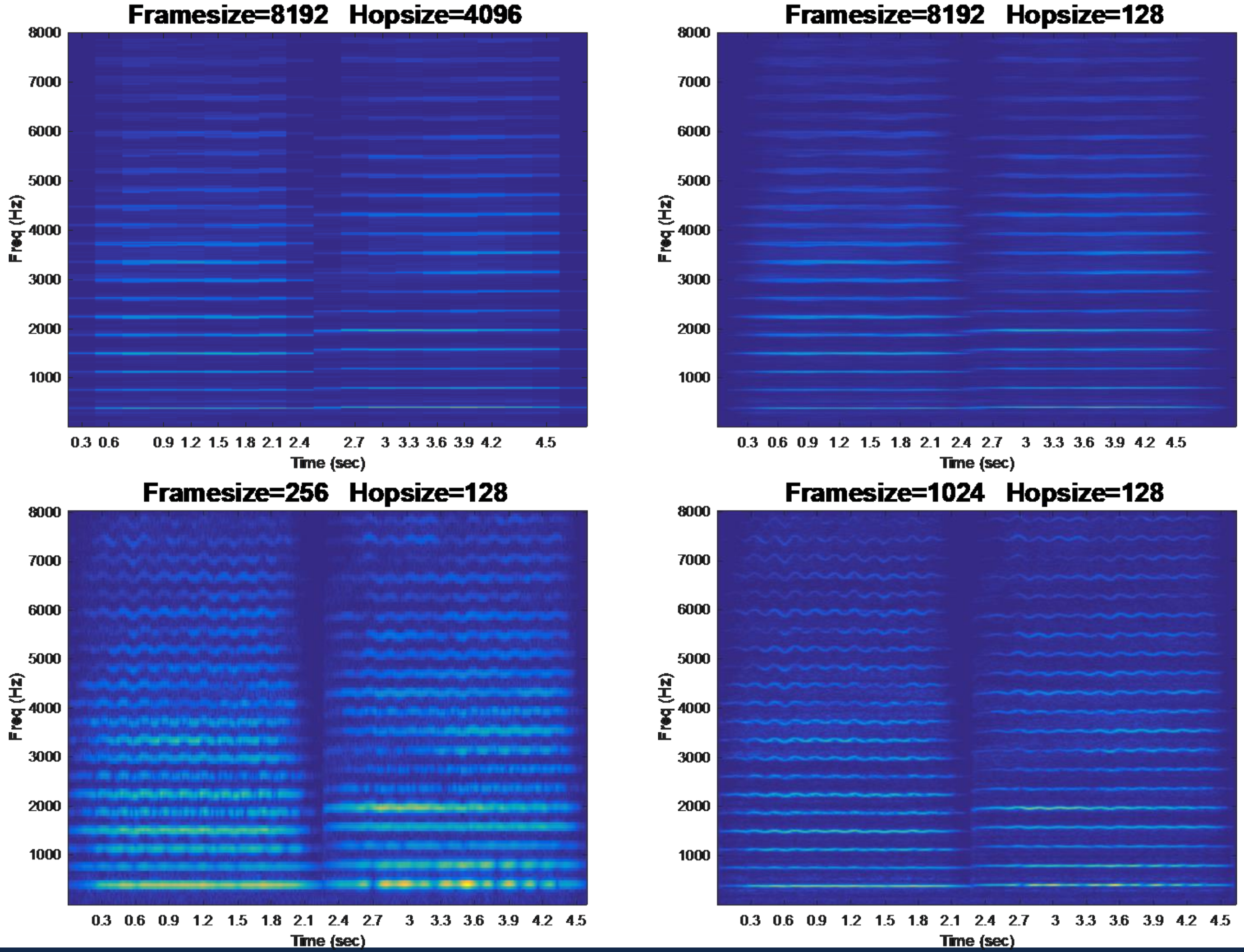
- Overlap-and-Add



Short-Time Fourier Transform

- Resolution control

- Trade-off between
  time and frequency resolutions



- Which one do you like
  the best?

Audio and Machine Learning

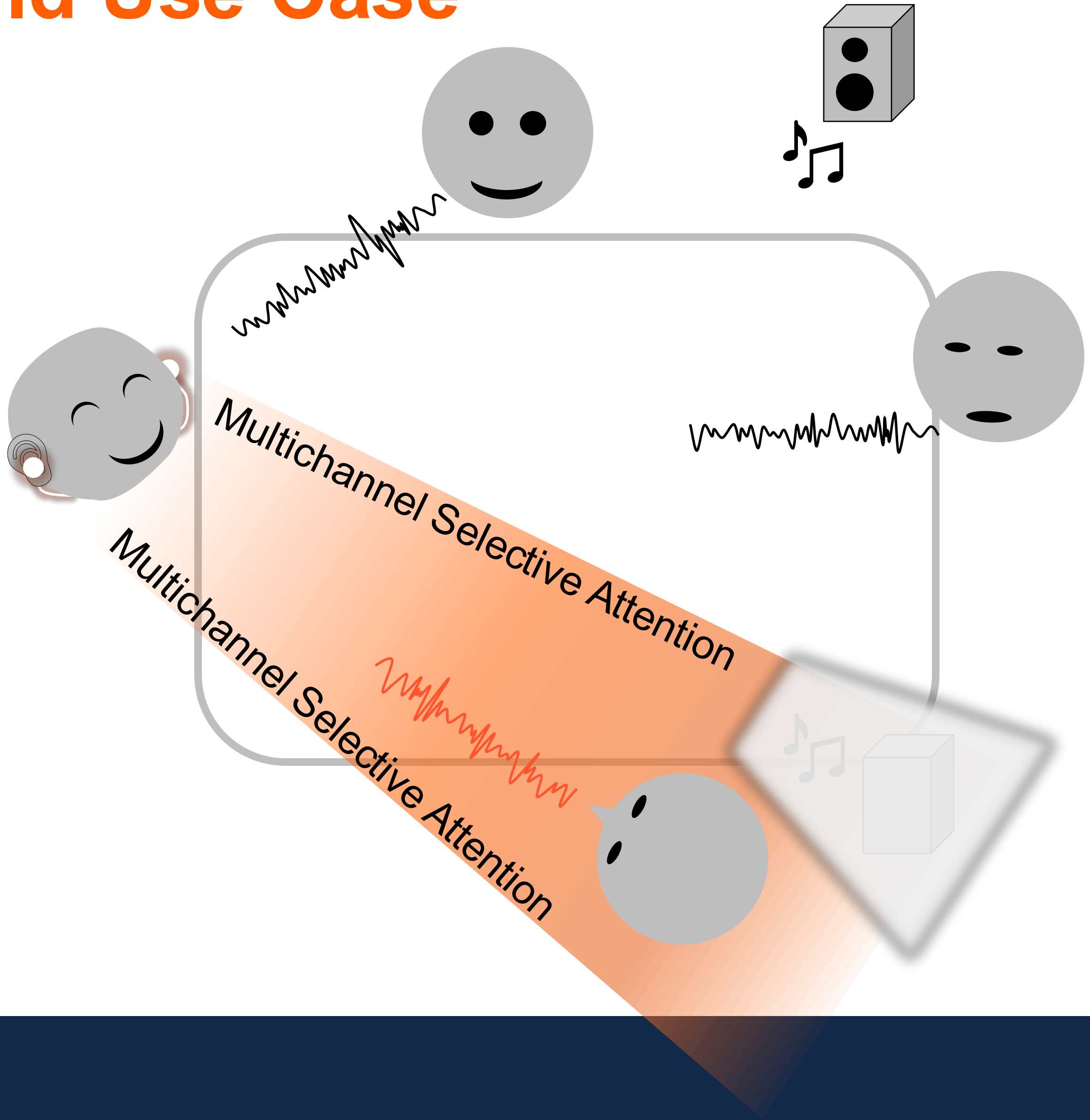
- From IEEE's perspective: EDICS

- IEEE Signal Processing Society; Audio and Acoustic Signal Processing Technical Committee

Audio signal processing	
	Signal enhancement, restoration, and extraction
	Audio and speech source separation
	Audio and speech coding, transmission, and representations
	Audio and speech quality and intelligibility measures
	Auditory modeling and hearing instruments
	System identification and dereverberation
	Acoustic sensor array processing
	Fundamental theory and algorithms for audio and acoustic signal processing
Acoustic scenes and events	
	Audio captioning, retrieval, and understanding
	Sound event and anomaly detection and sound scene classification
	Sound generation and synthesis
Acoustic environment processing	
	Modeling, analysis, and synthesis of acoustic environments
	Spatial audio recording and reproduction
	Active noise control; acoustic echo and feedback cancellation
Music analysis, processing, and generation	
	Music analysis
	Music signal processing, production, and separation
	Audio- and symbolic-domain music generation and content creation
Applications and other topics in audio and acoustic signal processing	
	Bioacoustics and medical acoustics
	Audio security
	Audio for video and multimedia
	Data and open source for audio and acoustic signal processing

Audio Signal Processing Problems

A Real-World Use Case



Speaker Localization

Speech Separation

Speaker Diarization

Speech Enhancement

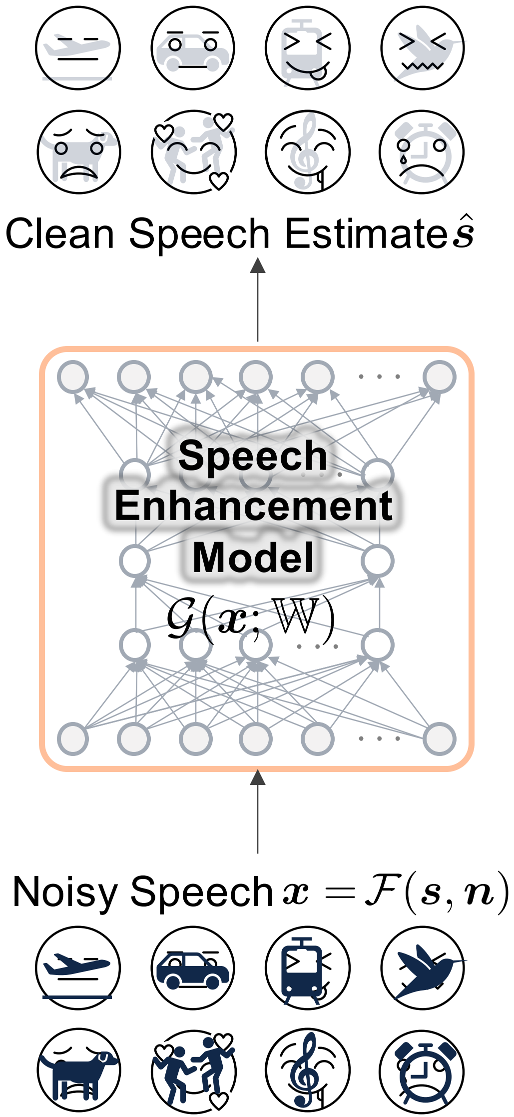
ML/NN for Speech Enhancement

- Learning a generalist

- 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)

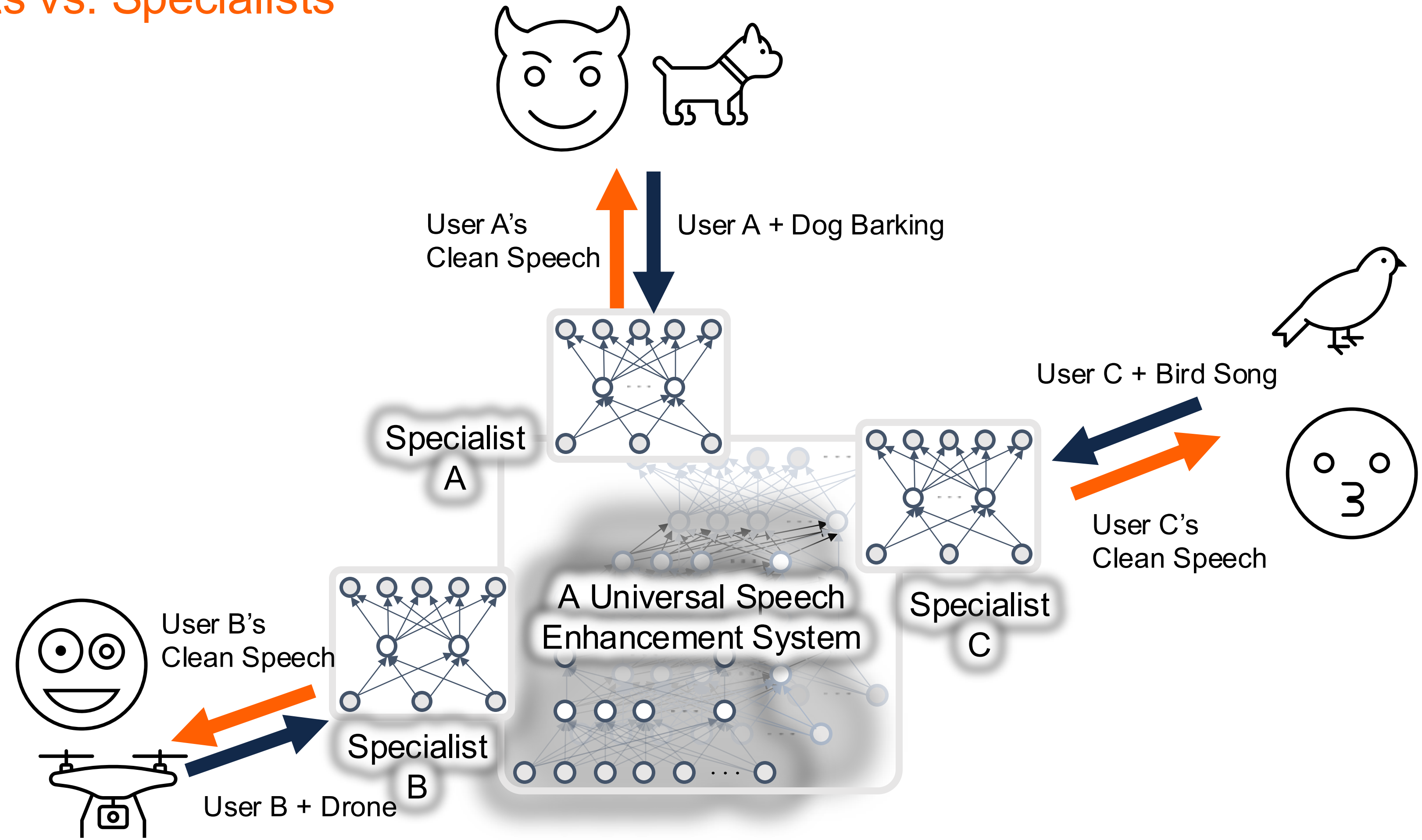
    $\boldsymbol{s} \approx \hat{\boldsymbol{s}} = \mathcal{G}(\boldsymbol{x}; \mathbb{W})$

- 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?



ML/NN for Speech Enhancement

- 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.

ML/NN for Speech Enhancement

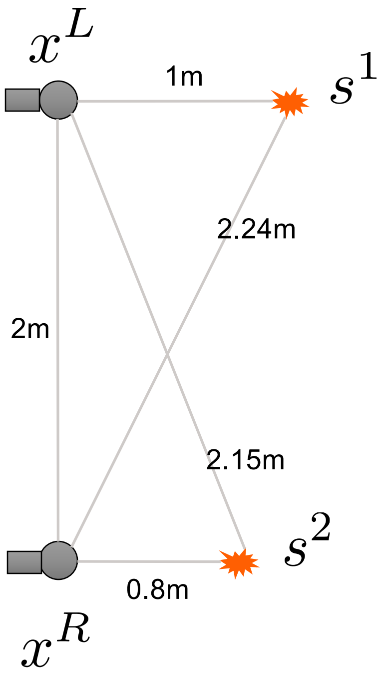
- Specialist Results

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

ML/NN for Speech Enhancement

- SPL and the geometry of the sources and sensors

- Sound Pressure Level (SPL) is inverse-proportional to the distance from the source



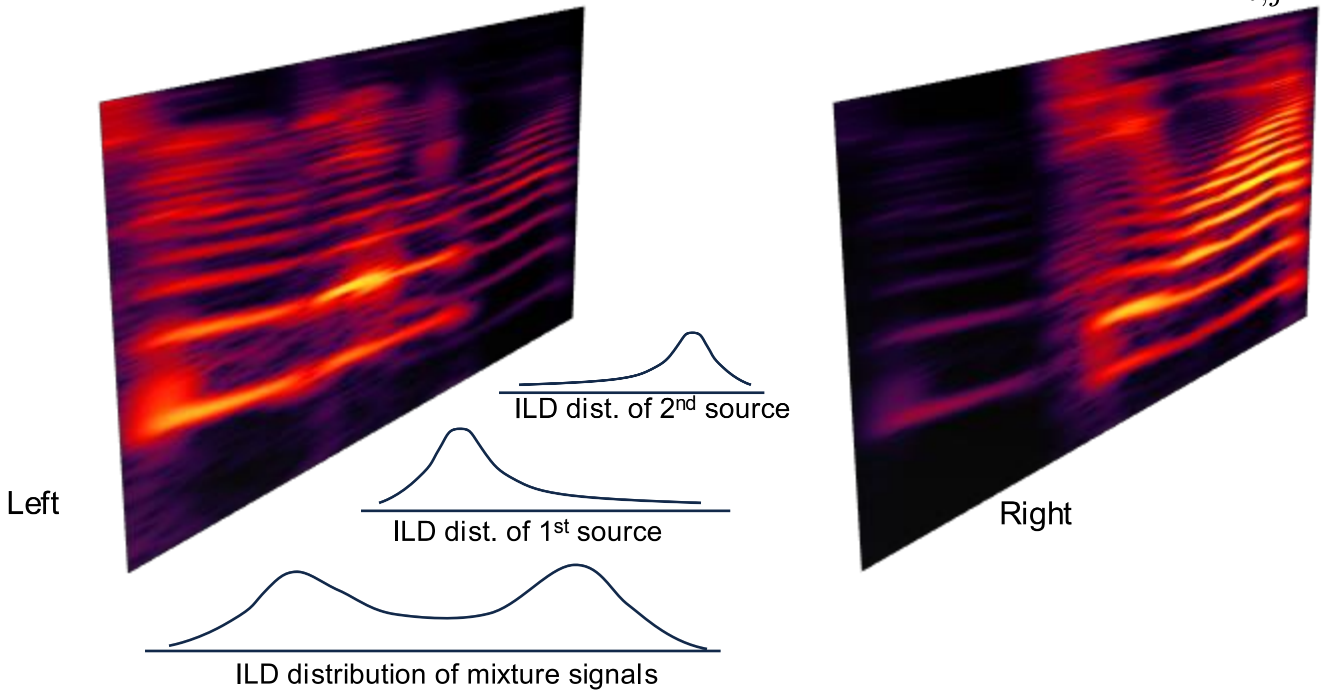
$$a_1 = \frac{\mathrm{SPL}(\hat{\boldsymbol{S}}^{1,L})}{\mathrm{SPL}(\hat{\boldsymbol{S}}^{1,R})} = \frac{\mathrm{dist}(\boldsymbol{X}^R, \boldsymbol{S}^1)}{\mathrm{dist}(\boldsymbol{X}^L, \boldsymbol{S}^1)} = \frac{2.24}{1} = 2.24$$

$$a_2 = \frac{\mathrm{SPL}(\hat{\boldsymbol{S}}^{2,L})}{\mathrm{SPL}(\hat{\boldsymbol{S}}^{2,R})} = \frac{\mathrm{dist}(\boldsymbol{X}^R, \boldsymbol{S}^2)}{\mathrm{dist}(\boldsymbol{X}^L, \boldsymbol{S}^2)} = \frac{0.8}{2.15} = 0.37$$

ML/NN for Speech Enhancement

- A clustering approach

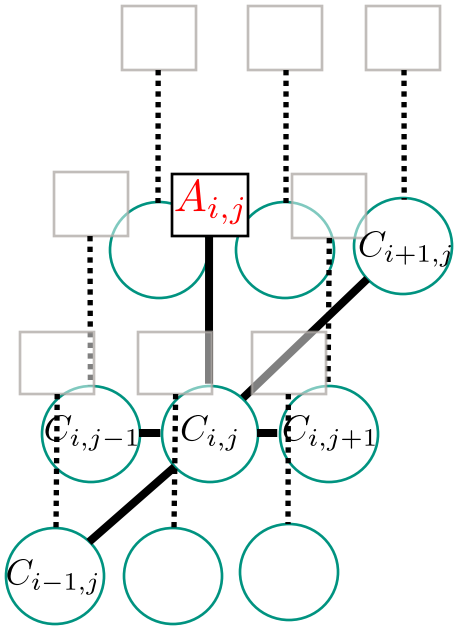
- Inter-channel Level Differences (ILD) can serve as a feature $A_{i,j} = 20 \log \frac{X^R_{i,j}}{X^L_{i,j}}$

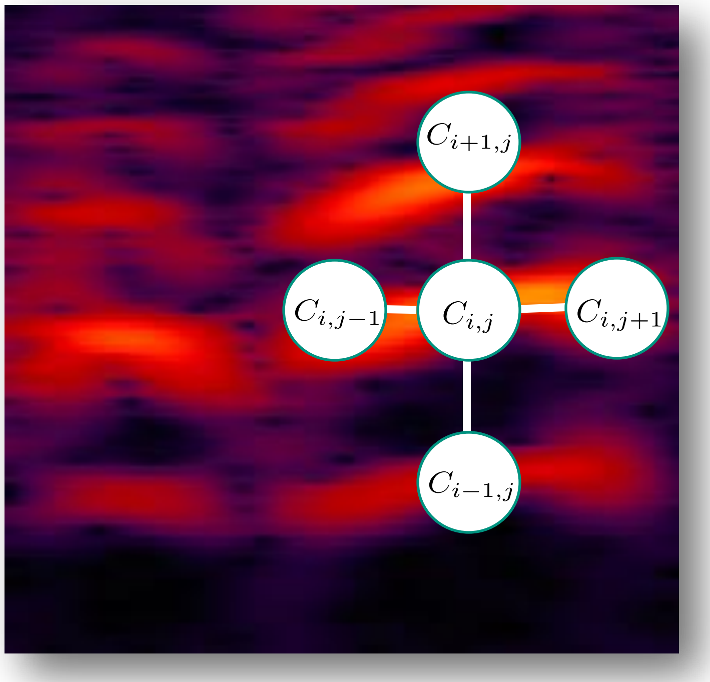


- The goal is to estimate source-wise distributions from their mixture
  - What kind of problem is it?
  - Clustering!

ML/NN for Speech Enhancement

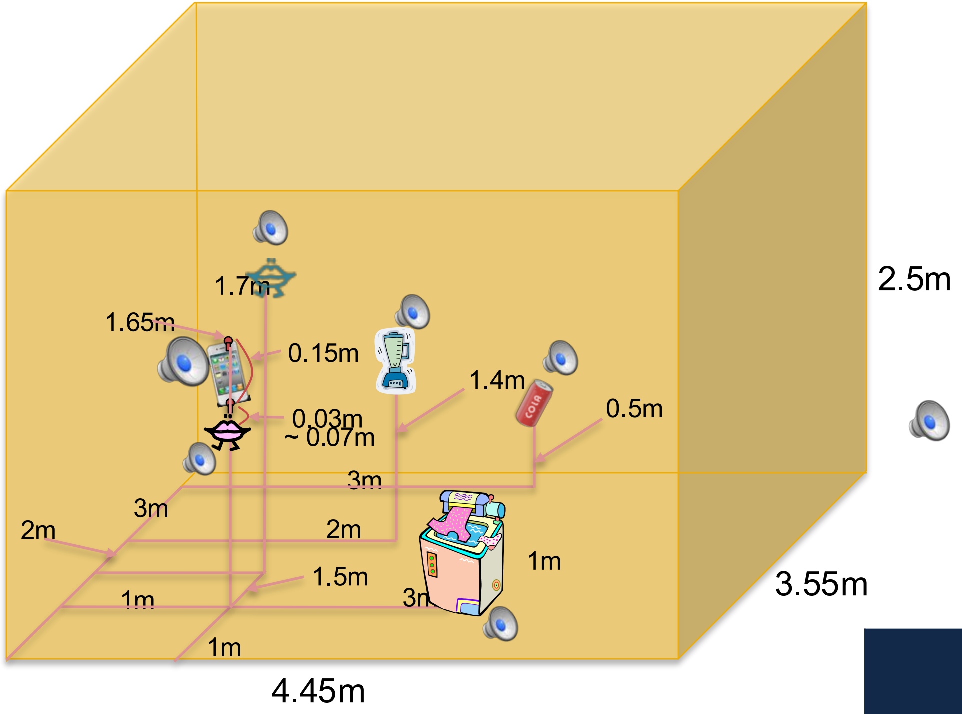
- The same pairwise MRF design





ML/NN for Speech Enhancement

- The mixing environment (multiple sources)

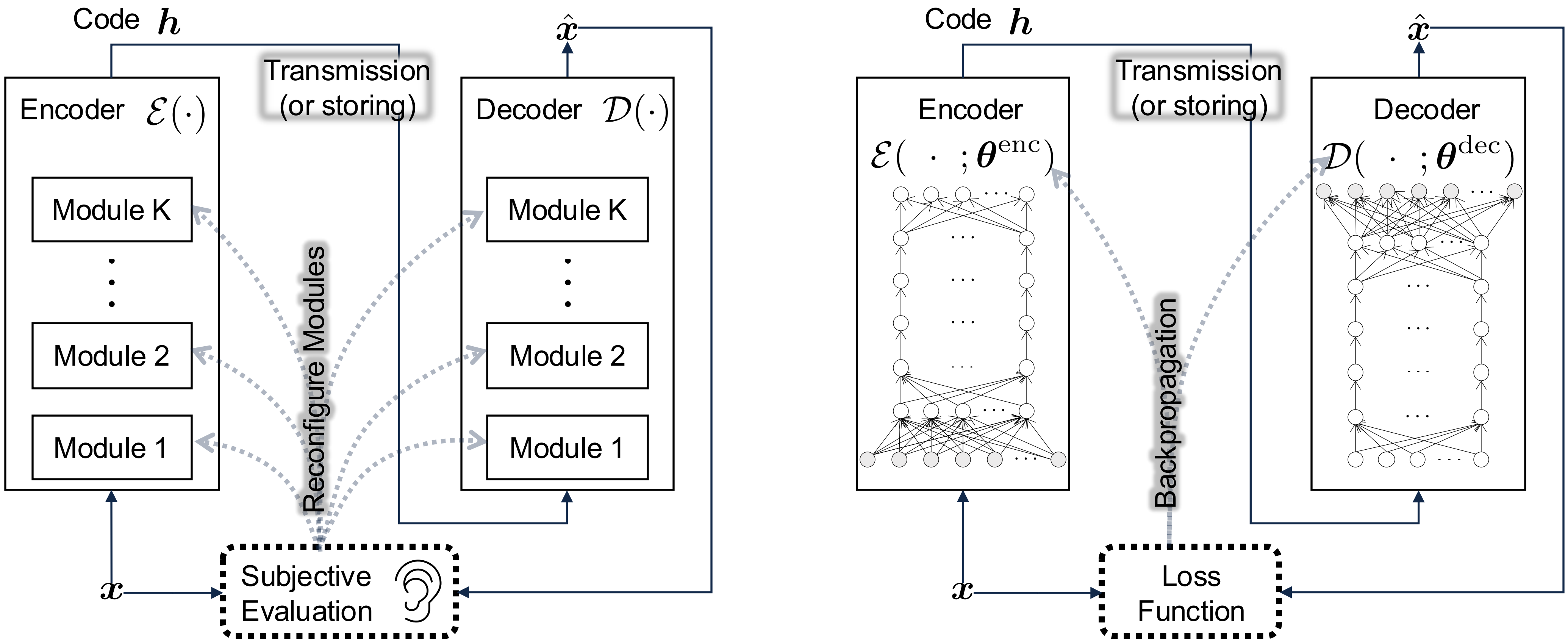


+Improvement from mixture
+Improvement by MRF smoothing

	Mixture	Vanilla GMM	MRF Smoothing
SDR	0.06	8.08 +8.02	10.42 +10.36 +2.34

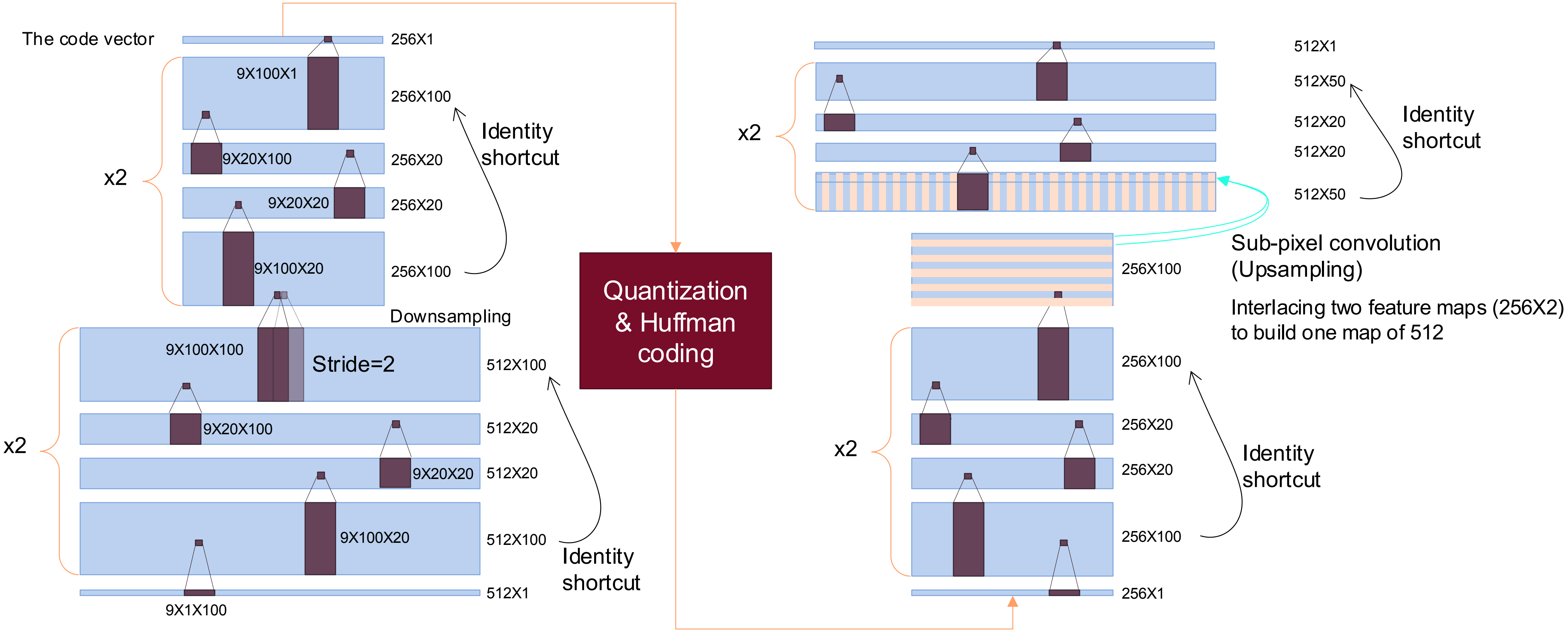
Neural Speech and Audio Coding

- Autoencoders vs. traditional codecs



Neural Speech and Audio Coding

- End-to-end CNN autoencoder

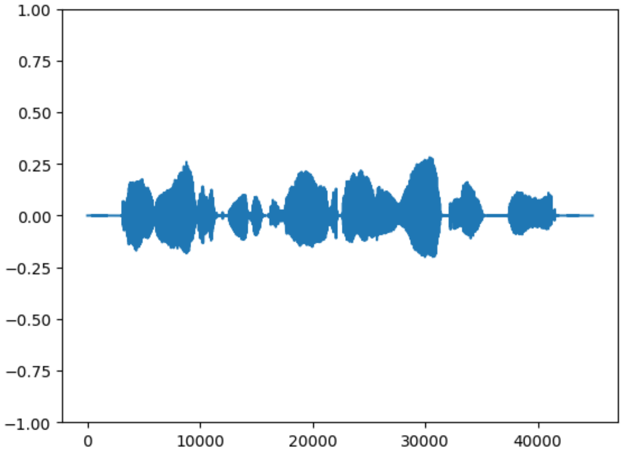


Perceptual Nature of Audio

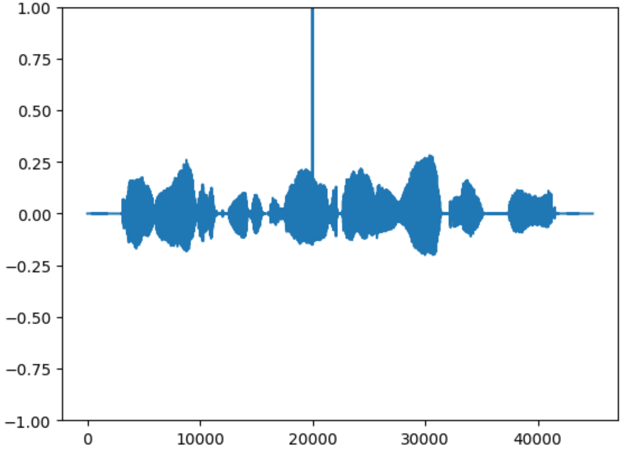
- Objective metrics are not good enough

- Time domain loss functions $\sum_t \mathcal{L}(s_t || \hat{s}_t)$

$$-10\log_{10}\frac{\sum_t s_t^2}{\sum_t (s_t - \hat{s}_t)^2} \qquad \sum_t (s_t - \hat{s}_t)^2 \qquad \sum_t |s_t - \hat{s}_t| \qquad \cdots$$

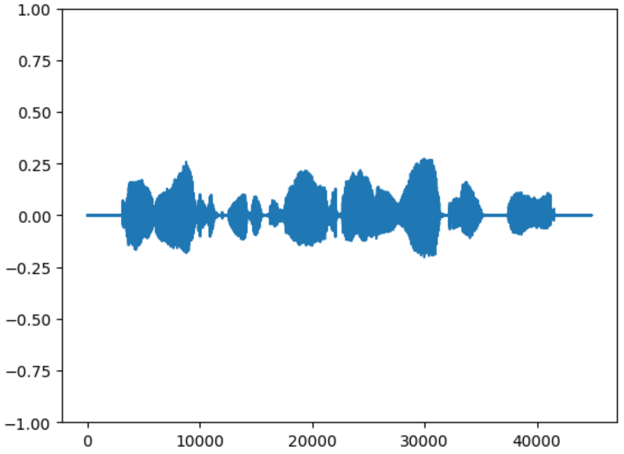


Original



Contaminated (version 1)

SNR = 20.14



Contaminated (version 2)

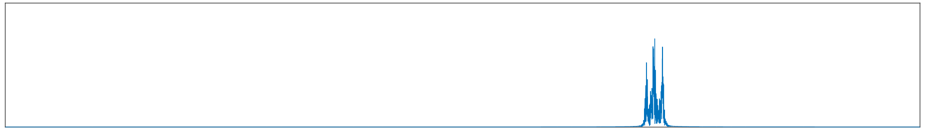
SNR = 18.30

Perceptual Nature of Audio

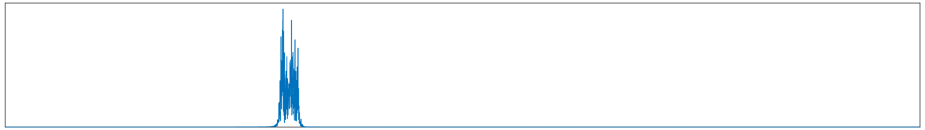
- You can't hear some tones!

- Which one is completely silent?

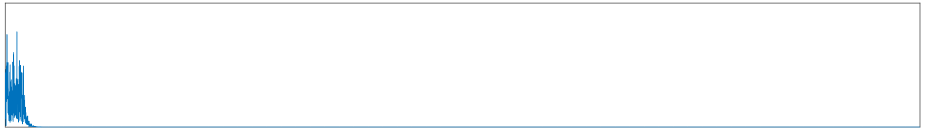
5600 – 5760Hz



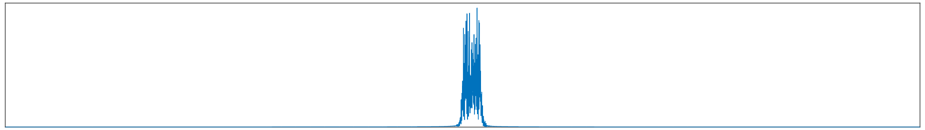
2400 – 2560Hz



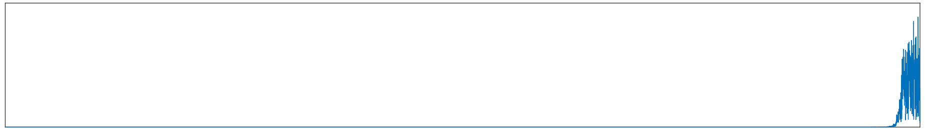
0 – 160Hz



800 – 960Hz



7840 – 8000Hz



Perceptual Nature of Audio

- You can’t hear some tones!

- Which one doesn’t have an interfering beep?

$$s(t) + n(t)$$

785Hz +20Hz

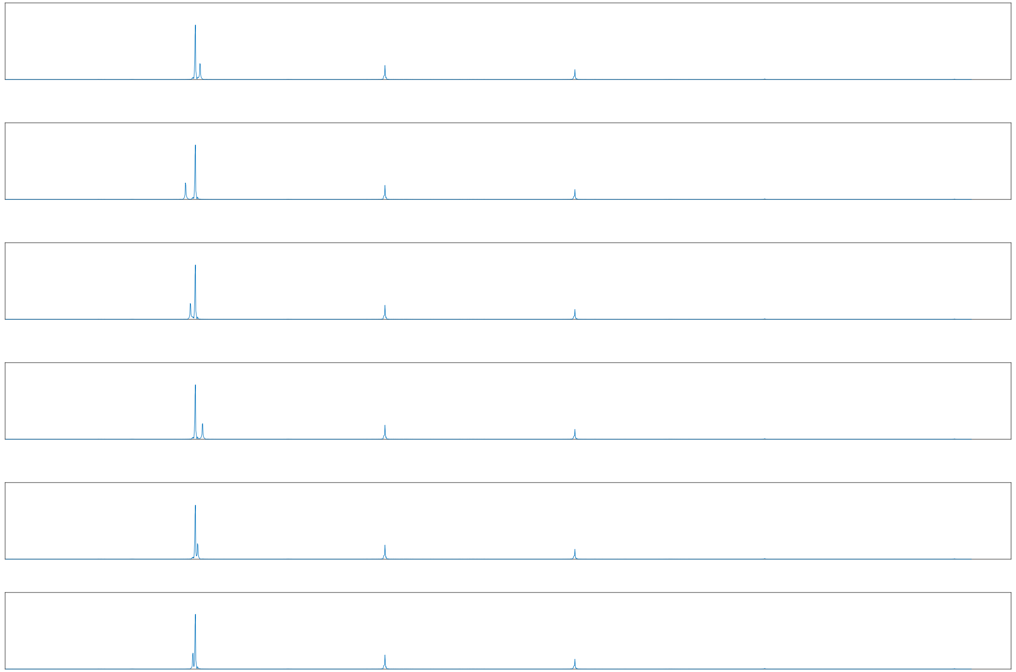
785Hz –40Hz

785Hz –20Hz

785Hz +30Hz

785Hz +10Hz

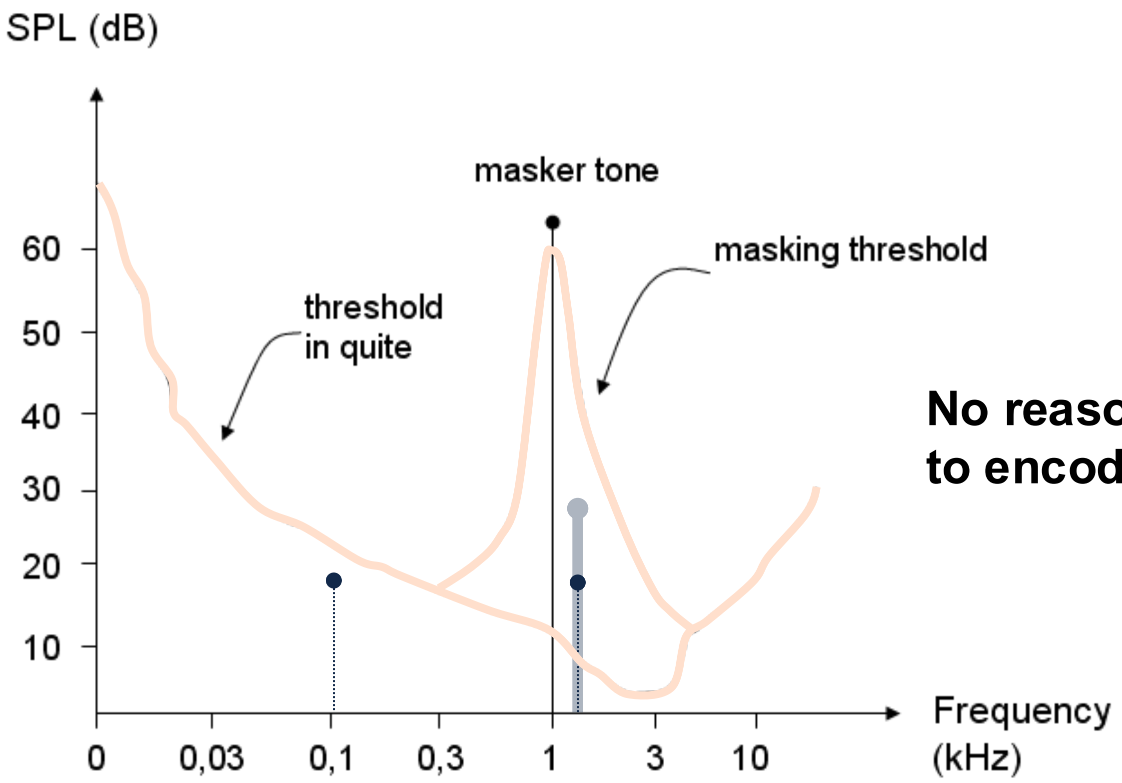
785Hz -10Hz



Perceptual Nature of Audio

- Psychoacoustics

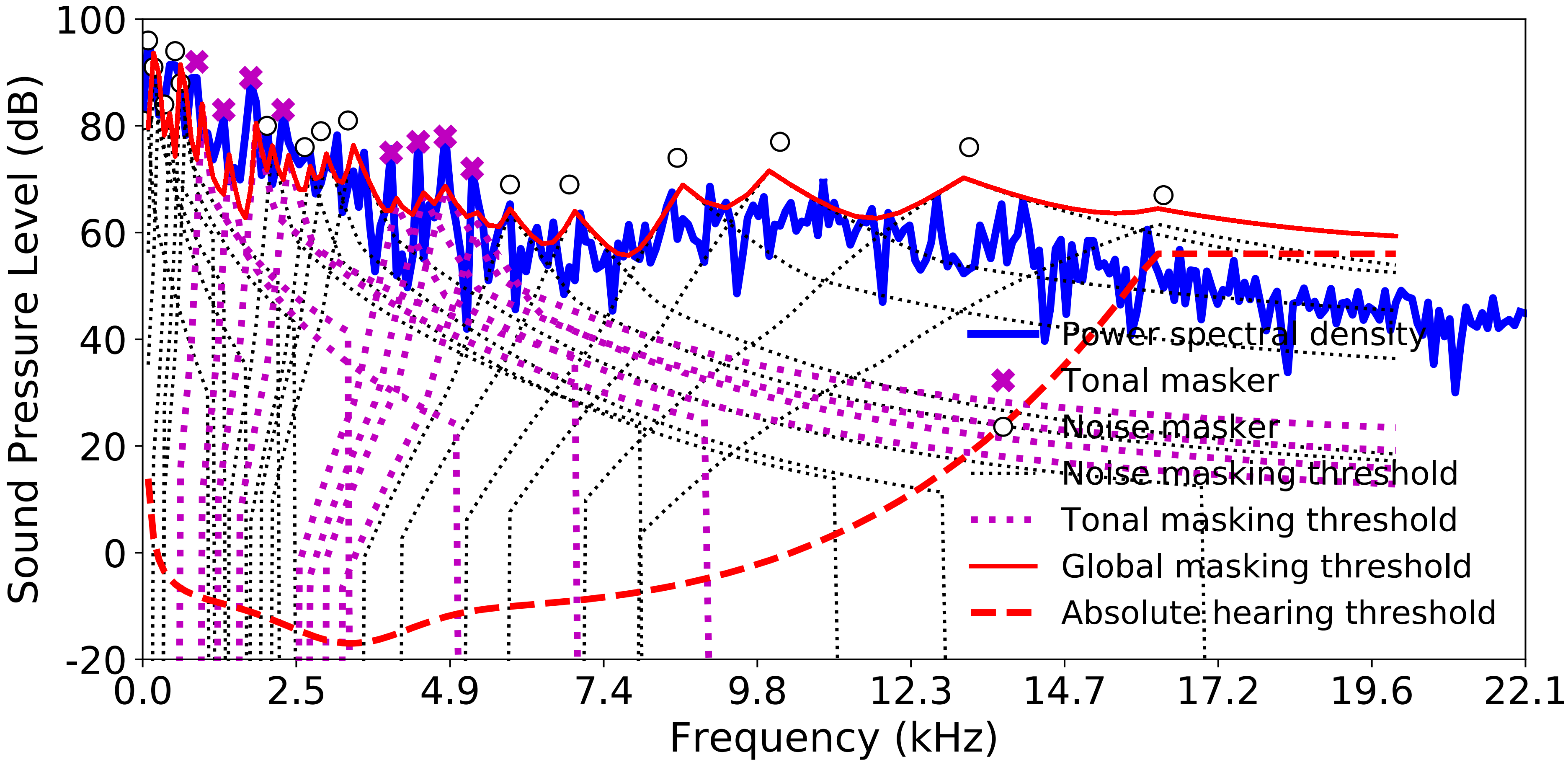
○ Psychoacoustics for MPEG audio coding technology (simultaneous masking)



No reason to spend many bits to encode these tones!

Perceptual Nature of Audio

- Our PAM-1 implementation (in TensorFlow)



Perceptual Nature of Audio

- Psychoacoustic loss for neural audio coding

- Priority weighting $\mathcal{L}_3(\boldsymbol{s}||\hat{\boldsymbol{s}}) = \sum_i \sum_f w_f \left(x_f^{(i)} - \hat{x}_f^{(i)} \right)^2$

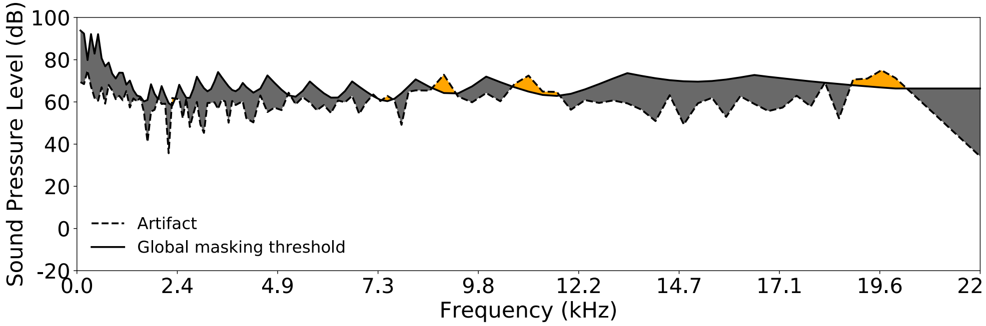
PAM weights

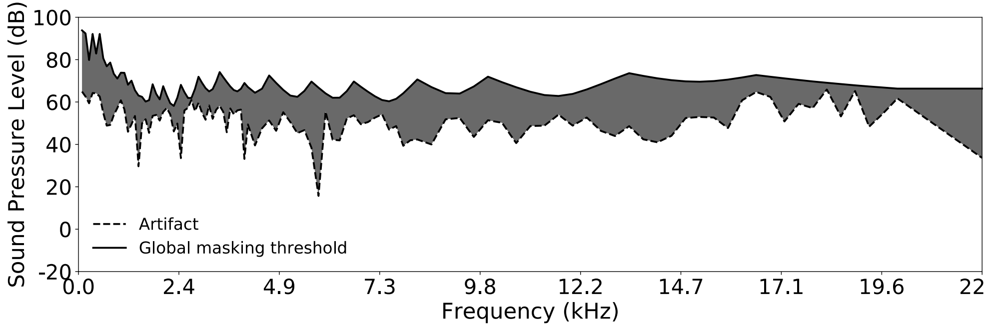
Log PSD

$$\boldsymbol{w} = \log_{10} \left(\frac{10^{0.1\boldsymbol{p}}}{10^{0.1\boldsymbol{m}}} + 1 \right)$$

Mask

  - Allows error in the masked area
    - Can reduce bitrates; Can reduce model sizes
- Noise modulation
  - Iteratively penalizes the highest NMR $\mathcal{L}_4 = \max_f \left(\mathrm{ReLU} \left(\frac{n_f}{m_f} - 1 \right) \right).$

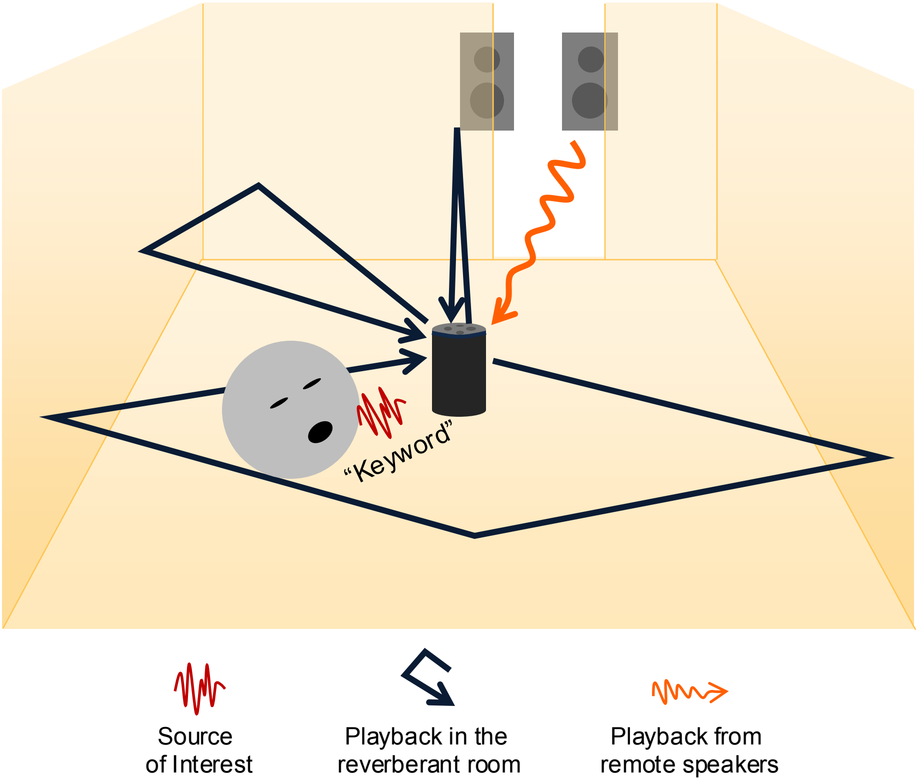




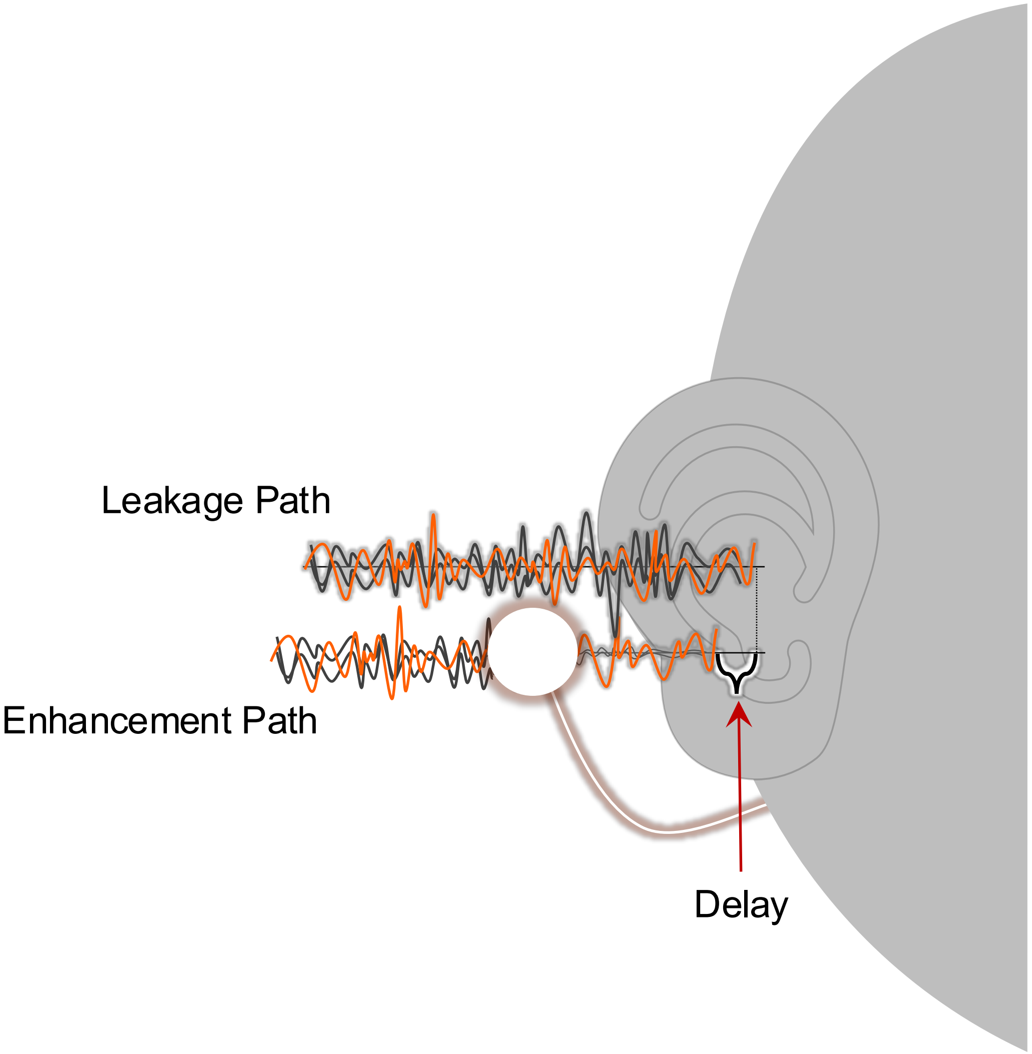
- Competes with MP3

Acoustic Echo Cancelation
Active Noise Cancellation
Spatial Audio

AEC and ANC



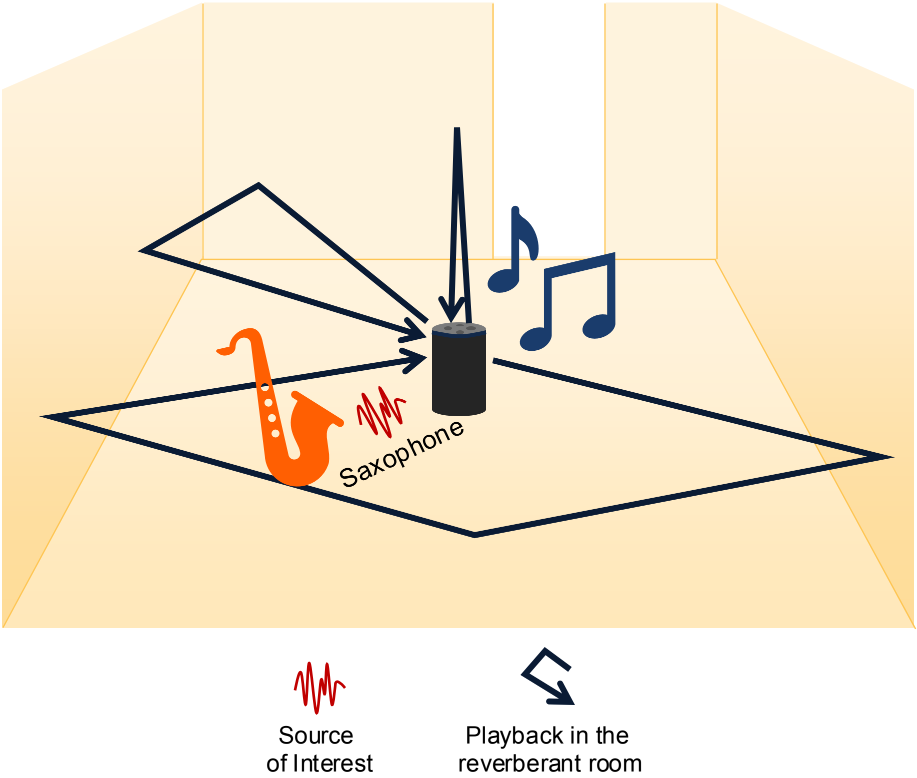
"Keyword"
Source
of Interest
Playback in the
reverberant room
Playback from
remote speakers

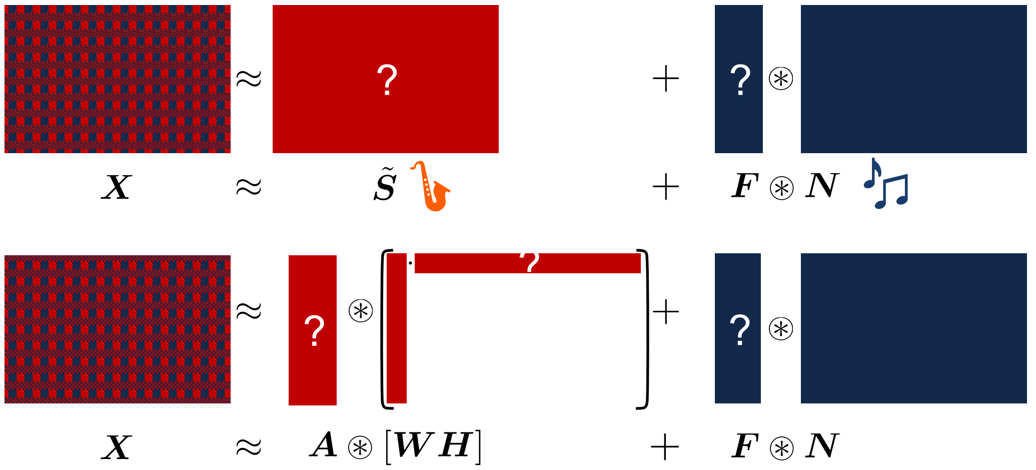


Leakage Path
Enhancement Path
Delay

AEC and ANC

- Music-version of AEC

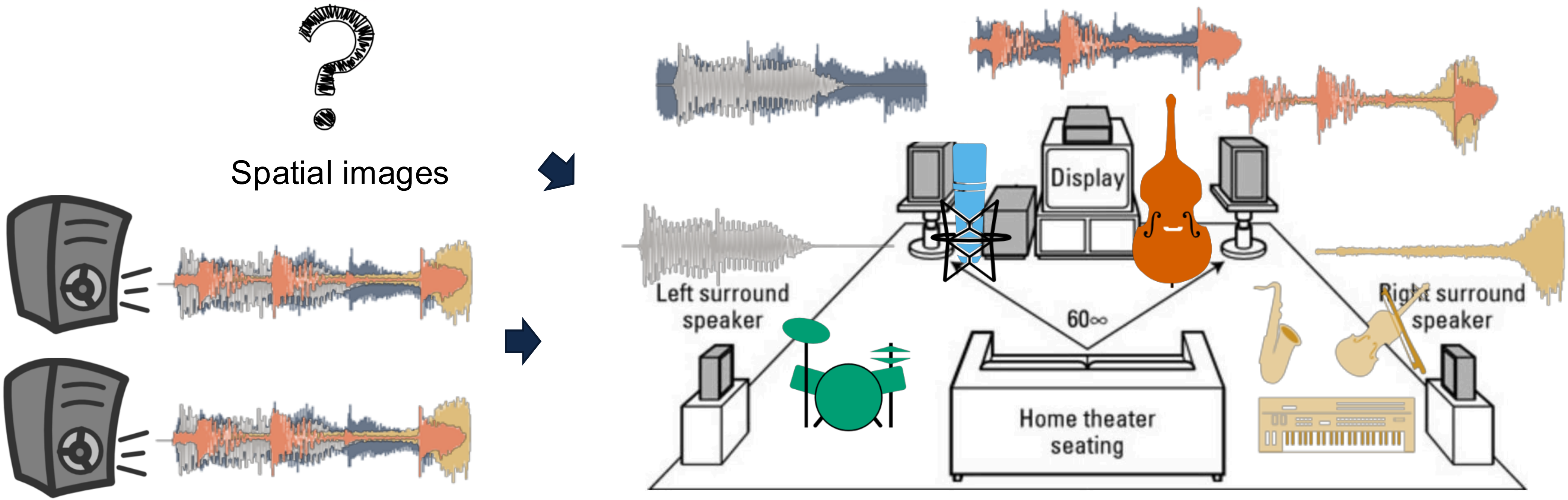




Spatial Audio

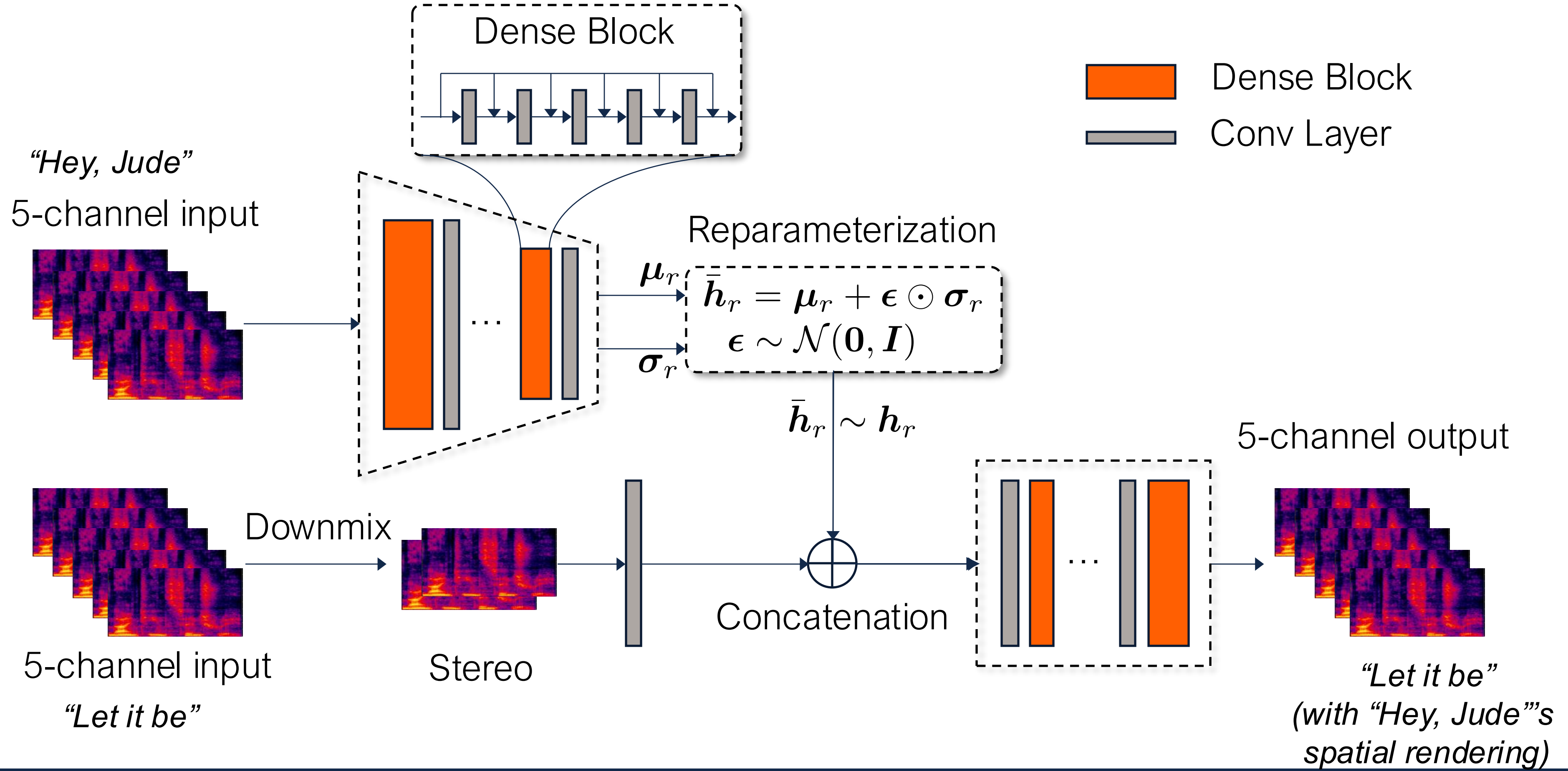
- Stereo to surround extension

- Music upmixing is not well defined
- Our goal: to disentangle music and spatial information in the latent space



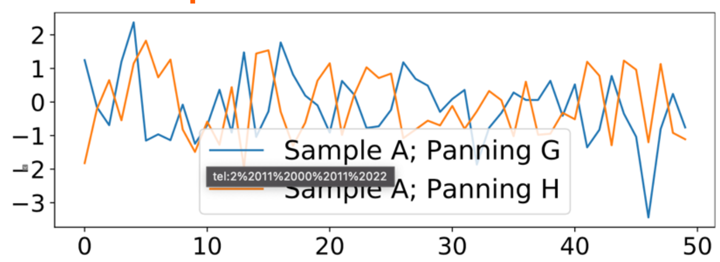
Spatial Audio

- Disentanglement in the latent space: a VAE-based approach

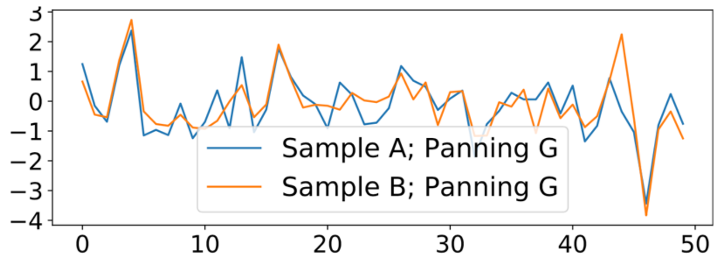


Upmixing via Style-Transfer

- Latent space visualization

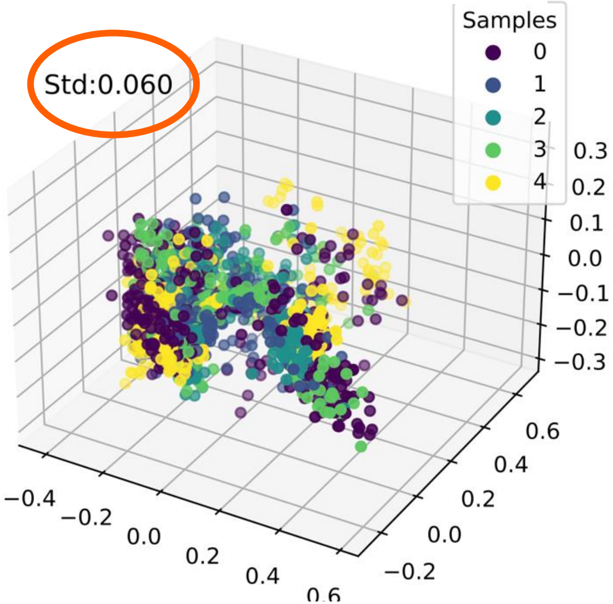


(a) Same music, different spatial map

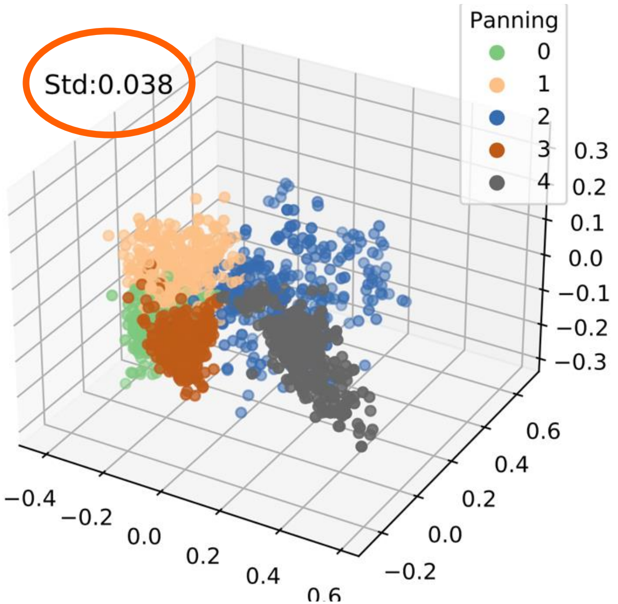


(b) Different music, same spatial map

A line represents a 50-dimensional latent vector.



(c) Colored by music content



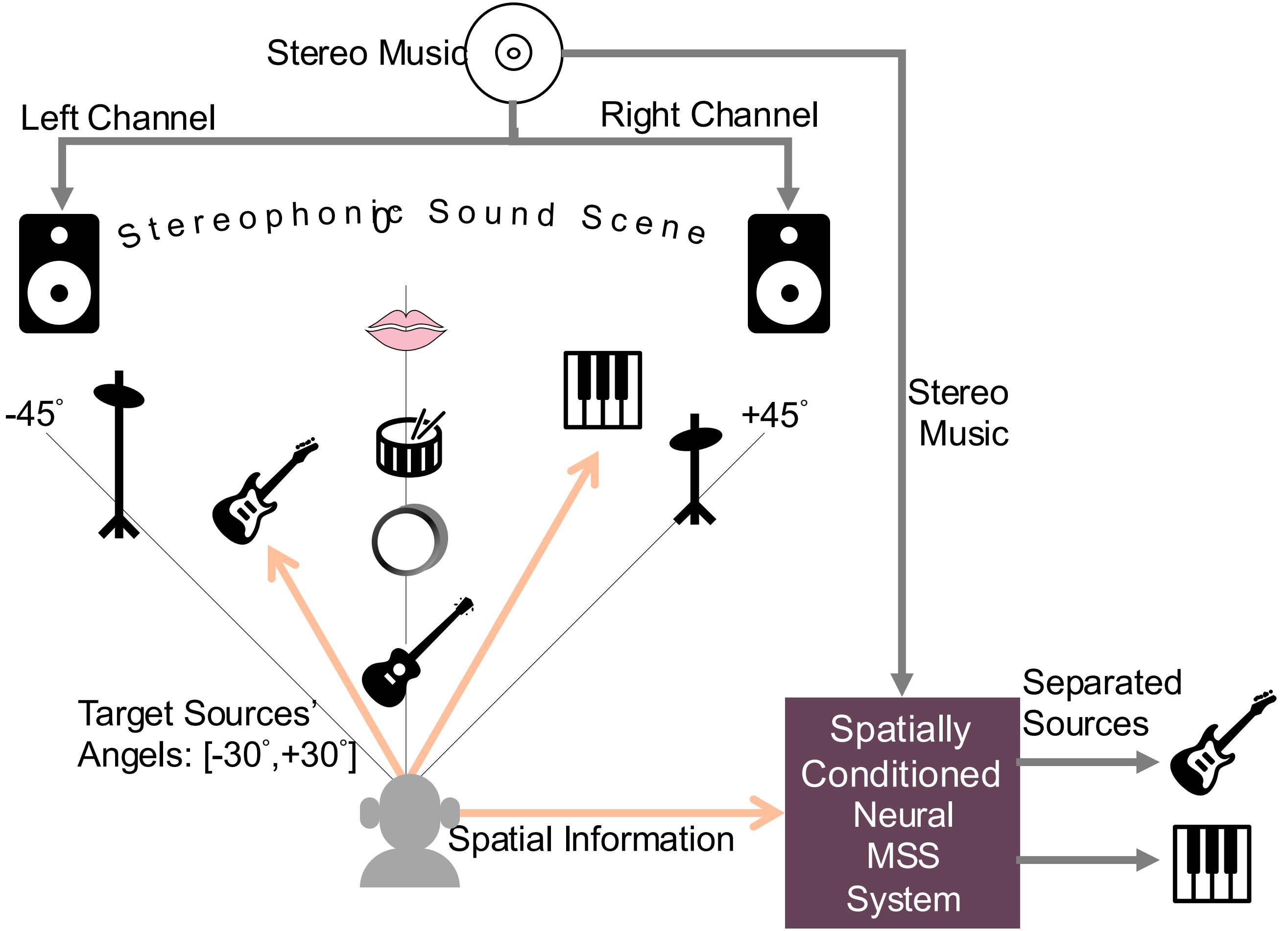
(d) Colored by spatial images

Each dot represents a dimension-reduced latent vector

Music Signal Processing

SpaIn-Net

- Spatially-Informed Stereophonic Music Source Separation



Spaln-Net
Audio Examples

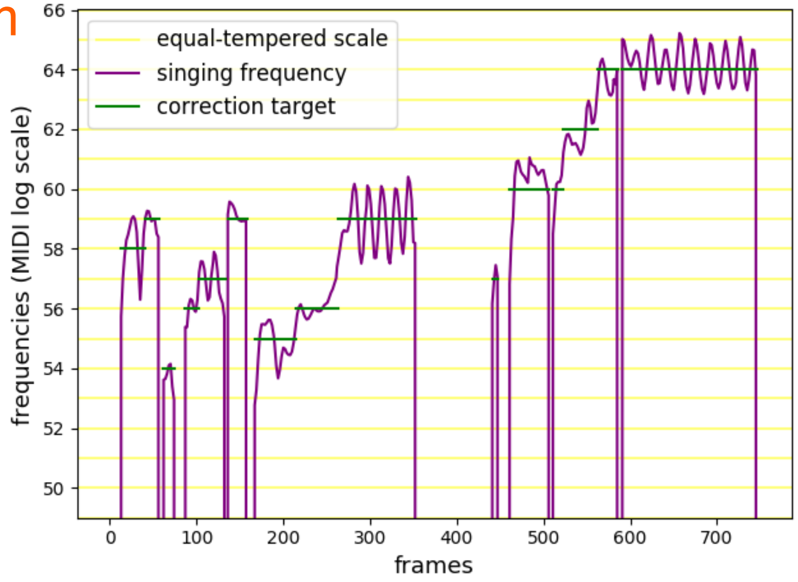
Input Mixture:

	Gtr1	Gtr2	Piano	Bass
Ground-Truth				
XUMX Baseline				
D1-CAT (Proposed)				

Neural Pitch Correction?

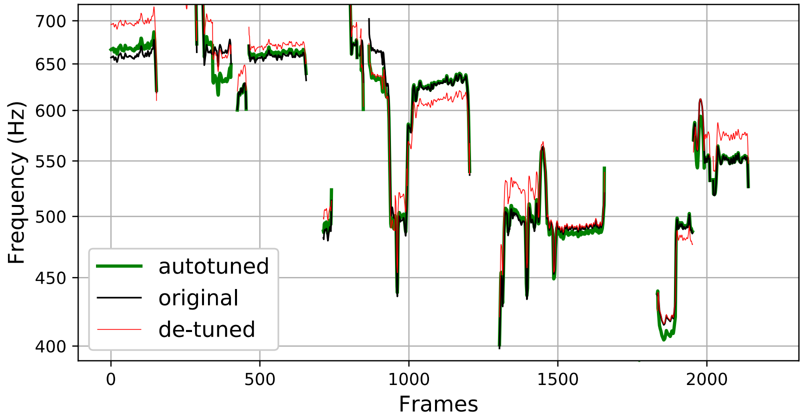
- It's a data-intensive regression problem

- Traditional autotuners:
  - Require specifying the pitches of the melody beforehand
  - Snap pitches to a grid
  - Robotic and musically limiting



Lights, Ellie Goulding performed by Smule, Inc. user applying Auto-Tune effect

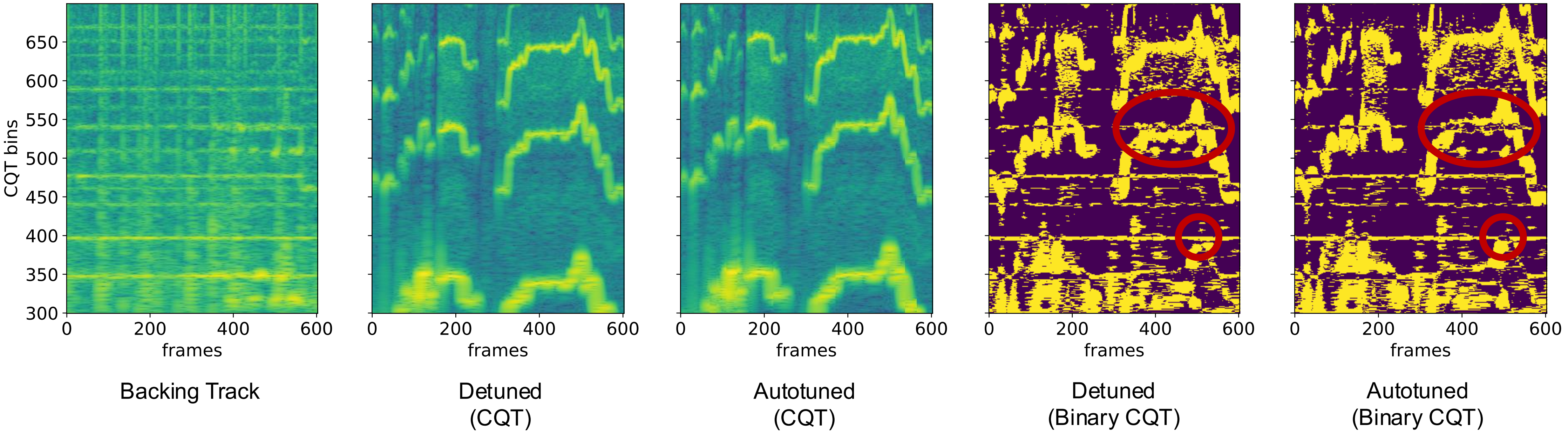
- Proposed approach:
  - Doesn't require reference pitches
    - Uses backing and vocal track overtones
  - Preserves nuances while detecting unintended pitch shifts



The Neural Pitch Correction System

- Input to the system

- Input consists of three CQT spectrograms
  - Backing track
  - De-tuned singing voice
  - The mismatch of the binarized CQT



Experimental Results

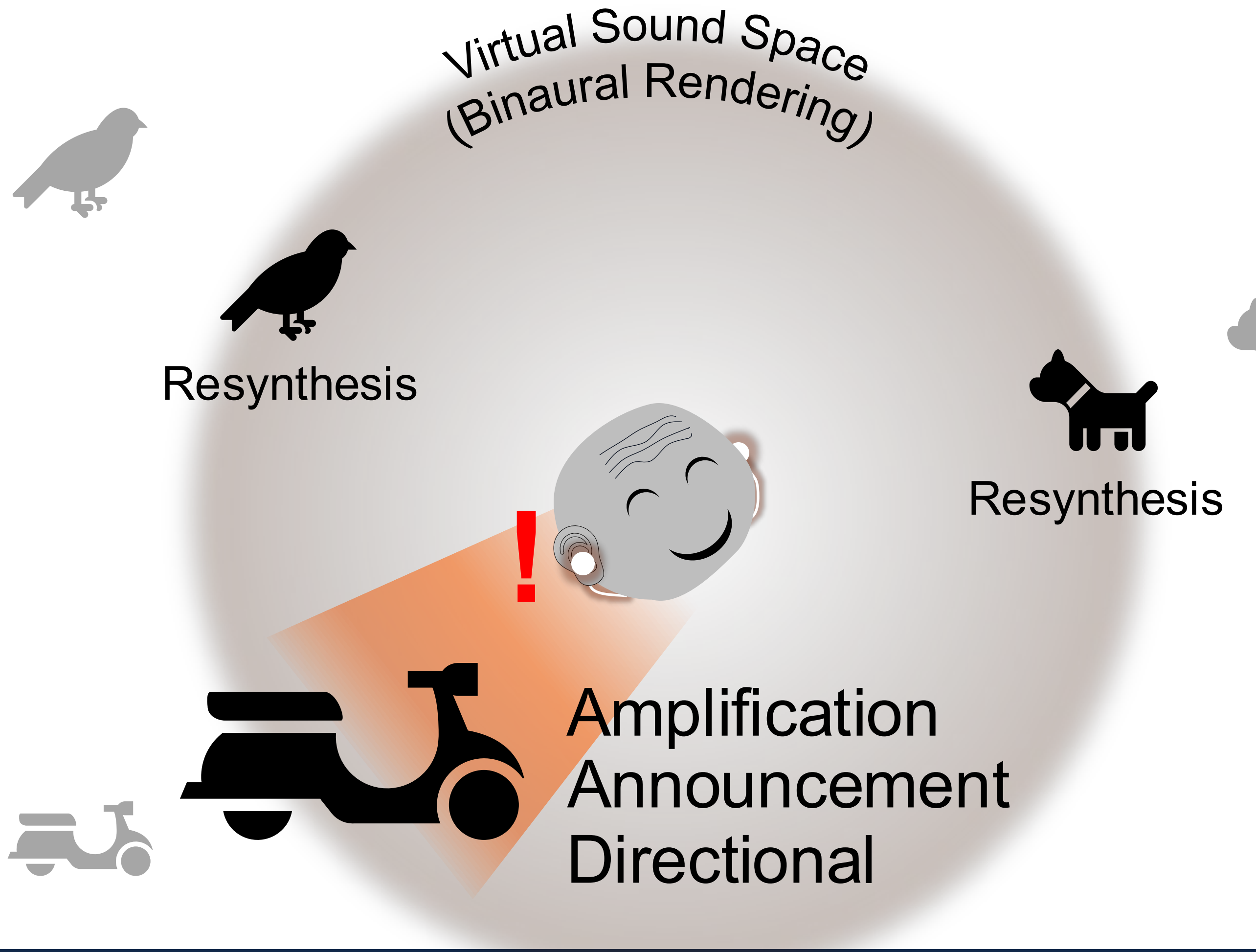
- Sound examples

- On artificially detuned test signals
  - Original intonation sample→ detuned version → autotuned version

- Real-world examples

Detection and Classification of Acoustic Scenes and Events

DCASE



Event Detection

Anomaly Detection

Source Tracking

Reference

- 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] [WASPAA 2021 supplementary material: code, demo, presentation video]
- Aswin Sivaraman and *Minje Kim*, "**Efficient Personalized Speech Enhancement through Self-Supervised Learning**," ***IEEE Journal of Selected Topics in Signal Processing***, vol. 16, no. 6, pp. 1342-1356, Oct. 2022 [pdf, demo, presentation video]
- *Minje Kim*, Paris Smaragdis, Glenn G. Ko, and Rob A. Rutenbar, "**Stereophonic Spectrogram Segmentation Using Markov Random Fields**," in Proceedings of the *IEEE International Workshop on Machine Learning for Signal Processing (**MLSP**)*, Santander, Spain, Sep. 23-26, 2012 [pdf, demo, bib]
- Kai Zhen, Mi Suk Lee, Jongmo Sung, Seungkwon Beack, and *Minje Kim*, "**Psychoacoustic Calibration of Loss Functions for Efficient End-to-End Neural Audio Coding**," ***IEEE Signal Processing Letters***, vol 27, pp. 2159-2163, 2020. [pdf, demo, code, presentation video]
- Haici Yang, Sanna Wager, Spencer Russell, Mike Luo, Minje Kim, and Wontak Kim, "**Upmixing Via Style Transfer: a Variational Autoencoder for Disentangling Spatial Images and Musical Content**," ***ICASSP*** 2022 [pdf, demo, presentation video].
- Darius Petermann and Minje Kim, "**SpaIn-Net: Spatially-Informed Stereophonic Music Source Separation**," ***ICASSP*** 2022 [pdf, demo, code, presentation video].
- Sanna Wager, George Tzanetakis, Cheng-i Wang, and Minje Kim, "**Deep Autotuner: A Pitch Correcting Network for Singing Performances**," ***ICASSP*** 2020 [pdf, demo, code, presentation video]

Thank You!

Minje Kim, Ph.D.
Associate Professor
Dept. of Computer Science
https://minjekim.com
minje@illinois.edu