Musical Source Separation Using Time-Frequency Source Priors

Presented by
Adam Cohen

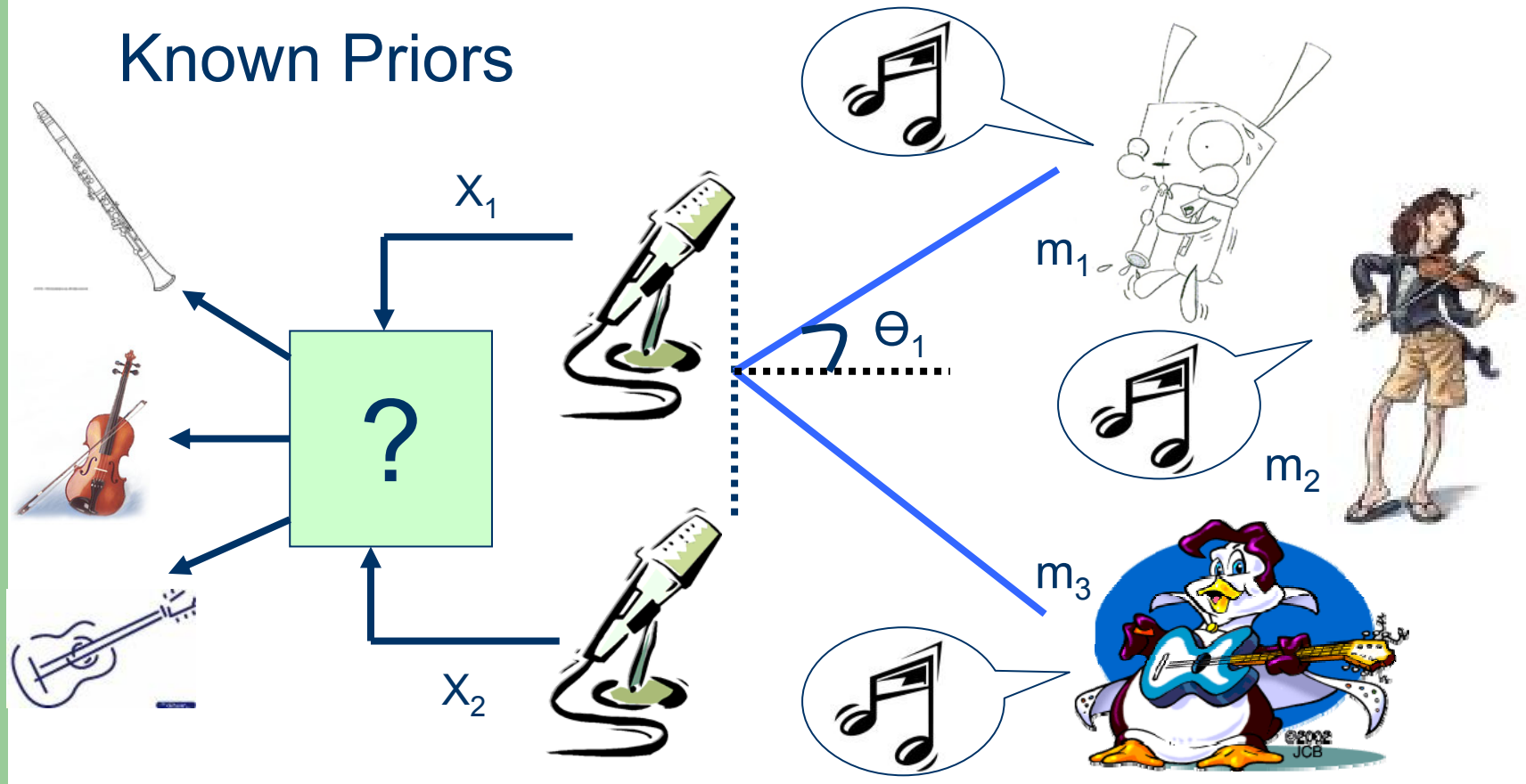
Scope

- Paper Review for “Speech Processing in Noisy Environment” course.
- “*Musical Source Separation Using Time-Frequency Source Priors”* by Emmanuel Vincent

Agenda

- Problem Definition
- Introduction
- Solution Overview
  - Model Presentation
  - Model Estimation
  - Instrument Parameter Learning
  - Filtering
- Algorithm Evaluation
- Conclusion

Problem Definition



Introduction

- Applications
- Main Problems
- Existing Methods
  - CASA
  - Statistical Spatial models
  - Statistical Spectral models

Introduction - Basic Assumptions

- Known instrument and locations (priors)
- Two microphone (technical)
- Channel Model:

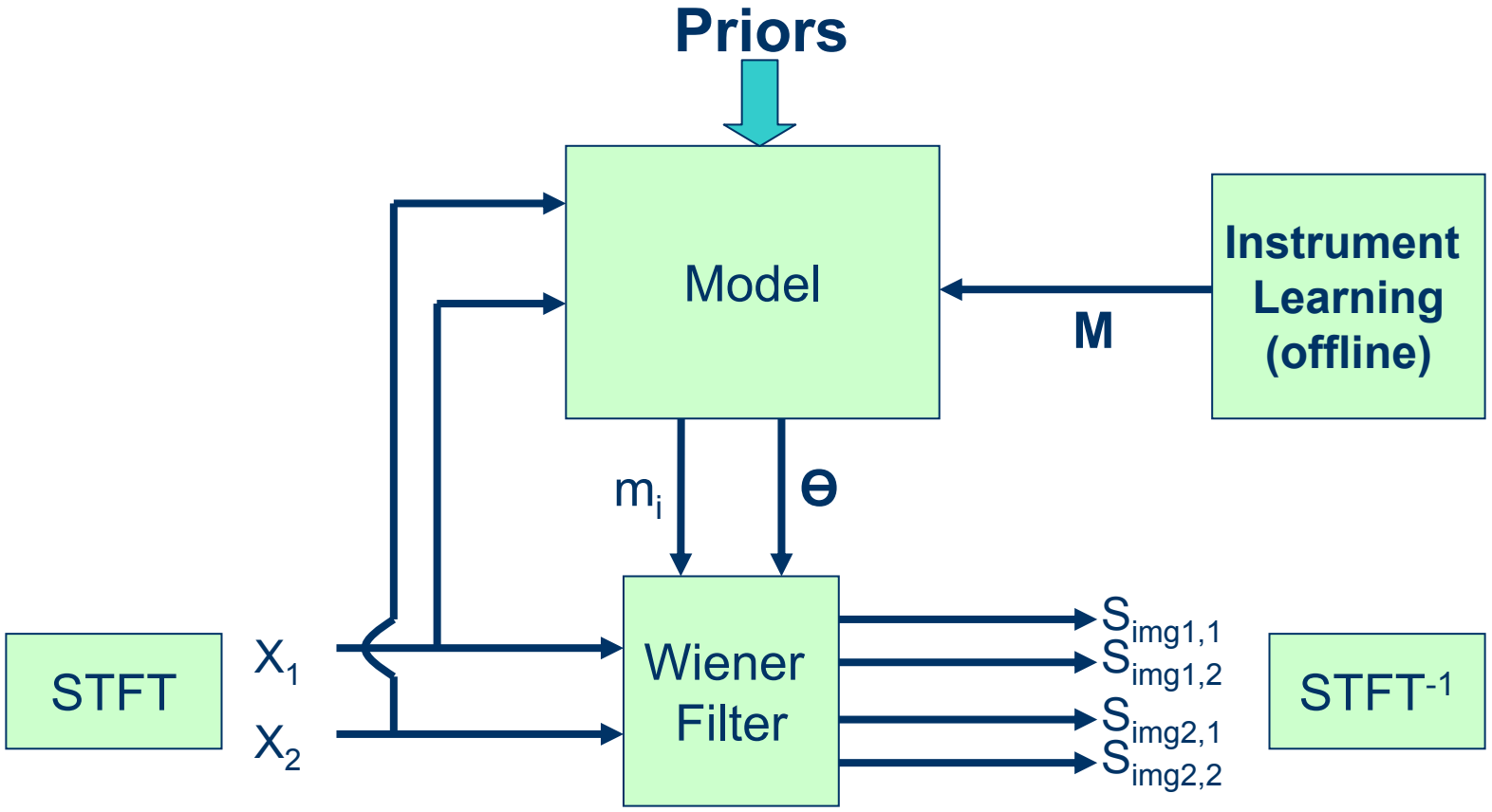
$$x_i = \sum_{j=1}^{J} a_{i,j} * s_j + n_i$$

  - J number of instruments
  - n_i is Gaussian noise
- Analysis in the STFT domain

Solution Overview

- Source Modeling for spectrum estimation
- Model require parameter learning
  - Offline instrument parameter learning
  - Mixture modeling to estimate instrument state
- Source filtering for separation

Solution Overview



Model Presentation

- Target:
  - Estimate instrument (source) spectrum
  - Estimate environment parameters
  - Estimation is used for filtering
- Means
  - Three Layer Bayesian Model
  - Learned Parameters
  - MAP
- Input: X_i

Model Presentation - Overview

State

Active Note

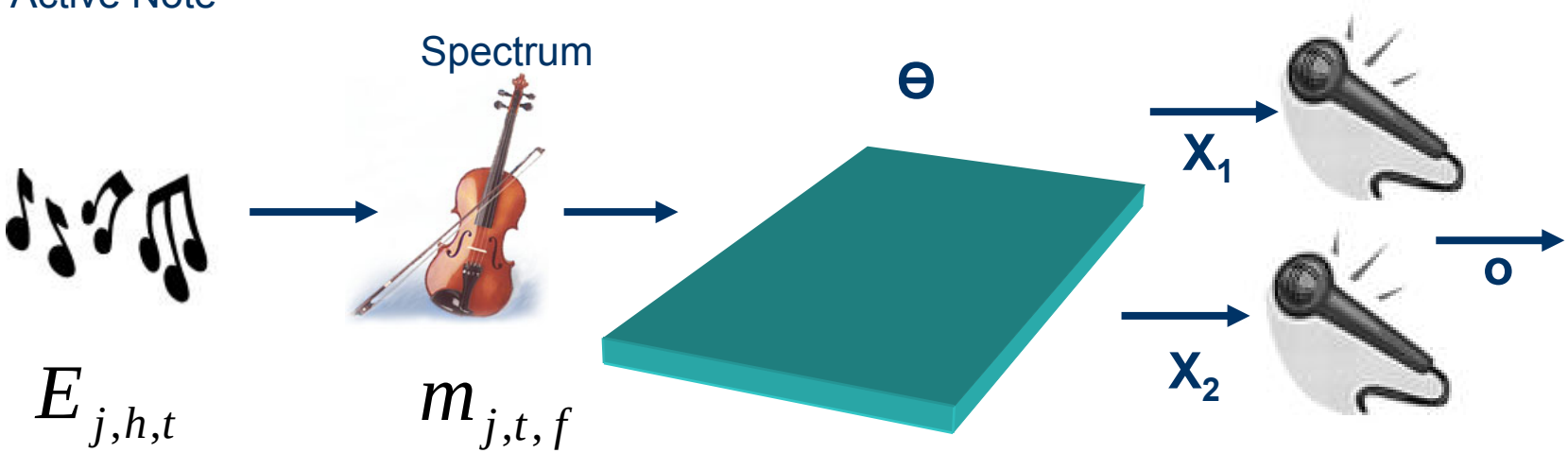
Source

Instrument

Spectrum

Mixture

Environment



$E_{j,h,t}$

$m_{j,t,f}$

$$P^{sta} = p(E_{j,h,t}) \qquad P_t^{src} = p(m_{j,t,f} \mid E_{j,t}, M) \qquad P_t^{mix} = p(o_t \mid m_{j,t}, \Theta)$$

Model Presentation - Preprocess

- Input:

$$o^{pow} = \log\left(\frac{\|x_1\|^2 + \|x_2\|^2}{g_f} + 1\right)$$

$$o^{pha} = \prec\langle x_1, x_2 \rangle$$

$$o^{coh} = \frac{\left|\langle x_1, x_2 \rangle\right|}{\|x_1\| \cdot \|x_2\|}$$

- Filter Bank: logarithmic bandwidth

Model Presentation - State

- Target: probability of active note

$$P^{sta} = p(E_{j,h,t})$$

- Factorial Model:
  - State is independent Bernoulli Process

$$P^{sta} = \prod_{j=1}^{J} \prod_{t=0}^{T-1} (1-Z)^{\#A_{j,t}} Z^{\#I_{j,t}}$$

  - Z is fixed
  - Higher Z implies high probability for low number of active note

- Segmental Model 

Model Presentation - Source

- Target: Probability of spectra given State
- Assumptions:
  - Different note are uncorrelated
  - Note Spectrum is stationary
  - Active note are log-Gaussian independent

$$m_{j,t,f} = \sum_{h=H_j} e_{j,h,t} \Phi_{j,h,f}$$

$$P_t^{src} = p(m_{j,t,f} \mid E_{j,t}, \Phi_{j,h}, \mu_{j,h}^e, \sigma_{j,h}^e) = \prod_{j=1}^{J} \prod_{h \in A_{j,t}} N(\log e_{j,h,t}; \mu_{j,h}^e, \sigma_{j,h}^e)$$

- Φ the spectrum of note (instrument learn)
- μ,σ Model parameters (instrument learn)

Model Presentation - Mixture

- Target: Channel model – $P_t^{mix} = p(o_t \mid m_{j,t}, n, a^{pow})$
- Mixing filters are modeled via a^{pow} & a_j^{pha}
  - a_j^{pha} phase response is known from the direction:

$$a_j^{pha} = 2\pi \sin\theta_j \bmod 2\pi$$

  - a^{pow} power response - environment parameter (estimate)
  - Experiments shows need to estimate $\mathbf{a^{phase}}$
  - n noise spectrum - environment parameter (estimate)

- Mono Model
- Stereo Model

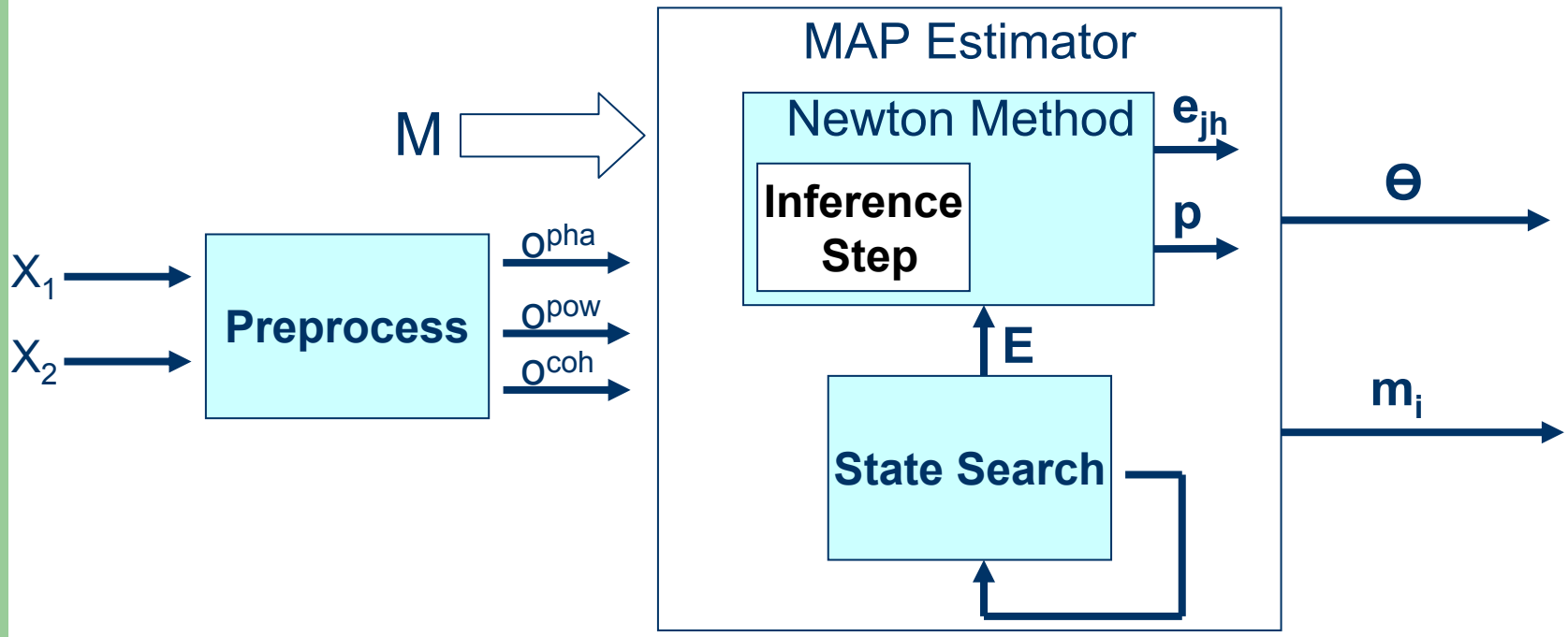
Model Estimation

- Problem: $\hat{s}_{i,j} = \arg\max P\left(s_{i,j} \mid x_i, \Theta, M\right)$
  - Integral can't be solved
- Approximation:
  - Algorithm target $\hat{s}_{i,j} \approx \arg\max P\left(s_{i,j} \mid x_i, \hat{\Theta}, \hat{m}\right)$
  - **Model target** $\hat{\Theta}, \hat{m}_{jt} = \arg\max P\left(\Theta, m_{jt} \mid o_t, M\right)$
- Weighted Bayes Law:

$$P^{tot} = P\left(o_t, m_{jt}, E_{jt}, \Theta, M\right) \propto \left(\prod_{t=0}^{T-1} P_t^{mix}\right)^{w_{mix}} \left(\prod_{t=0}^{T-1} P_t^{src}\right)^{w_{src}} P^{sta}$$

Model Estimation - Blocks

Known Instruments and directions



Model Estimation – Inference Step

- For each given state estimate best spectrum
- Estimate source spectrum by note power (e)
  - MAP is solved by Iterative Newton method:

$$\log e_{jht} \leftarrow \log e_{jht} + \Delta \log e_{jht}$$

  - ΔLog(e) is calculated by derivative of P^{tot}
- Spectrum is built using Φ
- Similar Method for Environment parameters

Model Estimation – State Search

- Target: Search $\mathbf{E}_{jht}$ for optimization
- State Space is to large (more then 10^8)
- Reducing the state search by heuristic
  - Search near previous state
  - Estimate state using anther method
  - Using beam search methods
  - Finding inactive state

Instrument Parameters Learning

- Target: learn $M=(\Phi,\mu^e,\sigma^e)$ for instrument
  - ML
  - Approximation:

$$\hat{M} = \arg\max P\left(M \mid o_t, \Theta, E_{jt}\right)$$

$$\hat{M} \approx \arg\max P\left(M \mid o_t, \Theta, \hat{m}_{jt}\right)$$

$$\hat{m}_{jt} = \arg\max P\left(m_{jt} \mid o_t, \Theta, , E_{jt}, M\right)$$

  - Solve by expectation maximization (EM) algorithm
- Offline learning from database
  - Single note learn

Filtering

- Motivation
  - Spectrum estimation is unnatural
  - Solving $\hat{s}_{i,j} \approx \arg\max P\left(s_{i,j} \mid x_i, \hat{\Theta}, \hat{m}\right)$
- Wiener Filter is used:

$$s_{img}(i,j) = \frac{a^{pow}\hat{m}_j}{\sum_{j=1}^{J} a^{pow} \cdot \hat{m}_j + n_f} x_i$$

- Every instrument have two outputs

Algorithm Evaluation

- Instrument parameters learn from music database
- Two instruments in two test Scenarios:
- Criteria:
  - SDR: Signal to Distortion Ration
  - SIR: Signal to Interference Ration
  - SAR: Signal to Artifact Ration
  - Human judgment
- First time for musical source separation ?

Algorithm Evaluation – Result Mix 1

Mixture	Clarinet (dB)			Violin (dB)		
	SDR	SIR	SAR	SDR	SIR	SAR
ICA	-1	4	4	-9	-3	3
Spatial Masking	4	15	4	1	6	8
Mono factorial	14	25	16	9	21	12
Mono segmental	18	37	19	14	28	15
Stereo factorial	15	28	16	11	25	12
Stereo segmental	16	37	16	13	28	15

Algorithm Evaluation – Result Mix 2

Mixture	Cello (dB)			Violin (dB)		
	SDR	SIR	SAR	SDR	SIR	SAR
ICA	-3	7	0	-3	2	4
Spatial Masking	4	14	5	5	12	7
Mono factorial	-17	12	-15	1	1	26
Stereo factorial	7	26	7	7	11	13
Stereo segmental	13	39	14	16	34	17

Conclusion

- The paper doesn’t suggest new approach
  - A new application for exciting methods
- The paper achieve good result on source separation
  - Doesn’t use all space information
  - Test scenarios aren’t complete
- Possible improvements
  - Note estimation
  - Online instrument modifications
  - More history/future use
- New approach is needed for complete solution
- Single sensor maximum performance

References

- E. Vincent "*Musical Source Separation Using Time-Frequency Source Priors*"
- T. Kinoshita, S. Saki and H. Tanaka "*Musical Sound Source identification based on frequency component adaptation*"
- L. Benaroya, R. Gribonval, and F. Bimbot, "*Non negative sparse representation for wiener based source separation with a single sensor*"
- L. Benaroya, Frederic Bimbot, Remi Gribonval "*Audio Source Separation With a Single Sensor*"
- E. Vincent and X. Rodet "*underdetermined source separation with structured source priors*"
- E. Vincent, R. Gribonval and C. Fevotte *"Performance Measurement in Blind Audio Source Separation"*

Model Presentation - State

- Segmental Model
  - Using prior temporal persistence
  - Markov chain model
  - No chords

$$P^{\text{sta}} = \prod_{j=1}^{J} \prod_{r=0}^{R_j-1} \mathcal{D}^{\text{not}}(d_r) \prod_{r=0}^{R_j-1} \mathcal{D}^{\text{seg}}(t_{r+1}-t_r)(1-Z)^{\#\mathcal{A}_{j,0}} Z^{\#\mathcal{I}_{j,0}} \times \prod_{r=1}^{R_j-1} (\#\mathcal{I}_{j,t_r}-1)^{-1}$$

Where

$$\mathcal{D}^{\text{not}}(d) = \begin{cases} \dfrac{\mathcal{N}(\log d; \mu^n, \sigma^n)}{\sum_{d' \geq d^n} \mathcal{N}(\log d'; \mu^n, \sigma^n)}, & \text{if } d \geq d^n \\ 0, & \text{otherwise} \end{cases}$$

Model Presentation – Mixture Mono

- Mono Model
  - Ignore phase information

$$o_t^{pow} = \log\left(\sum a^{pow} m_{jt} + n_f\right) + \varepsilon_t^{pow}$$

  - ε is the residual error: Gaussian
  - Fix std $\sigma^{\varepsilon pow}$

$$P_t^{mix}(\varepsilon_t^{pow}) \propto N(0, \sigma^{epow})$$

Model Presentation – Mixture Stereo

- Consider phase difference & azimuth cue
  - Different source are uncorrelated
  - Require estimate of noise phase

$$o_{tf}^{\text{pha}} = \angle\left(\sum_{j=1}^{J} a_f^{\text{pow}} m_{jtf} \exp\left(i a_{jf}^{\text{pha}}\right) + n_f \exp\left(i b_f^{\text{pha}}\right)\right) + \epsilon_{tf}^{\text{pha}} \mod 2\pi$$

$$\sigma_{tf}^{\epsilon\text{pha}} = \sigma^{\epsilon\text{pha}}\left(1 - o_{tf}^{\text{coh}}\right)^{\lambda^{\text{pha}}}$$

  - Low coherence less phase information

$$P_t^{\text{mix}} = \prod_{f=1}^{F-1} \mathcal{N}\left(\epsilon_{tf}^{\text{pow}}; 0, \sigma^{\epsilon\text{pow}}\right) \frac{\mathcal{N}\left(\epsilon_{tf}^{\text{pha}}; 0, \sigma_{tf}^{\epsilon\text{pha}}\right)}{\int_{-\pi}^{\pi} \mathcal{N}\left(\epsilon; 0, \sigma_{tf}^{\epsilon\text{pha}}\right) d\epsilon}$$

Model Presentation – Inference Step

- Derivative:

$$\frac{\partial \log P^{\text{tot}}}{\partial \log e_{jht}} = \frac{w_{\text{mix}}}{\sigma^{\epsilon\text{pow}2}} \sum_{f=0}^{F-1} \epsilon_{tf}^{\text{pow}} \pi_{jhtf} - \frac{w_{\text{src}}}{\sigma_{jh}^{e2}} \left(\log e_{jht} - \mu_{jh}^{e}\right)$$

- For the mono model

$$\Delta \log e_{jht} = \frac{\frac{w_{mix}}{\sigma^{\varepsilon pow2}} \varepsilon_t^{pow} \pi_{jht} - \frac{w_{src}}{\sigma_{jh}^{e2}} \left(\log e_{jht} - \mu_{jh}^{e}\right)}{\frac{w_{mix}}{\sigma^{\varepsilon pow2}} \pi_{jht} + \frac{w_{src}}{\sigma_{jh}^{e2}}}$$

Where

$$\pi_{jhtf} = \frac{a_f^{\text{pow}} e_{jht} \Phi_{jhf}}{\sum_{j=1}^{J} a_f^{\text{pow}} m_{jtf} + n_f}$$

- When note is masked π = 0: o is irrelevant