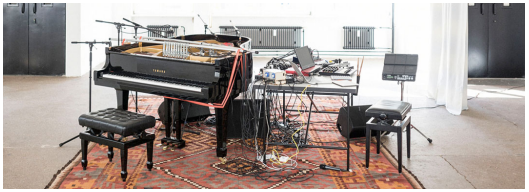
Deep Bayesian Inference for Audio Processor Parameter Estimation

Dominique Fourer
Joint work with Côme Peladeau and Geoffroy Peeters

IBISC Laboratory, Université Évry Paris-Saclay

20th of july, 2026
IEEE ARM 2026 Workshop @ Évry-Courcouronnes

Audio Quality (AQ) - Scientific Problem



What is audio quality?

As instrumental sound "timbre" is defined as all sound characteristics not related to pitch, loudness and duration, "audio quality" is everything related to the sound characteristics not related to the content sources.

Motivation:

- Detecting AQ is full of interest for audio indexing systems (AQ is related to listening experience)
- AQ can have an impact on the prediction accuracy of music information retrieval systems
- AQ results from recording conditions and studio practices
- Controlling AQ can enable robust machine learning methods (ie. data augmentation, models invariant to unwanted signal transformations)

The AQua-rius project - ANR-22-CE23-0022

- **Period:** from **jan. 2023** to **feb 2027**
- **Fundings:** 510 157.90 €
- **Partners:** 3 labs: IBISC (Univ. Evry) / UMR STMS (IRCAM) / LTCI (TelecomParis)
- **Coordinator:** D. Fourer (IBISC, SIAM)
- **Website:** `https://fourer.fr/aquarius`

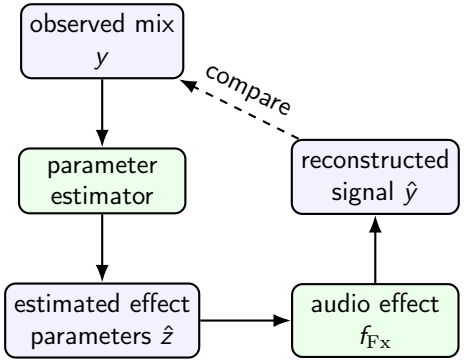
Goals

- **Exhaustive investigation of "audio quality": analysis/modeling with a focus to prediction/detection of the audio effects** applied during the audio signal production and diffusion chain
- Audio-quality control, including audio-effect removal or inversion, signal enhancement, and audio restoration.
- Simulation and synthesis of audio quality with possible applications to data augmentation and domain adaptation techniques using machine-learning-based methods

Why estimate audio effect parameters from a mix?

Given an observed processed or mixed signal y, we seek to infer the parameters z of the audio effect that produced it.

- **Automatic mixing:** reproduce or adapt an existing processing chain;
- **Preset retrieval:** recover interpretable plugin settings from a reference sound;
- **Audio restoration:** identify and compensate for unwanted processing;
- **Sound matching:** configure an effect or synthesizer to approach a target sound;
- **Content analysis:** describe production choices through meaningful parameters.



Challenge:

Several parameter vectors may explain the same observed signal.

1. Problem Formulation
2. Proposed Approach
3. Numerical Results
   - Experiment 1: estimating equalizer parameters
   - Experiment 2: FM synthesizer sound matching
4. Conclusion

Outline

Audio processors as forward models

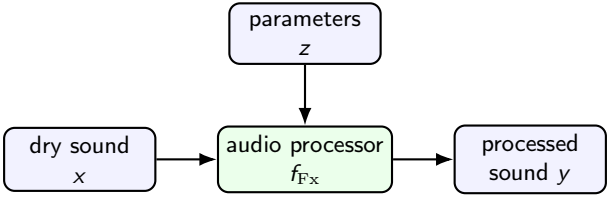
Forward problem

Given an input sound x and processor parameters z, synthesize or transform audio:

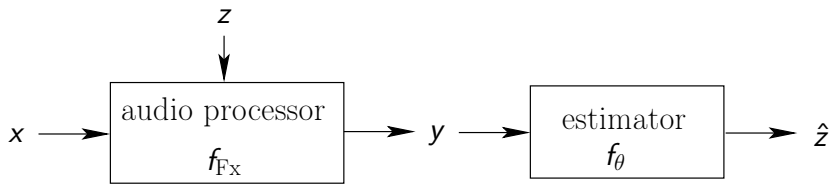
$$y = f_{\mathrm{Fx}}(x, z).$$

- x: Input discrete time audio signal ($x \in \mathbb{R}^N$)
- z: Audio processors parameter vector ($z \in \mathbb{R}^d$, $d \ll N$)
- y: Ouput processed signal ($y \in \mathbb{R}^N$)

- effects: EQ, compressor, distortion, reverb, etc.
- synthesizers: FM, subtractive, wavetable, etc.



Audio processor parameter estimation as an ill-posed inverse problem



Goal:

We aim to construct an estimator f_θ that predicts z from the observation y, such that

$$f_\theta(y) = \hat{z}$$

Problem:

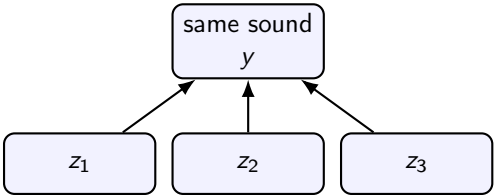
f_{Fx} is not injective $\Rightarrow$ given $x \in \mathbb{R}^N$, $\exists z' \neq z$, s.t. $f_{\mathrm{Fx}}(x, z') = f_{\mathrm{Fx}}(x, z)$
There exist several parameter vectors z that produce the same (perceptually) output sound y.

Hadamard definition: A well-posed problem satisfies 3 conditions: existence, **uniqueness**, and stability of the solution (continuous dependence of the solution on the data).

Why the inverse mapping is non-injective

Multiple parameter vectors can produce almost indistinguishable sounds:

- **Equalization:** gain changes in one band can be compensated by opposite changes in overlapping bands.
- **FM synthesis:** for low modulation indices, several modulation parameters have little or no audible effect.
- **Dynamic processing:** different combinations of threshold and ratio may produce nearly identical compression characteristics.



one observation,
many explanations

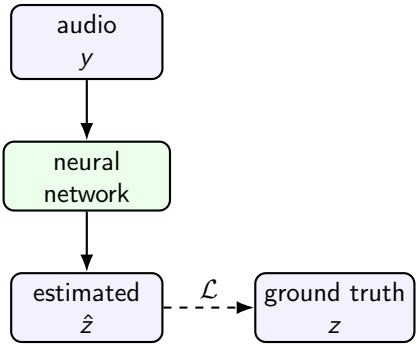
Consequence

A deterministic estimator hides ambiguity and uncertainty.

Baseline: supervised deterministic parameter regression

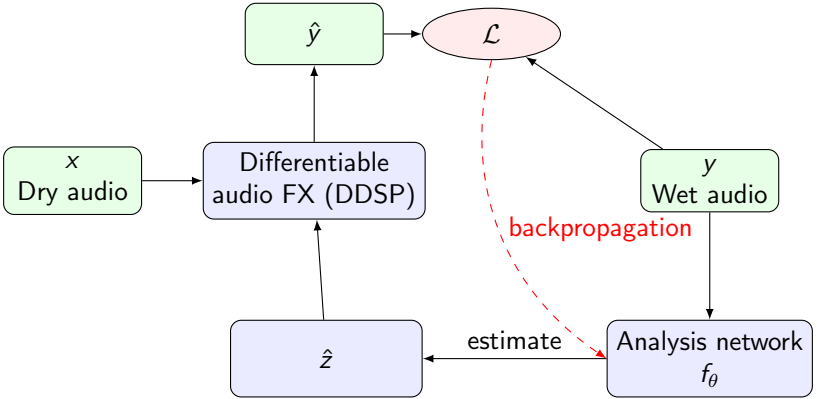
- Generate pairs (y, z) using known parameters.
- Train a neural network to predict z.
- Minimize a parameter loss, for instance

$$\mathcal{L}_{\text{param}}(z, \hat{z}) = \|z - \hat{z}\|^2$$



Limitation: Select one target in a possibly non-unique solution set.
The loss penalizes any prediction different from the single annotated target, even when several parameter vectors can explain the same observation.

Blind Estimation of Audio Effects (BE-AFX) [Peladeau, Peeters, ICASSP 24]



- An autoencoder-inspired architecture in which the latent representation corresponds to the parameter vector of a differentiable audio effect (DDSP), used as the decoder. The model is trained by minimizing the audio reconstruction loss $\mathcal{L}(y, \hat{y})$.
- Various analysis networks can be used: MEE (Music Effects Encoder), TFE (Time-Frequency Encoder), etc.
- No ground-truth effect parameters are required during training.

Outline

From deterministic estimation to Bayesian inference

Inverse problem

Given the observed processed signal y (and, when available, the input signal x), estimate the audio effect parameters z:

$$(x, y) \longmapsto z \qquad \text{or} \qquad y \longmapsto z.$$

Deterministic approach

$$\hat{z} = f_\theta(y)$$

Predict a single parameter vector.

Bayesian approach

$$p(z \mid x, y)$$

Estimate the posterior distribution over all plausible parameter vectors.

Question: Should we estimate one solution or all plausible solutions?

Proposed approach

Replace deterministic parameter estimation by posterior distribution estimation.

Deterministic estimation

$$\hat{z} = f_\theta(y)$$

One estimated parameter vector.

Bayesian estimation

$$p_\theta(z \mid y)$$

A posterior distribution over all plausible parameter vectors.

Once the posterior is estimated, we can

- sample multiple valid parameter vectors,
- compute a Maximum A Posteriori (MAP) estimate,
- quantify uncertainty and multimodality.

z becomes a random variable modeling the distribution of the parameters of the differentiable audio processor.

Bayesian perspective

Unknown processor parameters

Treat z as a random variable conditioned on the observation:

$$p(z \mid x, y).$$

The posterior should assign high density to parameters that reconstruct the observed sound:

$$\hat{y} = f_{\mathrm{DDSP}}(x, z) \approx y.$$

- narrow posterior: parameters are identifiable;
- wide posterior: many parameters are acceptable;
- structured posterior: parameters compensate each other.

Bayesian formulation

Boltzmann-Gibbs distribution model

We define the posterior from an audio reconstruction loss function ℓ:

$$p(z \mid x, y) = a \cdot e^{-\ell(f_{\mathrm{DDSP}}(x,z),y)},$$

where a is a real-valued normalization constant defined as: $a^{-1} = \int_{\mathcal{Z}} e^{-\ell(f_{\mathrm{DDSP}}(x,z),y)} dz$
We have:

$$\begin{aligned} \ln\left(p(z|x,y)\right) &= \ln(a) - \ell\left(f_{\mathrm{DDSP}}(x,z),y\right), && (1) \\ &= \ln(a) - \ell(\hat{y},y). && (2) \end{aligned}$$

Low reconstruction error $\Rightarrow$ high posterior probability.

Learning an approximate posterior

The true posterior $p(z \mid x, y)$ is generally intractable.

Learned posterior

We introduce a parametric distribution:

$$p_\theta(z \mid y)$$

which should approximate the true posterior:

$$p_\theta(z \mid y) \approx p(z \mid x, y).$$

Optimization criterion

We minimize the Kullback–Leibler divergence:

$$D_{\mathrm{KL}}\left(p_\theta(z \mid y) \parallel p(z \mid x, y)\right).$$

The model learns a posterior distribution, not a single estimate.

From KL divergence to the training loss

Starting from:

$$D_{\mathrm{KL}}\left(p_\theta(z \mid y) \parallel p(z \mid x, y)\right) = \mathbb{E}_{z\sim p_\theta(z|y)}\left[\ln\left(p_\theta(z \mid y)\right) - \ln\left(p(z \mid x, y)\right)\right].$$

Using the Boltzmann definition:

$$\ln p(z \mid x, y) = -\ell(f_{\mathrm{DDSP}}(x, z), y) + \ln(a)$$

Therefore:

$$\begin{aligned} D_{\mathrm{KL}}\left(p_\theta(z \mid y) \parallel p(z \mid x, y)\right) &= -\mathbb{H}_{p_\theta}[z \mid y] + \mathbb{E}_{z\sim p_\theta(z|y)}\left[\ell\left(f_{\mathrm{DDSP}}(x, z), y\right) - \ln(a)\right] && (3)\\ &= -\mathbb{H}_{p_\theta}[z \mid y] + \mathbb{E}_{z\sim p_\theta(z|y)}\left[\ell\left(\hat{y}, y\right)\right] + \mathit{const} && (4)\end{aligned}$$

Resulting training loss function

$$\mathcal{L}(x, y, \theta) = -\mathbb{H}[p_\theta(z \mid y)] + \mathbb{E}_{z\sim p_\theta(z|y)}\left[\ell(\hat{y}, y)\right].$$

with $\mathbb{H}[p_\theta(z \mid y)]$ the conditional differential entropy of z following the parametric distribution p_θ. The constant a can be ignored as it is independent from θ.

Entropy-regularized training objective

The training objective balances reconstruction accuracy and posterior diversity.

Entropy-regularized objective

$$\mathcal{L}(x, y, \theta) = -\beta\, \mathbb{H}_{p_\theta}[z \mid y] + \mathbb{E}_{z \sim p_\theta(z|y)}\left[\ell(\hat{y}, y)\right], \quad \forall \beta \in \mathbb{R}^+$$

Inspired by the trade-off mechanisms used in β-VAE objectives [Higgins et al, 2017], the coefficient β controls the contribution of the entropy term.

Reconstruction term

$$\mathbb{E}[\ell(\hat{y}, y)]$$

encourages every sampled parameter vector to reconstruct the observed sound.

Entropy term

$$-\mathbb{H}_{p_\theta}[z \mid y]$$

encourages a broader posterior and captures multiple plausible parameter vectors.

Increasing β favors diversity, while decreasing β favors reconstruction accuracy.

Modeling the posterior distribution $p_\theta(z \mid y)$

Base Gaussian distribution

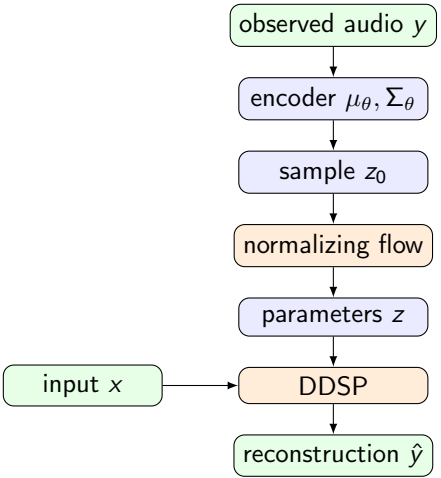
The encoder predicts the parameters of a Gaussian:

$$z_0 \sim \mathcal{N}(\mu_\theta(y), \Sigma_\theta(y)).$$

Normalizing Flow

A normalizing flow transforms the simple Gaussian sample into a more expressive parameter distribution:

$$z = NF_\theta(z_0).$$

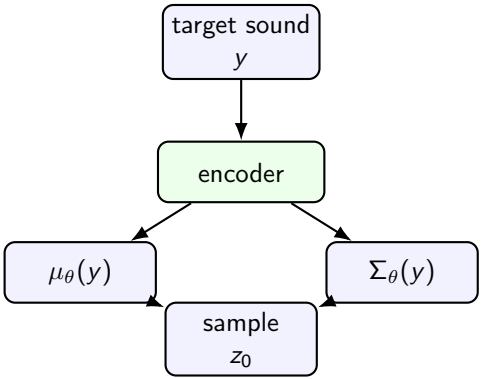


VAE-style encoder: parameterized base distribution

The encoder maps the observed sound to a Gaussian base distribution:

$$z_0 \sim \mathcal{N}\big(\mu_\theta(y), \Sigma_\theta(y)\big).$$

- Similar to a VAE encoder.
- The latent space is constrained to represent processor parameters.
- The decoder role is played by a differentiable DSP model.



Conditional normalizing flows: expressive posteriors

A Gaussian posterior cannot capture multimodal or highly nonlinear parameter distributions.

Normalizing flow

$$z = z_T = f_{\mathrm{NF}}(z_0) = f_T \circ f_{T-1} \circ \cdots \circ f_1(z_0),$$

with invertible transformations f_t, $\forall t \in 1, 2, ..., T$.

$$\ln p_\theta(z_T \mid y) = \ln p_\theta(z_0 \mid y) - \ln |\det J_{f_{\mathrm{NF}}}(z_0)| \tag{5}$$

where $J_{f_{\mathrm{NF}}}(z_0)$ is the Jacobian of the transformation. The second term follows directly from the change-of-variable theorem.

Result

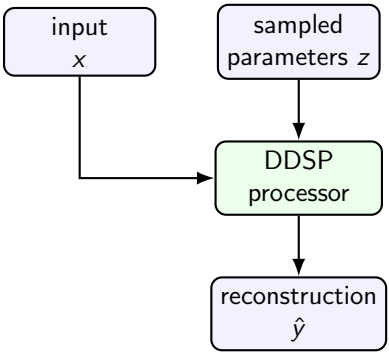
The normalizing flow transforms a simple Gaussian into a flexible posterior distribution while preserving the exact density evaluation.

DDSP as decoder / differentiable forward model

The sampled latent variable is not an abstract code: it is the parameter vector of a processor.

$$z \longrightarrow f_{\mathrm{DDSP}}(x, z) = \hat{y}.$$

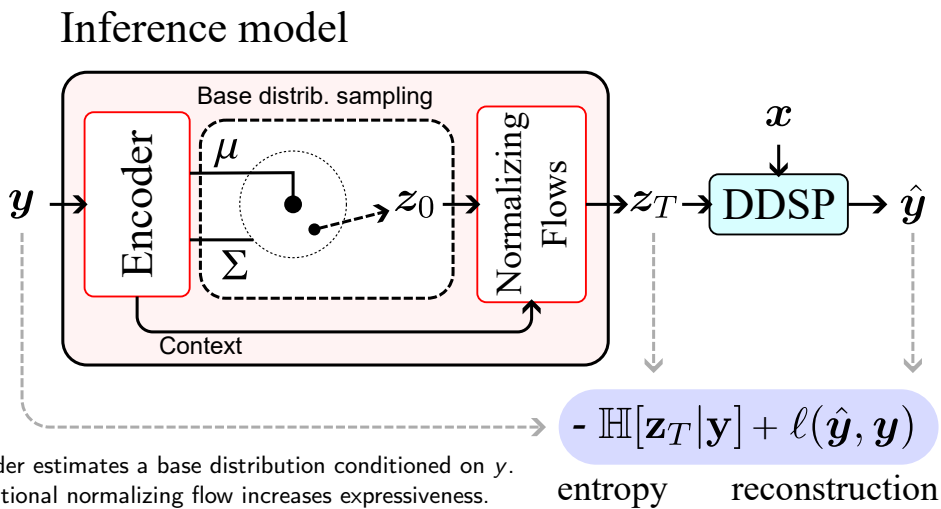
- physically interpretable latent space;
- gradients propagate through the processor;
- training is driven by audio similarity.



Important distinction

The decoder is not a black-box neural network: it is a differentiable signal-processing model.

Complete inference model



- Encoder estimates a base distribution conditioned on y.
- Conditional normalizing flow increases expressiveness.
- DDSP maps sampled parameters to audio.
- Loss balances reconstruction and entropy.

Outline

Experimental setups

Experiment 1: Equalizer parameter estimation

- Input: log-Mel spectrogram
- Encoder: ConvNeXt (1-D)
- DDSP: two high-shelf filters
- Parameters: $z \in [0, 1]^6$
- Audio loss: log-Mel distance

Experiment 2: FM synthesizer inversion

- Input: DFT magnitude spectrum
- Encoder: MLP
- DDSP: FM synthesizer
- Parameters: $z \in [0, 1]^2$
- Audio loss: SOT + MR-STFT

Common inference model

Both experiments use the same Bayesian inference framework:
encoder $\rightarrow$ Gaussian posterior $\rightarrow$ normalizing flow $\rightarrow$ differentiable processor $\rightarrow$ audio loss $-\beta$ entropy.

Experiment 1: estimating equalizer parameters

Controlled ambiguity

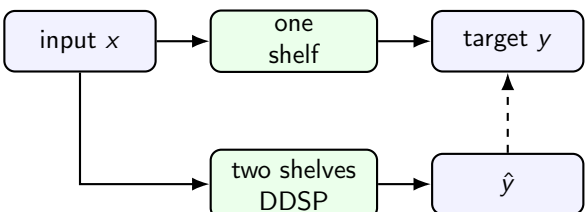
Data are generated with one high-shelf filter, but the model reconstructs it with two differentiable high-shelf filters in cascade.

Each shelf has 3 parameters:

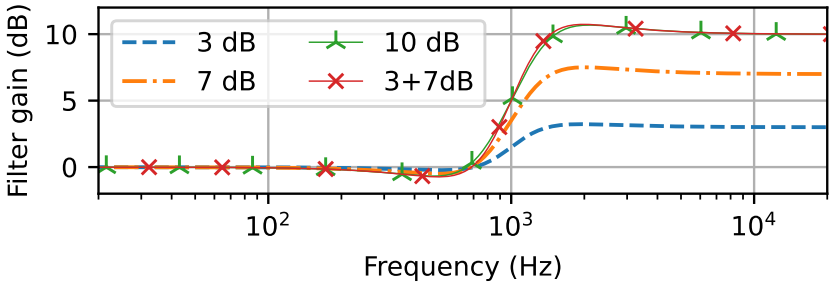
- center frequency f
- gain g
- quality factor Q

Expected ambiguity:

$$g_1 + g_2 \approx g.$$



Why the EQ setup is ambiguous



- The cascade of two shelves can reproduce the response of a single shelf.
- Several combinations of gains, cutoff frequencies and slopes produce nearly identical frequency responses.
- Consequently, different parameter vectors generate almost indistinguishable output signals.
- This controlled ambiguity makes the equalizer an ideal benchmark for posterior estimation.

Experiment 1: Experimental setup

Training data

- 50,000 audio excerpts from the Million Song Dataset[a]
- 40k / 5k / 5k, train/validation/test split
- 5-second excerpts
- Randomly generated EQ parameters

[a] Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., & Lamere, P. *The Million Song Dataset*. ISMIR, 2011.

Compared methods

- Deterministic regression baseline (BE-AFX)[a]
- Bayesian model with one flow layer ($T = 1$);
- Bayesian model with two flow layers ($T = 2$).

Training

- Entropy weight: $\beta = 0.1$;
- Same encoder and DDSP model for all methods.

[a] C. Peladeau and G. Peeters. *Blind estimation of audio effects using an auto-encoder approach and differentiable digital signal processing*, Proc. IEEE ICASSP 2024

Experiment 1: Quantitative results

Overall performance

Model	Mel ↓	MR-STFT ↓	SI-SDR ↑	Entropy bits/dim ↑
Deter.	0.235	0.412	14.8	–
Infer. ($T = 1$)	0.286	0.477	13.5	−1.51
Infer. ($T = 2$)	0.289	0.481	13.6	−1.41
Input	2.154	4.302	3.0	–

- Deterministic regression gives the best single prediction.
- One flow layer is sufficient; adding a second layer brings almost no improvement.

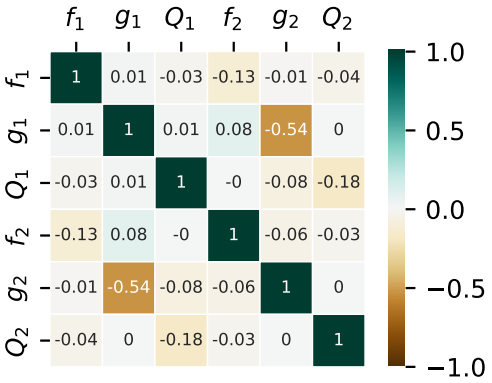
Best-of-N sampling ($T = 1$)

Model	N	Mel ↓	MR-STFT ↓	SI-SDR ↑
Deter.	–	0.235	0.412	14.8
Infer.	1	0.286	0.477	13.5
	2	0.241	0.401	15.3
	3	0.226	0.373	16.2
	5	0.206	0.340	17.3
	10	0.186	0.307	18.7
Uniform	10	0.667	1.141	11.8

Main result

Although a single probabilistic sample is less accurate, sampling only a few candidates already outperforms the deterministic estimator.

Experiment 1: Learning parameter dependencies



Average posterior correlation matrix computed over the test set.

Method

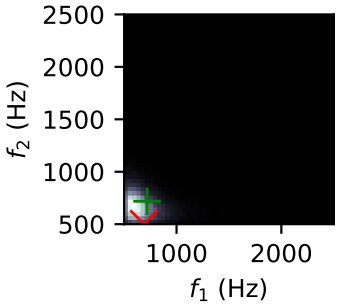
- Sample multiple parameter vectors from the learned posterior.
- Compute the parameter correlation matrix for each test example.
- Average the correlation matrices over the test set.

Key observation

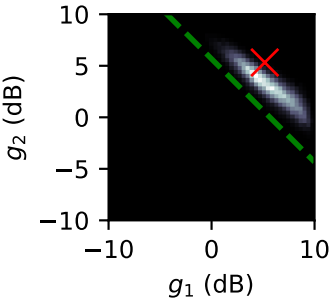
A strong negative correlation is learned between the two shelf gains (g_1 and g_2).
This reflects the compensation mechanism:

$$g_1 + g_2 \approx \text{constant},$$

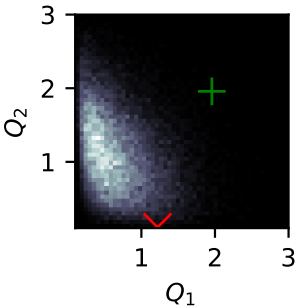
Experiment 1: Posterior visualization



(1) Frequency parameters



(2) Gain parameters



(3) Quality-factor parameters

Each plot shows the posterior distribution of one parameter pair for a single observation.

- Background: posterior density estimated by the inference model.
- Green marker: ground-truth parameter vector.
- Red $\times$: deterministic estimate.

Experiment 2: FM synthesizer sound matching

Inverse problem

Given a synthesized audio signal y, estimate the parameters of a differentiable frequency-modulation (FM) synthesizer to reproduce its timbre.

Differentiable forward model

$$y_c(t) = I_c \sin(2\pi f_c t + I_m \sin(2\pi f_m t)) .$$

- f_c: carrier frequency;
- f_m: modulation frequency;
- I_m: modulation index;
- I_c: output amplitude.

Estimated parameters

$$\boldsymbol{z} = \left(I_m, \frac{f_m}{f_c} \right) \in [0, 10]^2.$$

- Fixed: carrier frequency f_c and output amplitude I_c.

Training objective

$$\mathcal{L}_{\text{audio}} = \mathcal{L}_{\text{SOT}} + \mathcal{L}_{\text{MR-STFT}}.$$

Why is the inverse problem ambiguous?

For small modulation indices ($I_m \approx 0$), the modulation has little effect on the generated spectrum. As a result, many values of the modulation ratio f_m/f_c produce nearly indistinguishable sounds.

Experiment 2: Experimental protocol

Dataset

- Synthetic dataset generated with our differentiable FM synthesizer.
- 125 ms stationary sounds.
- Parameters sampled uniformly:

$$\mathrm{v} \in [0,1]^2.$$

- Physical parameters obtained through the mappings

$$I_m = 4\pi\, 10^{\frac{3}{4}\, 99(v_1 - 1)/20},$$

$$\frac{f_m}{f_c} = 9.5\, v_2 + 0.5.$$

Model architecture

- Input: magnitude spectrum (DFT).
- MLP encoder with Swish activations.
- $\approx$ 170k trainable parameters.
- Same deterministic and Bayesian architectures as in Experiment 1.

Training

- Audio loss:

$$\mathcal{L}_{\mathrm{SOT}} + \mathcal{L}_{\mathrm{MR\text{–}STFT}}.$$

- AdamW optimizer.

Why Spectral Optimal Transport (SOT)?

Unlike standard spectral losses, SOT is approximately convex with respect to frequency errors, making it suitable for estimating the modulation ratio.

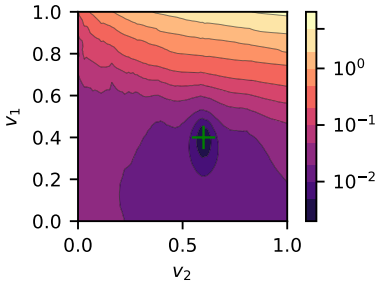
Experiment 2: Quantitative results

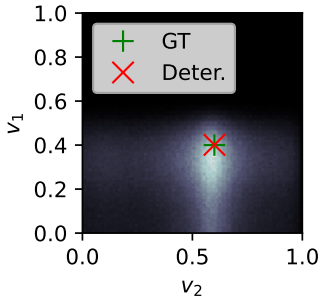
Model	T	Entropy ↑	SOT ↓	Mel ↓	SI-SDR ↑
Deterministic	–	–	**0.006**	**0.122**	**26.8**
Inference	1	-2.40	0.022	0.513	15.0
Inference	2	-2.43	0.022	0.502	14.6
Uniform (random)	–	–	0.514	3.308	2.08

Observations

- Deterministic regression achieves the lowest reconstruction error.
- Bayesian inference gives similar performance for $T = 1$ and $T = 2$.
- The learned posterior nevertheless provides uncertainty estimates unavailable to deterministic models.

Experiment 2: Posterior at low modulation index





Low modulation index ($I_m \ll 1$)

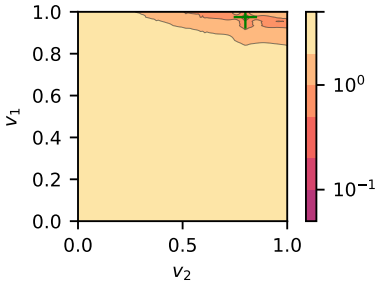
- Only weak harmonics are generated.
- The modulation ratio has little influence on the spectrum.
- Many parameter configurations produce similar sounds.

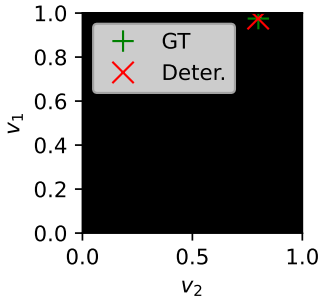
Posterior behaviour

The inferred posterior spreads over the whole low-loss region, reflecting the ambiguity of the inverse problem.

$$\mathbb{H}(z|y) \approx -0.40 \text{ bits/dim.}$$

Experiment 2: Posterior at high modulation index





High modulation index

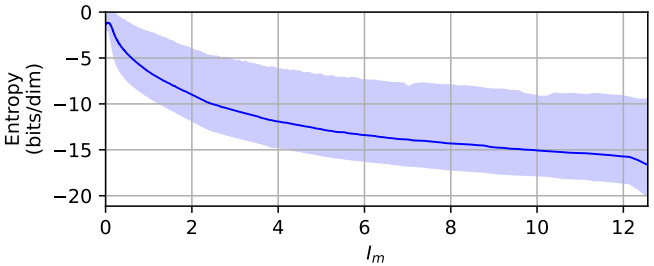
- Strong harmonics are generated.
- The modulation ratio strongly affects the spectrum.
- Only a small region of the parameter space matches the sound.

Posterior behaviour

The inferred posterior concentrates around a single solution, indicating a much less ambiguous inverse problem.

$$\mathbb{H}(z|y) \approx -5.64 \text{ bits/dim.}$$

Experiment 2: Posterior adapts to ambiguity



Conditional entropy of the estimated posterior as a function of the modulation index I_m. The solid line shows the mean and the shaded area the 10–90% percentile range.

Observation

- Low I_m:
  - weak harmonics;
  - many plausible parameter vectors.
- High I_m:
  - strong harmonics;
  - parameters become identifiable.

Take-home message

As the modulation index increases, the inverse problem becomes better posed.
The learned posterior automatically reflects this by reducing its conditional entropy.

Outline

Conclusions, limitations and future work

Take-home messages

- Audio processor inversion is often an ill-posed inverse problem with multiple valid solutions.
- Estimating a posterior distribution provides uncertainty information and captures parameter ambiguity.
- Conditional normalizing flows enable expressive posterior distributions while remaining compatible with differentiable audio models.

Current limitations

- Experiments are limited to relatively low-dimensional parameter spaces (EQ and FM synthesis).
- Performance relies on the accuracy of the differentiable forward model (DDSP).
- Only audio reconstruction is considered; perceptual criteria could better reflect human judgments.

Future directions

- Scale to more complex audio processors and larger parameter spaces.
- Integrate perceptual or task-specific audio losses.
- Explore more specific Bayesian inference methods and posterior representations.

Thank you for your attention!

Questions?

Reference

Peladeau, C., Fourer, D., Peeters, G.
Audio Processor Parameters: Estimating Distributions Instead of Deterministic Values.
Proceedings of the 28th International Conference on Digital Audio Effects (DAFx-25), Ancona, Italy, 2025, pp. 275–282.

Code
`https://github.com/peladeaucome/DAFx_Params_Distrib`

GitHub repository

Slides available upon request.

Appendix 1: High-shelf filters and parameter ambiguity

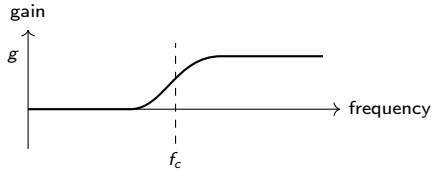
What is a high-shelf filter?
A high-shelf filter amplifies or attenuates frequencies above a transition frequency f_c.

$$H(\mathrm{z}) = \frac{b_0 + b_1\mathrm{z}^{-1} + b_2\mathrm{z}^{-2}}{1 + a_1\mathrm{z}^{-1} + a_2\mathrm{z}^{-2}}, \quad \forall \mathrm{z} \in \mathbb{C}$$

It is typically controlled by three parameters:

$$z_i = (f_{c,i},\, g_i,\, Q_i),$$

where g_i is the high-frequency gain and Q_i controls the transition shape.



Why can two filters resemble one?
For two filters in cascade,

$$H_{\mathrm{tot}}(\mathrm{z}) = H_1(\mathrm{z})H_2(\mathrm{z}).$$

Therefore, their gains add in decibels:

$$20\log_{10}|H_{\mathrm{tot}}| = 20\log_{10}|H_1| + 20\log_{10}|H_2|.$$

If the two transition frequencies are close, their combined response can approximate one broader high-shelf response:

$$g_{\mathrm{equiv}} \approx g_1 + g_2.$$

For example,

$$g_1 = 2\ \mathrm{dB}, \qquad g_2 = 4\ \mathrm{dB}$$

may produce a response similar to a single shelf with approximately

Appendix 2: FM synthesis and Bessel functions

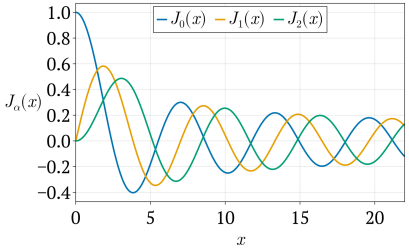
FM synthesis model

$$y_c(t) = I_c \sin(2\pi f_c t + I_m \sin(2\pi f_m t)) .$$

Possible extension using the Jacobi–Anger expansion:

$$y_c(t) = I_c \sum_{n=-\infty}^{+\infty} J_n(I_m) \sin(2\pi(f_c + n f_m)t) ,$$

where J_n is the Bessel function of the first kind of order n.



Bessel coefficients $J_n(I_m)$ as functions of the modulation index.

Spectral interpretation

The FM spectrum contains components at

$$f_c + n f_m, \qquad n \in \mathbb{Z},$$

whose amplitudes are controlled by

$$J_n(I_m).$$

Connection with the inverse problem

For $I_m \ll 1$,

$$J_0(I_m) \approx 1, \qquad J_n(I_m) \approx 0 \quad (|n| \geq 1).$$

The signal is therefore dominated by the carrier f_c, and the modulation frequency f_m is difficult to identify.