



Improving Inference-Time Optimisation for Vocal Effects Style Transfer with a Gaussian Prior

Chin-Yun Yu¹ Marco A. Martínez-Ramírez² Junghyun Koo² Wei-Hsiang Liao²Yuki Mitsufuji^{2,3} György Fazekas¹¹Centre for Digital Music, Queen Mary University of London, London, UK²Sony AI, Tokyo, Japan³Sony Group Corporation, Tokyo, Japan

Problem Statement: Style Transfer for Vocal Effects

The goal of this work is to transfer the *effect style* from a reference vocal track to another raw vocal track using inference time optimisation (ITO). We formulate the problem as follows:

Definition (The Problem). Given a reference track $\bar{y} \in \mathbb{R}^N$, C raw tracks $\tilde{x} \in \mathbb{R}^{N \times C}$, and a content-invariant style encoder $g : \mathbb{R}^N \rightarrow \mathbb{S}^{D-1}$, process \tilde{x} so the resulting \tilde{y} has the same *effect style* as \bar{y} .

Assumptions:

- An effects model $f : \mathbb{R}^{N \times C} \times \mathbb{R}^M \rightarrow \mathbb{R}^N$ is known.
- $\bar{y} = f(\bar{x}, \theta)$ and $\tilde{y} = f(\tilde{x}, \theta)$ for some unknown \bar{x} and θ .
- θ contains the information of *effect style*.

Theorem (Style Transfer = Parameter Estimation). Transferring the reference style is equivalent to modelling the posterior distribution

$$p(\theta | \bar{y}, \tilde{x}) = p(\theta | \bar{z} = g(\bar{y}), \tilde{x}).$$

Existing Approaches: The Maximum Likelihood Estimation

Prior works [1, 2] achieve ITO by minimising the distance \mathcal{D} between the style embeddings $\tilde{z} = g(\tilde{y})$ and $\bar{z} = g(\bar{y})$. It is equivalent to the maximum likelihood estimation (MLE) in Eq. (1) with $\alpha = 0$. Since the prior is ignored, the estimated parameters θ^* are not bounded and may be unreasonable.

Key Insight. The prior matters because a perfect style encoder g is unattainable.

Proposed Method: The Maximum-A-Posteriori Estimation

We propose to include a prior term to regularise the ITO, which equals the maximum-a-posteriori (MAP) estimation of parameters θ :

$$\theta^* = \arg \max_{\theta} \log p(\bar{z} | \theta, \tilde{x}) + \alpha \log p(\theta | \tilde{x}) \quad (1)$$

previous approaches

$$p(\bar{z} | \theta, \tilde{x}) = p(\bar{z} | \tilde{z} = (g \circ f)(\tilde{x}, \theta)) \quad (2)$$

$$p(\bar{z} | \tilde{z}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\arccos(1 - \mathcal{D}(\tilde{z}, \bar{z}))^2}{2\sigma^2}\right) \quad (3)$$

$$p(\theta | \tilde{x}) \approx p(\theta) = \mathcal{N}(\bar{\theta}, \Sigma_{\theta}) \quad (4)$$

- f : differentiable effects from [3] consist of parametric equalisers, a compressor, a ping-pong delay, and a reverb.
- $\bar{\theta}, \Sigma_{\theta}$: estimated from 365 vocal presets in [3].
- \mathcal{D} : cosine distance [2].

Experimental Setup

- **Evaluation set:** 65 vocal tracks from MedleyDB [4, 5].
- **Baselines:**
 - Oracle: parameters θ derived from paired data (\tilde{x}, \bar{y}) by [3].
 - Mean: always predict $\bar{\theta}$.
 - Regression: a five-layer convolutional neural network that predicts θ from \bar{y} directly. It was trained on the same proprietary dataset as [3].
 - Nearest Neighbour (NN): find the closest \bar{z} in the vocal presets and use its corresponding θ .
- **Encoders g :** AFx-Rep from [2], 25 Mel-frequency cepstral coefficients (MFCC), and MIR features including RMS, crest factor, dynamic spread, etc. (MIR).
- **Evaluation metrics:**
 - MSS: Multi-scale spectral loss [3].
 - MLDR: Multi-scale loudness dynamic range loss [3].
 - PMSE: Parameter mean squared error.
- **Optimisation:** Adam with learning rate 0.01 for 1000 iterations.

Overview and Contributions

- Converting the effects style transfer task into a MAP problem and outperforming baselines using a Gaussian prior.
- Exploration of multiple encoders g for style representation.
- Scalability to limited paired data regimes (< 400 tracks).

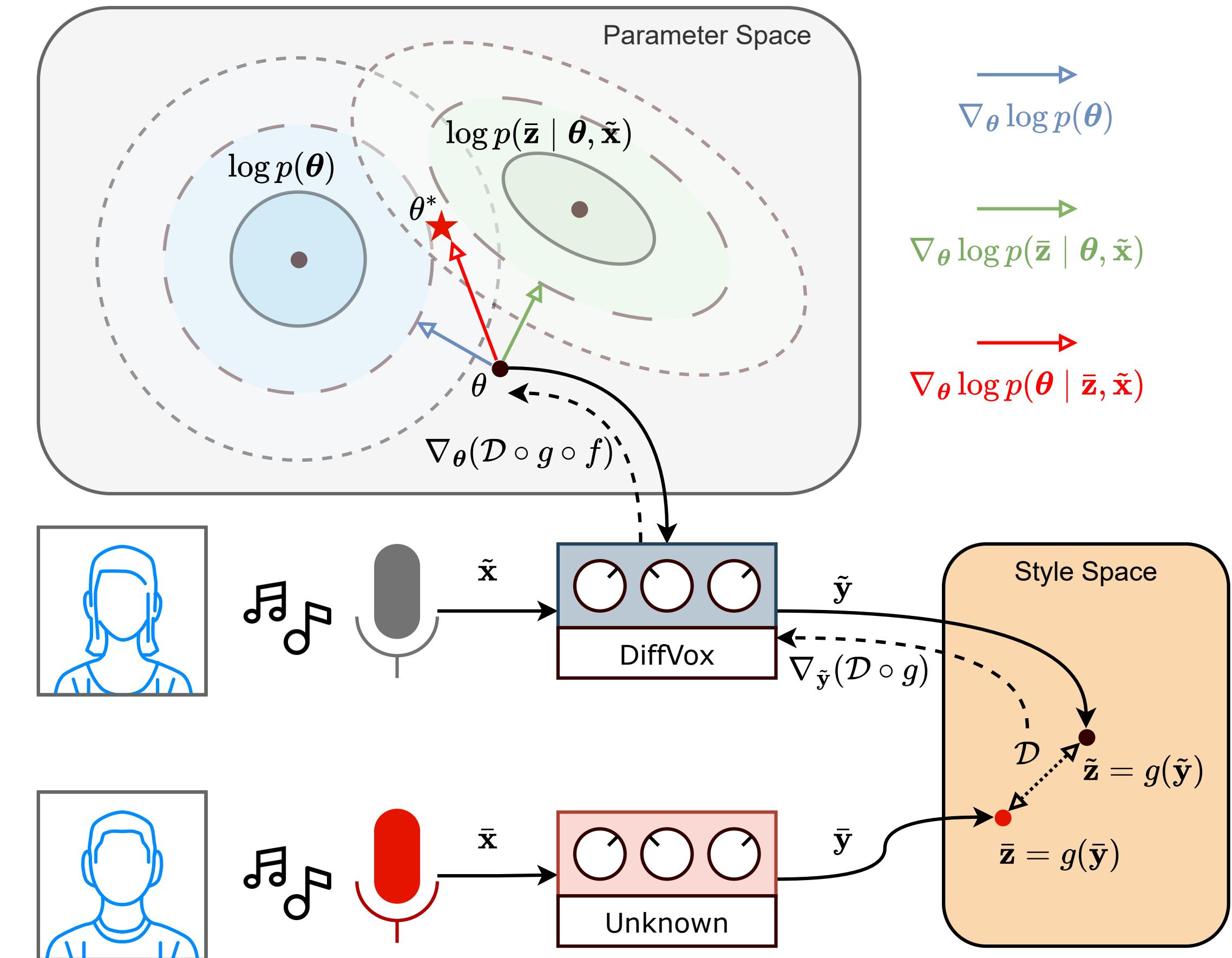


Figure 1. Overview of the proposed calibration method. The red star is the optimal parameters θ^* for the vocal effects style transfer.

Interesting Fact. The optimisation process is similar to non-stochastic diffusion sampling with classifier guidance [6].

Experimental Results

Table 1. Median scores of the proposed methods and baselines.

Method	MSS ↓		MLDR ↓		PMSE ↓
	I/r	m/s	I/r	m/s	
Oracle	0.775	1.012	0.313	0.383	0.0
Mean	+0.354	+0.836	+0.503	+0.692	+5.310
Regression	+0.281	+0.574	+0.480	+0.695	+5.002
NN-θ	+0.381	+0.675	+0.518	+0.629	+4.145
NN-AFx-Rep	+0.321	+0.672	+0.320	+0.504	+9.463
NN-MFCC	+0.274	+0.464	+0.424	+0.559	+8.374
NN-MIR	+0.424	+0.803	+0.561	+0.706	+10.019
Encoder α					
0.0	+0.435	+0.570	+0.343	+0.424	+7.756
0.01	+0.221	+0.606	+0.249	+0.402	+5.924
0.1	+0.211	+0.513	+0.321	+0.445	+5.168
1.0	+0.318	+0.795	+0.427	+0.629	+5.339
0.0	+0.761	+0.897	+1.047	+0.977	+9.255
0.01	+0.507	+0.531	+0.720	+0.765	+6.706
0.1	+0.333	+0.469	+0.514	+0.621	+5.661
1.0	+0.312	+0.563	+0.459	+0.547	+5.250
0.0	+0.782	+1.105	+0.873	+0.797	+7.103
0.01	+0.598	+1.505	+0.856	+0.854	+5.622
0.1	+0.490	+0.807	+0.778	+0.778	+5.359
1.0	+0.363	+0.714	+0.508	+0.695	+5.319

Subjective Evaluation: The regression model is rated the lowest, and AFx-Rep with ITO is rated the highest (not statistically significantly), averaged over 16 participants.

Conclusions and Next Steps

The results highlight the importance of incorporating prior knowledge into ITO for better performance. Future work includes extending it to more complex or non-differentiable effect chains and stronger priors.

- [1] Chu et al. Text2FX: Harnessing CLAP Embeddings for Text-Guided Audio Effects. IEEE. 2025.
- [2] Steinmetz et al. ST-ITO: Controlling Audio Effects for Style Transfer With Inference-Time Optimization. Nov. 2024. doi: 10.5281/zenodo.14877423.
- [3] Yu et al. DiffVox: A Differentiable Model for Capturing and Analysing Vocal Effects Distributions. 2025.
- [4] Bittner et al. MedleyDB: A multitrack dataset for annotation-intensive mir research. 2014.
- [5] Bittner et al. MedleyDB 2.0: New data and a system for sustainable data collection. ISMIR Late Breaking and Demo Papers (2016).
- [6] Song et al. Score-Based Generative Modeling through Stochastic Differential Equations. 2021.