VOCAL TIMBRE EFFECTS WITH DIFFERENTIABLE DIGITAL SIGNAL PROCESSING

David Südholt *

Centre for Digital Music
Queen Mary University of London
London, UK
d.sudholt@qmul.ac.uk

Cumhur Erkut

Multisensory Experience Lab
Aalborg University Copenhagen
Copenhagen, Denmark
cer@create.aau.dk

ABSTRACT

We explore two approaches to creatively altering vocal timbre using Differentiable Digital Signal Processing (DDSP). The first approach is inspired by classic cross-synthesis techniques. A pretrained DDSP decoder predicts a filter for a noise source and a harmonic distribution, based on pitch and loudness information extracted from the vocal input. Before synthesis, the harmonic distribution is modified by interpolating between the predicted distribution and the harmonics of the input. We provide a real-time implementation of this approach in the form of a Neutone model.

In the second approach, autoencoder models are trained on datasets consisting of both vocal and instrument training data. To apply the effect, the trained autoencoder attempts to reconstruct the vocal input. We find that there is a desirable "sweet spot" during training, where the model has learned to reconstruct the phonetic content of the input vocals, but is still affected by the timbre of the instrument mixed into the training data. After further training, that effect disappears.

A perceptual evaluation compares the two approaches. We find that the autoencoder in the second approach is able to reconstruct intelligible lyrical content without any explicit phonetic information provided during training.

1. INTRODUCTION

Neural singing voice synthesis has made great progress over recent years. Many efforts are focused on generating natural-sounding voices. The fame of the classic "vocoder" sound however, popularized by artists like Daft Punk or Kraftwerk shows the desire for creative timbre manipulation of vocals, where naturalness is not a desired characteristic.

Differentiable Digital Signal Processing (DDSP) [1] proposes an end-to-end learning approach for neural audio synthesis. Instead of generating signals sample-by-sample in the time domain, or time-varying spectra in the frequency domain, DDSP offers a library of synthesizer components implemented entirely within a framework supporting auto-differentiation. In the case of timbre transfer, an autoencoder model is trained to reconstruct a monophonic sound source based on into pitch and loudness information by generating time-varying control parameters for the synthesizers. The loss is calculated by comparing the spectrogram of the generated audio from the synthesizers to that of the original audio on multiple timescales. The auto-differentiable implementation allows the gradient of the loss to backpropagate through the synthesizers to update the model weights of the autoencoder.

The synthesizers are based on the spectral modeling synthesis (SMS) [2] framework. The harmonic components of the sound are generated by a sum of K sinusoids. The decoder predicts K time-varying amplitudes $A_k(n)$, referred to as the *harmonic distribution*, since the sinusoids are defined to oscillate at integer multiples of the (also time-varying) fundamental frequency $f_0(n)$ extracted by the encoder. Thus, the output of the harmonic component x_h can be formulated as

$$x_h(n) = a(n) \sum_{k=1}^{K} A_k(n) \cdot \sin(\phi_k(n)) \,, \tag{1}$$

where $a(n)$ is a global amplitude, also predicted by the decoder, and $\phi_k(n) = 2\pi \sum_{m=0}^{n} k f_0(m)$ is the instantaneous phase of the k-th harmonic.

Additionally, the decoder predicts the time-varying magnitude responses of a finite impulse response (FIR) filter. The non-harmonic components of the sounds are generated by processing white noise through these FIRs.

Recent approaches extended the DDSP components and source waveforms. Masuda [3] proposed a novel approach to synthesizer sound matching by implementing a basic subtractive synthesizer using differentiable DSP modules. Shan [4] introduced Differentiable Wavetable Synthesis (DWTS), a technique for neural audio synthesis that learns a dictionary of one-period waveforms through end-to-end training. Lee [5] formulated recursive differentiable artificial reverberation components, allowing loss gradients to be back-propagated end-to-end, and implemented these models with finite impulse response (FIR) approximations. Finally, Wu [6] proposed a new vocoder called SawSing for singing voice, which synthesizes the harmonic part of singing voices by filtering a sawtooth source signal with a linear time-variant finite impulse response filter whose coefficients are estimated from the input mel-spectrogram by a neural network.

Despite these achievements, the use of the classical DDSP for a vocal input with intelligible lyrics has not been explored or exploited, except in [7]. In this paper, we propose two methods of adapting DDSP to create vocal effects. We provide a real-time implementation and report perceptual experiments to evaluate our approaches. The structure of this short paper follows our approaches and experiments.

* Work performed as an M.Sc. student in Sound and Music Computing at Aalborg University Copenhagen

2. VOCAL EFFECTS WITH DDSP

Our first approach focuses on altering the predicted synthesizer parameters in a vocoding-inspired manner and will be referred to as the *vocoding approach*. The other makes use of a latent encoding of timbre information and will be referred to as the *latent approach*.

2.1. Vocoding Approach

The vocoding approach uses a model trained for timbre transfer. A decoder solely conditioned on pitch and loudness features is trained on audio recordings of a specific instrument, e.g. a trumpet. After training has completed, extracting pitch and loudness from any input audio can be used to generate the same melody line in the sound of a trumpet by using the trained decoder to predict corresponding synthesizer controls.

To create the effect of a "talking trumpet" from vocal input, we extract pitch and loudness information from the input. Before the synthesis step however, the harmonic distribution A_k is replaced by an altered distribution A_k^{out} by interpolating between the predicted distribution and the harmonics A_k^{in} of the input.

To generate A_k^{out}, a user-supplied interpolation factor $p \in [0, 1]$ is introduced. The modified distribution can then be calculated as

$$A_k^{\text{out}} = \begin{cases} (1-p)A_k + A_k^{\text{in}} & kf_0 < \frac{f_s}{2} \\ 0 & \text{else} \end{cases}, \tag{2}$$

taking care not to include oscillators at frequencies exceeding the Nyquist limit of $f_s/2$. Note that for $p = 0$, $A_k^{\text{out}} = A_k$. In this way, we can create a hybrid harmonic distribution containing both aspects of the timbre of the instrument the decoder was trained on, and the spectral contour of the phonetic content of the vocal input.

A real-time implementation of this approach is made available as a Neutone model at `https://github.com/dsuedholt/ddsp_xsynth`.

2.2. Latent Approach

In the latent approach, the encoder generates a vector z in addition to pitch and loudness information. We use an encoder provided in the DDSP library that calculates mel-frequency cepstral coefficients (MFCCs) of the input audio at every time step, and processes them through a recurrent layer before projecting them to the latent space.

In this approach, no modifications are applied to the decoder output. Instead, the effect is generated through selection of the training datasets. As explored previously [7] and confirmed through preliminary experiments, simply training a VAE on recordings of a singing voice can be sufficient to obtain a model capable of reconstructing a vocal input from a different singing voice in the style of the training data with intelligible lyrics.

The idea behind this approach is to add in other monophonic instruments, such as a trumpet or a synthesizer, to the training data, to influence the timbre of the reconstructed vocals in musically interesting ways.

During the experiments, it became clear that if the model is trained until the training loss converges, a decoder with a sufficient number of parameters learns to distinguish between vocal input and the additional instrument, and is able to reconstruct both accurately. This results in a model that is effectively just performing voice transfer.

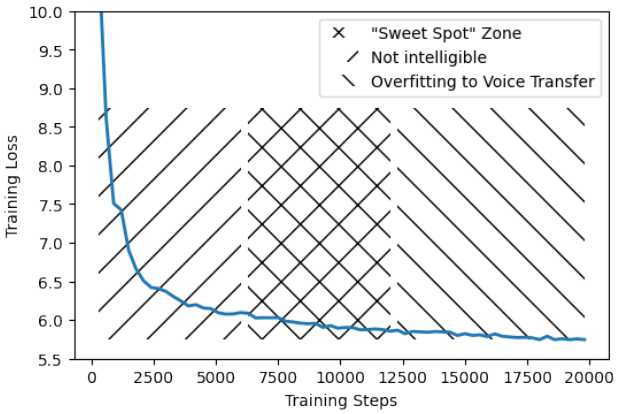

Figure 1: *The training process of a latent encoding model on a combined dataset of vocal performances and brass instruments. Early during training, it cannot reconstruct intelligible lyrics yet. Then it transitions into the "sweet spot" where lyrical content is preserved, but the timbre is affected by the additional instrument. After further training, that effect disappears, and the model performs regular voice transfer.*

However, there appears a "sweet spot" early on in training, where the model is already able to reproduce the lyrical content of the input, but has not yet learned to fully distinguish between the different input sources. At this point, the timbre of the reconstructed vocals is affected noticeably by the additional instrument. This is illustrated in Figure 1.

3. EXPERIMENTS

A dataset of vocal performances was created from the Children's Song Dataset (CSD) [8] and the MUSDB18 dataset [9]; instrument datasets were created by combining respective instrument recordings taken from the University of Rochester Multi-Modal Music Performance dataset (URMP) [10]. Additionally, a synthesizer performance was obtained by processing randomized MIDI at varying velocities and pitches through a software synthesizer.

Sound examples demonstrating the effect of the vocoding approach at various values for p, as well as the "sweet spot" effect of the latent approach, are available at `https://dsuedholt.github.io/ddsp-vocal-effects`.

We performed a perceptual evaluation to compare the two approaches. We trained and used the following four models:

VC-Synth: Timbre transfer model trained on the synth dataset, vocoding approach, $p = 0.7$

VC-Brass: Timbre transfer model trained on the brass dataset, vocoding approach, $p = 0.7$

Z-Vocals: Latent encoding voice transfer model trained on a single singer from the CSD dataset

Z-Mixed: Latent encoding model trained on a mixed dataset from the MUSDB18 medley vocals (multiple singers) and the synth dataset

Two vocal samples, one performed by a male, one by a female singer, were processed by all four models. 15 participants rated the output in a multi-stimulus test under the following three aspects:

1. Perceived audio quality
2. Intelligibility of the lyrics
3. How musically interesting the effect is

The results of the subjective evaluation are shown in Figure 2.

The clearest result can be found in the rating of the lyrical intelligibility aspect on the female input sample, where the latent encoding models clearly outperform the vocoding models. The same trend, although to a lesser degree, is shown in the evaluation of the male input sample. This seems to confirm that the MFCC + RNN encoder is already capable of reproducing intelligible lyrics without any explicit phonetic information.

None of the models are rated particularly favorably under the aspect of perceived audio quality, although the latent encoding models perform slightly better than the vocoding models. This could potentially be improved by working sampling rates greater than 16 kHz.

The highly subjective rating according to "musical interest" shows the highest variance of the ratings, although a slight trend favoring the latent encoding models seems to exist.

4. CONCLUSION

We presented two methods of creating vocal effects that show how the model training and the synthesis stage of the DDSP pipeline can be manipulated for creative effect. We demonstrated that no conditioning on explicit phonetic information is needed to preserve lyrical intelligibility while altering the timbre of the vocal input. These results pave the way towards synthetic "talking" instruments, as well as better understanding of the DDSP training mechanisms and strategies. Still, implementing a unified voice synthesis framework such as NANSY++ [11] remains a future challenge for our field in general.

5. REFERENCES

[1] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts, "DDSP: Differentiable Digital Signal Processing," in *International Conference on Learning Representations*, 2020.

[2] Xavier Serra and Julius O. Smith, "Spectral modeling synthesis. A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[3] Naotake Masuda and Daisuke Saito, "Synthesizer Sound Matching with Differentiable DSP," in *Proc. Intl. Soc. Music Information Retrieval Conf. (ISMIR)*, 2021.

[4] Siyuan Shan, Lamtharn Hantrakul, Jitong Chen, Matt Avent, and David Trevelyan, "Differentiable Wavetable Synthesis," in *Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Proc. (ICASSP)*. IEEE, may 23 2022.

[5] Sungho Lee, Hyeong-Seok Choi, and Kyogu Lee, "Differentiable Artificial Reverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2541–2556, 2022.

[6] Da-Yi Wu, Wen-Yi Hsiao, Fu-Rong Yang, Oscar Friedman, Warren Jackson, Scott Bruzenak, Yi-Wen Liu, and Yi-Hsuan Yang, "DDSP-based singing vocoders: A new subtractive-based synthesizer and a comprehensive evaluation," *arXiv*, 2022.

[7] Juan Alonso and Cumhur Erkut, "Explorations of Singing Voice Synthesis using DDSP," in *18th Sound and Music Computing Conference*, 2021, vol. 2021-June, pp. 183–190.

[8] Soonbeom Choi, "Children's Song Dataset for Singing Voice Research," in *International Society for Music Information Retrieval Conference*.

[9] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner, "The MUSDB18 corpus for music separation," Dec. 2017.

[10] Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma, "Creating a Multitrack Classical Music Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 522–535, Feb. 2019.

[11] Hyeong-Seok Choi, Jinhyeok Yang, Juheon Lee, and Hyeongju Kim, "NANSY++: Unified voice synthesis with neural analysis and synthesis," in *Intl. Conf. Learning Representations (ICLR)*, 2023, Accepted as a poster.

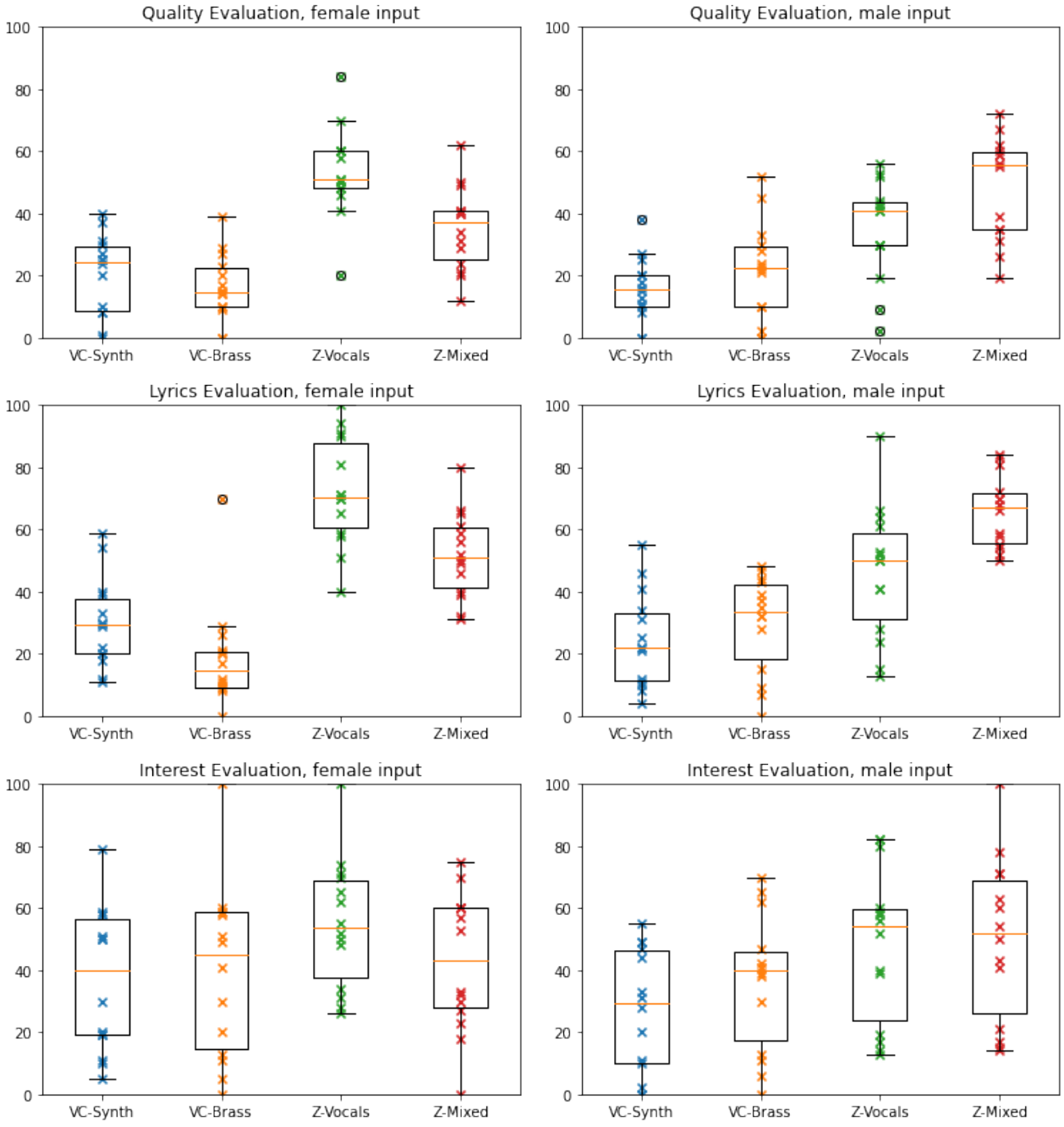

Figure 2: *Results of the perceptual evaluation. All individual ratings are displayed as a scatter plot. A box plot marks the median rating with a horizontal line. The box itself extends from the first to the third quartile of the ratings.*