EVALUATING INPUT REPRESENTATIONS FOR TIMBRE TRANSFER

Ryan Bowering
University of Rochester
rbowerin@ur.rochester.edu

ABSTRACT

Timbre transfer refers to changing the timbre of one recording to the timbre of another while preserving all other musical characteristics. Recent developments have shown the potential of diffusion models for this task, although there is no agreed upon best practice for other aspects such as data representation. This work compares models using raw waveforms to models using mel spectrograms under the same generative paradigm to determine which yields the highest performance. Measured by Fréchet Audio Distance (FAD), the performance of the mel spectrogram models are found to far surpass that of the raw waveform models.

1 Introduction

The task of timbre transfer refers to changing the timbre of a music recording to a different timbre while preserving all other musical characteristics. This task is challenging because timbre is very hard to define and model. For this work, it can generally be thought of as the perceived characteristics of a sound that are independent of pitch and dynamics. There are many possible applications for timbre transfer, such as tools for music production or faster creation of educational materials.

Recent work [5, 6, 4] has demonstrated the potential of using guided diffusion models for timbre transfer. However, there is still no consensus on best practice for other aspects of the problem, such as input representation or model architecture. For example, [5] generates mel spectrograms, [6] generates constant Q transforms (CQT), and [7] generates raw waveforms.

This work will compare generating mel spectrograms with generating raw waveforms using the same generative paradigm. Specifically, it will introduce a custom diffusion transformer [1] architecture into a flow matching generative framework [2]. The mel spectrograms are converted to waveforms using the BigVGAN neural vocoder [3]. This work does not investigate the CQT representation, as there are no high quality, public CQT vocoders available.

The models will be trained on the StarNet dataset [8], which contains recordings of instrument pairs playing the same piece (Clarinet/Strings and Vibraphone/Piano). One model will be trained for each instrument pairing, and each data representation, for a total of 8 models. Results will be evaluated by measuring the Fréchet Audio Distance (FAD) [9] between the generated samples and the targets, a measure of the distance between two distributions. The FAD will be measured separately for each pairing and representation.

2 Background

The key challenge for any generative model is learning and sampling from a distribution of observed data. One popular generative paradigm, called Continuous Normalizing Flows [10], learns an ODE mapping from a simple distribution such as $\mathcal{N}(0, \mathbf{I})$ to the data distribution as follows:

$$\frac{d}{dt}\phi_t(x) = v_t^{\theta}(\phi_t(x)) \tag{1}$$

$$\phi_0(x) = x, \tag{2}$$

where v_t^θ is a learnable time-dependent vector field parameterized by some neural network, and $\phi_t(x)$ is the trajectory at a point x. To avoid expensive simulations of the ODE, the model can be trained to regress onto a target velocity field conditional on the observed dataset, which is called flow matching [2]. Once the model is trained, new data can be generated by sampling from the initial distribution and simulating the ODE with any method such as Euler's method.

The generation process can be guided with some context signal c by amplifying the difference between an unconditional model and a conditional model as follows:

$$\tilde{v}_t^\theta(x, c) = (1 + w)v_t^\theta(x, c) - wv_t^\theta(x, \emptyset), \tag{3}$$

where w is the guidance strength, and $\emptyset$ represents no context signal. Since the vector field controls the transport direction, amplifying the conditional–unconditional difference guides samples toward regions consistent with c. This approach is called Classifier-Free Guidance [11]. It is worth noting that the same parameters can be used for both the conditional and unconditional model by randomly zeroing the context signal during training.

3 Methods

Guided flow matching is used in this work with a DiT backbone. Specifically, the model is given one of the instruments in a pair as guidance, and attempts to reconstruct the signal from the other instrument. The model is trained to match the linear interpolation field described in [2].

The key addition to the DiT architecture from [1] is an extra cross-attention module with the conditioning signal to more aggressively preserve musical context information. Additionally, due to computational constraints, the raw waveform was passed through downsample blocks that halved the length a total of 5 times, and the block output was fed through upsample blocks to restore the original waveform shape. Due to time constraints, the raw waveform models were trained for only a quarter of the epochs of the mel spectrogram models, but the loss seems to reach near convergence regardless so the comparison is still mostly valid.

4 Results

The loss throughout training is plotted in Figure 2. The raw waveform models were trained with less epochs due to computational constraints, however they still clearly perform worse than their mel counterparts.

Plotting the FAD on the holdout set throughout training reveals that the raw waveform representation is very unstable. While the models were not trained to minimize FAD, it is still expected to follow roughly the same curve as the loss. This suggests that the model may be struggling with some aspect other than audio generation when given the raw waveform, possibly a result of aggressive downsampling. Also worth noting is that the FAD drops to 0 at a single epoch for the raw waveform strings model; this is simply a placeholder value that gets used if the FAD evaluation threw an error from failing to load the VGGish model.

The final results are summarized in Table 1. While there is no standardized test set against which models are measured these results suggest that the mel spectrogram model is performing near the state-of-the-art, while the raw waveform model is lagging further behind.

Perceptually, all the raw waveform generated samples have residual noise at least equally as strong as the music. This could suggest that the hyperparameters for model inference were not ideal, perhaps needing more steps. However, if the noise is disregarded, the music quality was very good. The mel spectrogram samples might not have suffered from this issue if the vocoder was able to reduce the residual noise, but this was unexplored.

	FAD ↓			
	Strings to Clarinet	**Clarinet to Strings**	**Vibes to Piano**	**Piano to Vibes**
Mel Spectrogram	4.06	4.49	9.47	3.31
Raw Waveform	31.82	15.07	25.58	14.52

Table 1: FAD Comparison

5 Conclusion

This work compared DiT based flow matching models using both raw waveforms and mel spectrograms for the task of timbre transfer. The mel spectrogram models significantly outperformed the raw waveform models, although the latter

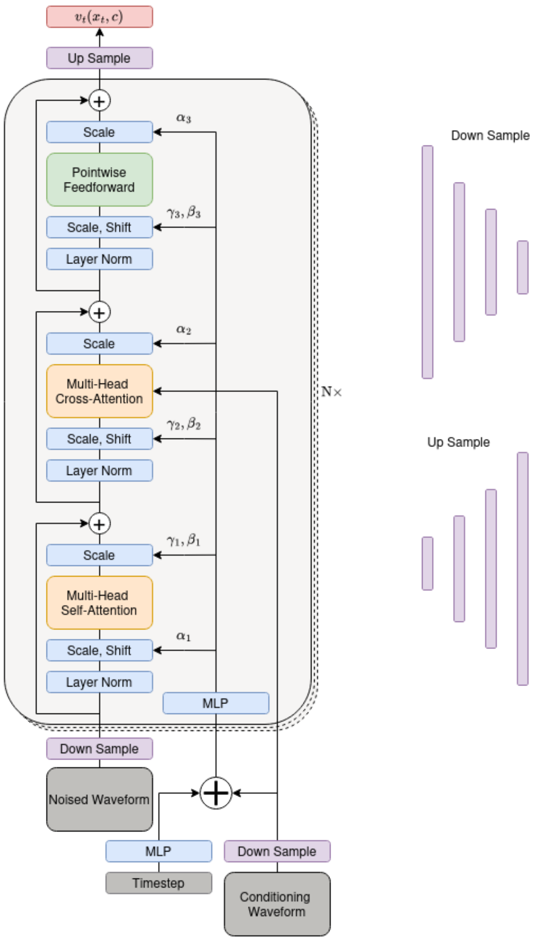

Figure 1: Model architecture for raw waveforms. The mel spectrogram model is the same but without the downsample/upsample blocks

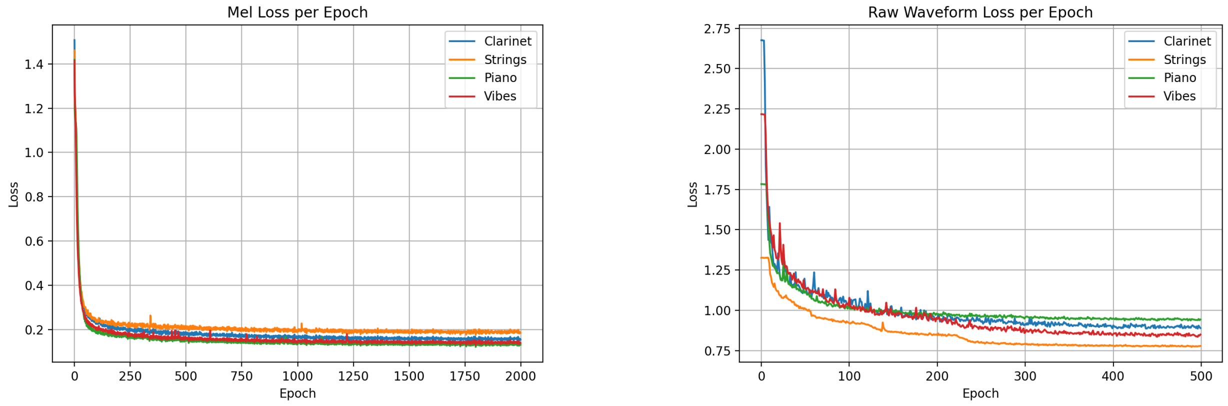

Figure 2: The losses for each pairing and representation seem to converge

does show promise. More work is needed to remove residual noise in raw waveform samples, but the baseline music quality is acceptable. Additionally, future work should compare the CQT against these models, as they were out of scope for this project.

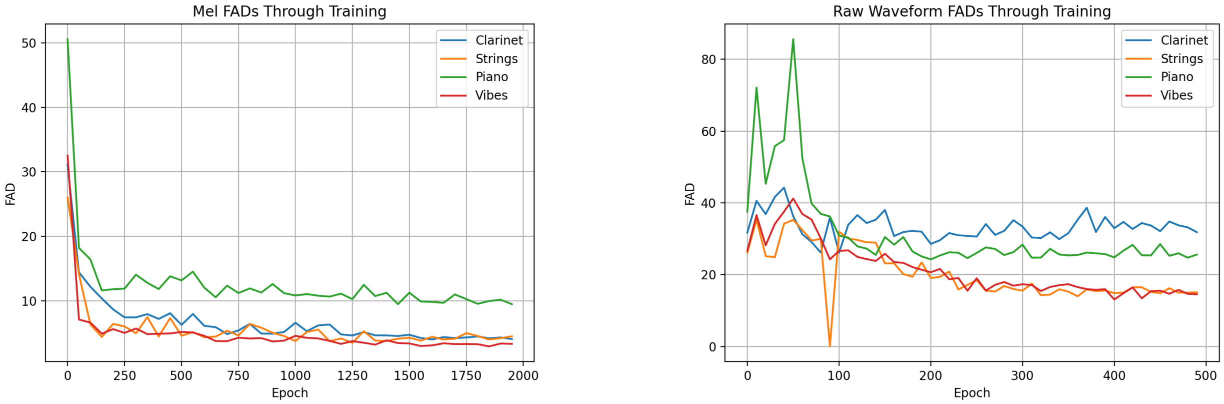

Figure 3: The FAD throughout training is far less stable than the loss

References

[1] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[2] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[3] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "Bigvgan: A universal neural vocoder with large-scale training," *arXiv preprint arXiv:2206.04658*, 2022.

[4] C.-H. Wu, P. Cheewaprakobkit, T. K. Shih, Y.-C. Lin, and B.-Z. Liu, "Automatic timbre transformation using enhanced diffusion model," *IEEE Access*, 2025.

[5] L. Comanducci, F. Antonacci, and A. Sarti, "Timbre transfer using image-to-image denoising diffusion implicit models," *arXiv preprint arXiv:2307.04586*, 2023.

[6] H. Huang, J. Man, L. Li, and R. Zeng, "Musical timbre style transfer with diffusion model," *PeerJ Computer Science*, vol. 10, p. e2194, 2024.

[7] G. FRATTICIOLI, "Mambatransfer: raw audio musical timbre transfer using selective state-space models," 2023.

[8] M. Alinoori and V. Tzerpos, "Music-star: a style translation system for audio-based re-instrumentation." in *ISMIR*, 2022, pp. 419–426.

[9] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," *arXiv preprint arXiv:1812.08466*, 2018.

[10] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," *Advances in neural information processing systems*, vol. 31, 2018.

[11] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.