Unified Timbre Transfer: A Compact Model for Real-Time Multi-Instrument Sound Morphing

Anders R. Bargum*† Naotake Masuda‡, Bogdan Teleaga‡, Andrew Fyfe‡ and Cumhur Erkut*
* Multisensory Experience Lab, Aalborg University, Copenhagen, Denmark
E-mail: {arba, cer}@create.aau.dk
† Heka VR, Copenhagen, Denmark
‡ Qosmo & Neutone AI, Tokyo, Japan
E-mail: {bogdan, masuda, andrew}@qosmo.jp

***Abstract*—Recent advances in transformer-and diffusion-based deep-generative models have significantly impacted the field of music and audio synthesis. However, controllable and real-time interactive models, such as those used for timbre transfer in music production, remain largely dominated by auto-encoders and generative adversarial networks. In pursuit of efficient and flexible timbre morphing and multi-instrument timbre transfer, we propose a simplified modeling approach, utilizing an upsampled two-dimensional timbre space in conjunction with engineered and instrument-dependent excitation signals. Different from many other works, our model enables any-to-many timbre transfer with added control over timbre, pitch, and loudness. We additionally allow for seamless interpolation between instruments, eliminating the need for separate model training. Our evaluation shows performance comparable to specialized models, making it highly relevant for the broader creative audio community.**

I. Introduction

Advancements in deep-generative modeling have significantly expanded the possibilities for manipulating audio, particularly for music generation and timbre transfer i.e the task of altering a sound's tonal quality while preserving pitch and dynamics. Early models like WaveNet [1] has enabled high-quality waveform synthesis, while methods such as NSynth [2] and RAVE [3] have applied frameworks build on variational autoencoders (VAE) and generative adversarial networks (GAN) to learn individual instrument representations. More recently, diffusion models like Stable Audio [4] have improved generation quality, whereas DDSP [5] has introduced a physics-based approach, enhancing interpretability and efficiency for monophonic timbre transfer.

Despite the progress, only DDSP and RAVE have seen practical use due to their training stability, real-time efficiency, and generalization to out-of-domain input. Yet, both face limitations: DDSP's source-filter design constrains quality, while RAVE's latent compression entangles pitch and timbre, limiting control. Moreover, both methods require individual model training for each target instrument, limiting scalability.

In this work, we propose a novel model topology for real-time, multi-instrument timbre transfer that generalizes across instruments without retraining. Contrary to most related work, our method departs from encoder-based latent representations and instead decomposes input audio into explicit features such as pitch (F0), loudness, and timbre embeddings. The key contributions are:

- A simple projection that maps perceptually grounded, low-dimensional timbre embeddings into harmonic excitation signals allowing for continuous control.
- A conditioning-based decoder that filters pitch-aligned excitation signals via instrument-specific filters, inspired by DDSP but extended to generalization across several instruments.
- An efficient, streamable F0 predictor trained via teacher-student distillation for accurate pitch tracking in real-time.

The paper is organized as follows: Section II reviews related work and current limitations. Sections III and IV present our methodology and training strategy. Sections V–VII offer evaluation, real-time considerations, and concluding insights. Audio examples, supporting code and model weights are available on our project page[1].

II. Background

The main challenge for deep generative models in creative tasks has shifted from quality to interpretability and control. A common solution is conditioning, either via text using transformers [6] or self-supervised style transfer [7]. However, abstract concepts like timbre and musical style limit the precision of such models while computational demands restricts usability in real-time settings. Efficient, controllable timbre transfer thus remains an underexplored niche that only a few works have investigated. FaderRAVE [8] extends RAVE with continuous control over descriptors using fader networks and a latent discriminator. P-RAVE [9] adds pitch conditioning to the RAVE decoder using feature-wise linear modulation (FiLM) layers [10] and multi-band excitation. DDSP [5] introduces explicit conditioning on pitch and loudness, predicting harmonics and filter coefficients via a neural source-filter model. A sawtooth-based excitation [11] later improved the DDSP output for singing voices.

Other approaches focus on learned timbre spaces. As an example, the authors in [12] map MIDI-based note data into an instrument-conditioned mel-spectrogram and uses WaveNet

The work was done during an internship at Neutone AI. Funding has been provided by the Innovationfund of Denmark, 2052-00035B.

[1] Webpage: unified-timbre-transfer.github.io

979-8-3315-7206-8/25/$31.00

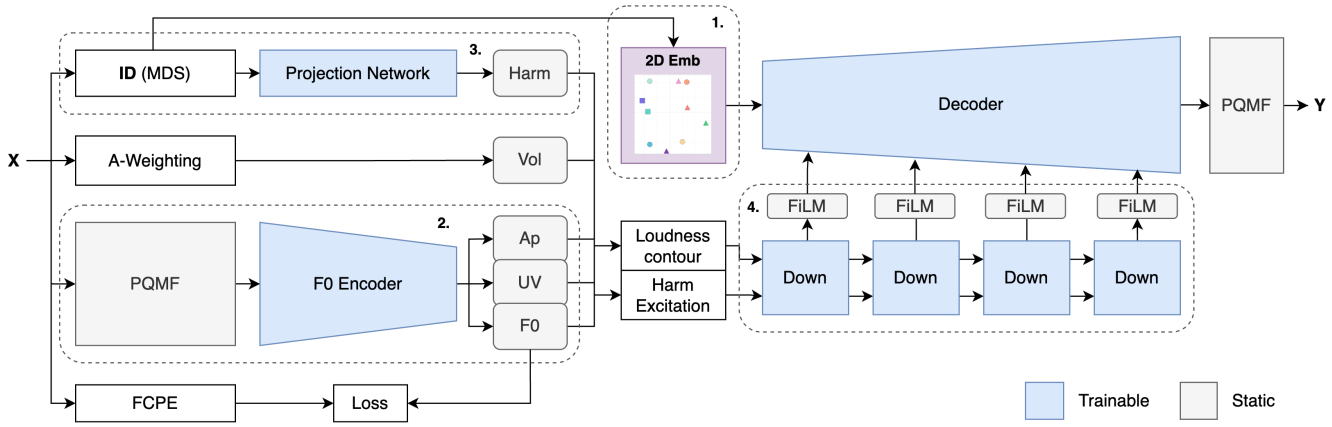

Fig. 1. Overview of the proposed method divided into the four key components described: 1. Perceptual 2D instrument/timbre space retrieved from instrument ID, 2. Supervised neural F0 extraction, 3. Prediction of instrument/timbre based harmonics for the excitation, 4. Excitation downsampling and conditioning.

for synthesis. While this work enables multi-instrument timbre morphing, it relies on MIDI, lacks real-time capability, and is constrained by vocoder quality.

Lastly, recent diffusion-based models such as [13] unify control and timbre transfer, extracting semantic features for joint musical and timbre representation. Though it achieves strong results and generalizes well, it lacks an interpretable timbre space and is not targeted streamable usage.

III. Proposed Method

We approach multi-instrument timbre transfer based on the assumption that a monophonic music instrument can be divided into three individual global features: timbre, pitch, and dynamics (loudness). Specifically, we exploit an inductive bias by guiding the audio generation through instrument-dependent excitation signals. As mentioned in [9], this mechanism enhances the learning process compared to strictly F0 conditioned models. Additionally, excitations have been popular choices in the field of singing voice conversion [14]. Building on similar work, we introduce a new strategy for learning an excitation signal, where its harmonics are predicted based on the instrument label. We then downsample this signal to condition the upsampling layers of our decoder.

Our model is built on four key components illustrated in Figure 1: 1. A perceptual 2D timbre embedding space (instrument ID) enabling real-time sound morphing. 2. Streamable neural extraction of F0, voicing (UV) and aperiodicity (AP). 3. Harmonic estimation for additive excitation synthesis, and 4. Double FiLM conditioning incorporating both pitch and loudness into the decoder/generation process. We demonstrate that this unified model, trained on a multi-timbre dataset, achieves output quality comparable to that of baselines trained individually. Additionally, we present a thorough ablation analysis highlighting the effects of our additions.

A. Harmonic Excitation Generation

The perceptual and timbral characteristics of sounds, such as ”brightness”, are closely linked to their harmonic content or ”overtones”. To enhance this inductive bias, we develop an instrument-specific excitation method that learns different harmonics for each target. The approach leverages the fundamental frequency (F0), aperiodicity, and the harmonic amplitudes which are directly predicted from the 2D timbre embedding. The excitation signal is generated using an additive sinusoidal oscillator, defined by:

$$E[n] = \begin{cases} A_p \eta[n], & \text{if } f_0[n] = 0, \\ A[n] \sum_{k=1}^{K} c_k[n] \sin(\phi_k[n]), & \text{otherwise,} \end{cases} \quad (1)$$

$$\phi_k[n] = 2\pi \sum_{m=0}^{n} \frac{k f_0[m]}{f_s} + \phi_{0,k}. \quad (2)$$

Here, equation (1) is the excitation signal $E[n]$, which is either scaled noise $A_p\eta[n]$ in unvoiced regions or a harmonic sum in voiced regions. Each sinusoid is defined by a harmonic coefficient $c_k[n] \in [0, 1]$, a global amplitude $A[n]$, and a phase $\phi_k[n]$. The phase $\phi_k[n]$ in equation (2) approximates the integral of instantaneous frequency $k f_0[n]$, ensuring phase continuity, with f_s as the sampling rate. $\eta \sim \mathcal{N}(0, 1)$ defines Gaussian noise with zero mean and unit variance.

The instrument-specific harmonic amplitudes $c_k(n)$ seen in equation (1) are predicted using 3 stacked projection blocks consisting of a linear-layer, layer-normalization and leaky relu activations. The blocks are followed by a sigmoid activation limiting the amplitudes to be within the range of 0 and 1.

B. Neural F0 Extraction

Accurate F0 estimation is critical for high-quality excitation generation. Traditional real-time methods like YIN [15], based on DSP heuristics, often underperform on noisy inputs, uncommon instruments, or fast pitch modulations. To overcome these issues, we adopt a data-driven approach by training the encoder from [3] with a pitch classification objective, yielding a context-aware model akin to [16]. Unlike frame-independent models such as CREPE [17], our method avoids the need for viterbi decoding.

The encoder produces pitch logits over cent-scale bins, optimized using categorical cross-entropy. Aperiodicity is estimated via the entropy of the predicted distribution:

$$\hat{h} = -\frac{1}{\ln P} \sum_{i=0}^{P-1} p(y = c_i \mid x) \ln p(y = c_i \mid x), \tag{4}$$

where P is the number of pitch bins. To improve efficiency and focus the frequency range, audio is first processed using Pseudo-Quadrature Mirror Filters (PQMF), as in [3].

Rather than relying on labeled pitch datasets, which are often limited in size and domain relevance, we distill knowledge from a larger teacher model during training. At inference, pitch is extracted via argmax, while voiced/unvoiced decisions are made by thresholding the entropy:

$$v/uv_t = \begin{cases} 1, & \text{if } \hat{h} > \alpha \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

C. Timbre Embedding

Timbre embeddings of musical instruments are in most studies learned directly or represented through pretrained self-supervised models, as an example for zero-shot conversion or downstream MIR tasks [18]. We use a perceptually relevant 2D timbre space allowing for straightforward interaction. Similar to the work in [19], we use the notion of timbre spaces [20] to form perceptual 2D timbre embeddings. Timbre spaces are created by presenting pairs of sounds to test subjects that rate the perceptual dissimilarities between a set of instruments. The ratings are compiled into a dissimilarity matrix that through multi-dimensional scaling (MDS) can be projected onto a lower dimensional space. Rather than using the MDS scaled ratings for regularization as done in [19], we use the same 2D space directly as timbre embeddings as illustrated in figure 2.

D. Conditional generator

We upsample the two-dimensional timbre embedding space using an architecture analogous to the decoder presented in [3]. This architecture comprises four upsampling blocks, each consisting of transposed one-dimensional (1D) convolutions and residual layers of 1D convolutions. Between each upsampling block, conditioning information is incorporated via Feature-wise Linear Modulation (FiLM) modules [10]. The decoder is conditioned on two sources of information, namely the harmonic excitation signal and an upsampled loudness contour.

In the FiLM conditioning mechanism, each conditioning feature is downsampled sequentially through a 1D convolution to generate a scale and shift vector. These vectors are subsequently used to apply a linear transformation to the hidden features within each upsampling block. Let E denote the sine-excitation signal, L the loudness features, and U the hidden upsampled features. FiLM is then defined by Equation (6):

$$\hat{U} = (\gamma_E + \gamma_L) \odot U + \varepsilon_E + \varepsilon_L. \tag{6}$$

Here γ_E and ε_E represent the affine parameters associated with the excitation signal, while γ_L and ε_L correspond to the parameters associated with the loudness contour. The operator $\odot$ denotes the Hadamard (element-wise) product.

IV. Experiments

We evaluate our model's ability to match target timbre and structure while maintaining audio quality, and conduct an ablation study to assess the impact of key components.

The unified model is trained at 44.1 kHz for 2M steps with a batch size of 8 and a frame-length of 2048 samples. We use 32 harmonics for the excitation and initialise decreasing decoder channel-widths of 512, 256, 128 and 64 (≈13.2M parameters). We use the Adam optimizer with a learning rate of 1e-3 and train both a non-causal version for comparison with baselines and ablations, and a causal version for real-time use. For the causal version all non-causal convolutions are substituted with a causal alignment block. Training begins with a 200k-step warm-up using multi-scale STFT loss, followed by 1.8M steps with an added adversarial objective from [3].

While pitch extraction may be added to the global loss and optimized during the warm-up stage, we handle it by pretraining the F0-encoder (channels: 128, 64, 32, 16). The encoder outputs 1440 bins at 5 cents/bin, following [21] and is trained for 400k steps on similar data as the main model. Pitch targets are derived using the Fast Context-Based Pitch Estimator (FPCE)[2] as shown in figure 1.

A. Datasets

The main model is trained on the URMP dataset [22] only, using 2 hours and 45 minutes of monophonic audio from 10 instruments that overlap with the MDS timbre ratings [19]: *bassoon, cello, clarinet, flute, horn, oboe, saxophone, trombone, trumpet*, and *violin*. For evaluation, we create a test subset by selecting two recordings per instrument (one for bassoon due to limited data), totaling 19 unseen samples. To assess generalization to out-of-domain data, we use the CocoChorales validation split [23]. Since the training data is exclusively monophonic, our model, and particularly the F0-encoder, is not designed to handle polyphonic audio.

To enable comparison with (m)any-to-one baseline models, we split the CocoChorales test audio into two subsets: (1) 10 randomly selected samples from 10 training-similar instruments for the ablation study, and (2) 100 randomly chosen 10-second clips from the *violin* and *trumpet* sets (n=200) for baseline evaluation.

B. Ablations

We examine the effect of model variations using the same training setup as the main model:

No-Ex: Baseline without excitation conditioning, matching the architecture in [3] but with a 64-dim timbre embedding concatenated to the latent space.

W-Enc: A P-RAVE-[9] based model trained with an encoder and variational objective, including loudness conditioning. The timbre embedding is also concatenated (dim=64).

No-Harm: A single sinusoid is used instead of a harmonic-based excitation to evaluate the role of the inductive bias.

[2]https://github.com/CNChTu/FCPE

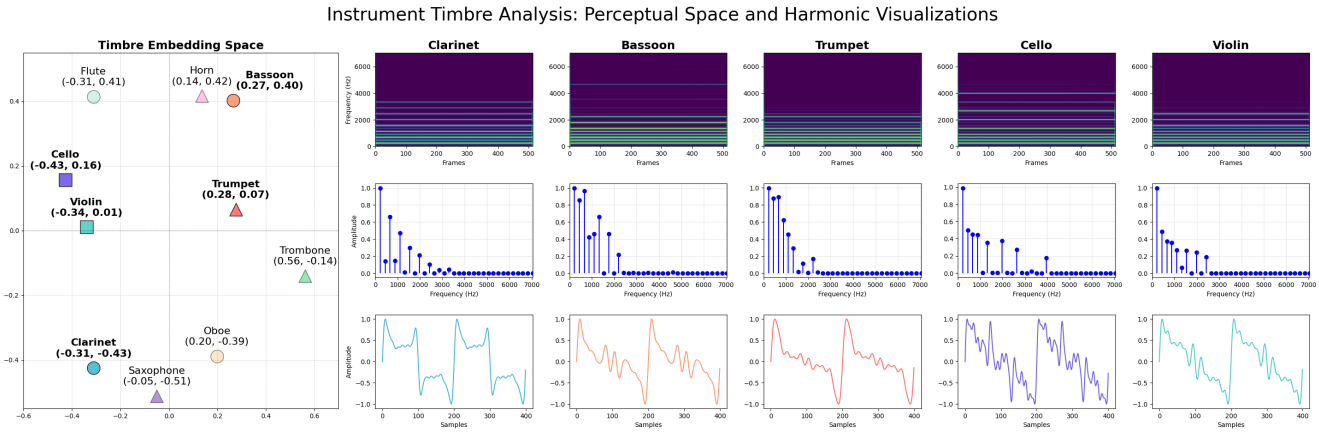

Fig. 2. **Left:** 2D instrument/timbre embeddings of MDS scaled dissimilarity metrics. Symbols provide instrument family: '◯' for woodwinds, '△' for brass, '□' for string instruments. **Right:** Harmonics learned by the proposed method for the excitation signals at 220 Hz corresponding to each highlighted instrument.

	Quality ↓		**Structure ↓**				**Transfer URMP ↓**			**Transfer Coco ↓**		
Model	*mSTFT*	*JND*	$\Delta F0$ *(Hz)*	ΔL *(dB)*	*JD*	*DPD*	*FAD*	*MMD*	$\Delta F0$ *(Hz)*	*FAD*	*MMD*	$\Delta F0$ *(Hz)*
No-Ex	**5.8**	**0.18**	17.73	**1.71**	0.17	0.1	4.62	0.12	103.46	7.33	0.23	125.87
Proposed	6.31	0.27	**8.14**	2.01	0.15	0.09	3.18	**0.12**	**8.96**	**4.05**	**0.19**	**4.71**
Pitch-Aug	6.66	0.29	8.97	2.63	**0.14**	**0.07**	4.28	0.15	9.02	5.36	0.23	6.14
W-Enc	6.03	0.22	8.58	1.94	0.15	0.09	**3.11**	0.14	10.30	5.00	0.31	5.35
No-Harm	6.37	0.28	10.02	2.37	0.15	0.09	3.68	0.13	10.65	4.34	0.21	4.85

TABLE I
ABLATION ANALYSIS COMPARING THE EFFECT OF THE ADDED COMPONENTS ON BOTH RECONSTRUCTION AND TIMBRE TRANSFER.

Pitch Aug: Includes pitch augmentation by randomly shifting pitch ±12 semitones in 35% of the training batches to broaden register coverage.

Ablations are evaluated on a reconstruction task using the URMP test data, while the timbre transfer capabilities of each model is evaluated using unseen in-domain (URMP) and out-of-domain (CocoChorales) data. We transfer each sample to all target instruments (n=190 for URMP, n=1000 for CocoChorales).

C. Timbre Transfer

We compare timbre transfer performance against two baseline approaches:

VAE-GAN: Unsupervised many-to-many timbre transfer with a universal encoder. The model is trained on URMP but only trumpet and violin checkpoints are released [24].

P-RAVE: Inspired by [9], the model is structurally similar to our method but without multi-instrument generalization. This method uses YIN [15] for F0 extraction.

D. Evaluation Metrics

We asses model performance using three key metrics:

Audio Quality: Measured via perceptually relevant spectral metrics: M-STFT and Just Noticeable Difference (JND).

Structure: Melody preservation is evaluated using Basic Pitch [25]. We generate pitch class probabilities and compare original and transferred melodies by Jaccard distance and Dynamic Pitch Distance (DPD), using DTW for temporal alignment. Also, average absolute deviations in fundamental frequency ($\Delta F0$) and loudness (ΔL) are computed.

Timbre Similarity: We measure timbre quality with Frechet Audio Distance (FAD) [26] using VGGish embeddings, and Maximum Mean Discrepancy (MMD) [27] between target and transferred MFCC distributions. The reference distribution is based on the full URMP training set.

V. RESULTS

A. Ablations

We evaluate the ablations on abovementioned metrics and summarize the results in Table I for both in-domain (URMP) and out-of-domain (CocoChorales) inputs. The RAVE baseline without excitation (*No-Ex*) shows strong reconstruction (mSTFT = 5.8, JND = 0.18, ΔL = 1.71) but poor timbre transfer on out-of-domain audio, likely due to its lack of excitation conditioning and limited structure preservation. Adding excitation conditioning significantly improves performance across all pitch and melody metrics. As shown in Figure 2, learned excitations reflect instrument characteristics: e.g., clarinet shows odd-harmonic focus while strings exhibit higher frequency amplitudes. Removing the harmonic structure (*No-Harm*) similarly degrades performance, particularly

		Trumpet Transfer						Violin Transfer					
		Timbre ↓		Structure ↓				Timbre ↓		Structure ↓			
Input	**Model**	*FAD*	*MMD*	$\Delta F0$	ΔL	*JD*	*DPD*	*FAD*	*MMD*	$\Delta F0$	ΔL	*JD*	*DPD*
Coco (V / T)	*VAE-GAN*	4.74	0.27	177.42	8.19	0.41	0.25	**5.21**	0.23	28.49	**4.84**	0.16	0.23
	P-RAVE	7.67	0.13	70.42	7.18	0.38	0.30	11.26	0.23	37.47	6.11	0.23	0.17
	Proposed	**3.40**	0.13	**3.81**	**6.18**	**0.06**	**0.01**	8.22	**0.14**	**4.69**	6.03	**0.03**	**0.01**
URMP (Multi)	*VAE-GAN*	4.70	0.29	79.25	7.27	0.35	0.53	7.24	0.25	130.08	10.86	0.37	0.56
	P-RAVE	3.23	0.1	41.68	4.92	0.28	0.22	7.75	0.18	55.98	6.74	0.31	0.41
	Proposed	**2.88**	0.1	**9.68**	**3.30**	**0.13**	**0.1**	**6.68**	**0.12**	**11.26**	**2.94**	**0.18**	**0.11**

TABLE II
TIMBRE TRANSFER COMPARISON WITH BASELINES. (V / T) REPRESENT THE VIOLIN AND TRUMPET INPUT USED FOR THE OPPOSITE TARGET INSTRUMENT, WHEREAS (MULTI) REPRESENT THE VARIETY OF INSTRUMENTS PRESENT IN THE URMP DATA.

FAD scores. This confirms the benefit of instrument-specific excitation in capturing timbral nuances and enhancing pitch accuracy through richer harmonic conditioning.

While the proposed method generally performs best, the version that includes an encoder (*W-Enc*) slightly improves in-domain timbre transfer (FAD = 3.11), likely by leveraging learned structure. However, the model reduces quality and accuracy on out-of-domain audio due to entangled latent representations highlighting the value of explicit disentanglement.

Lastly, it is clear that augmenting pitch during training offers marginal melodic improvement without enhancing timbre transfer. This effect likely arises from greater exposure to melodic variation rather than expanded timbral diversity.

B. Timbre Transfer

We evaluate timbre transfer between violin and trumpet using in-domain and out-of-domain data. Results show that the proposed method consistently outperforms baseline models in both audio quality and timbre similarity. Among baselines, *P-RAVE* excels on in-domain data, while *VAE-GAN* performs better on out-of-domain data. Our method, trained across diverse instruments, combines the strengths of both baselines. This is especially evident when transferring audio to the trumpet target, which achieves strong FAD scores (2.88 for in-domain audio, 3.40 for out-of-domain audio).

Timbre transfers with violin targets are generally experienced to be more challenging. While *VAE-GAN* achieves the best FAD for trumpet-to-violin, analysis of the generated audio reveals that the *VAE-GAN* model sporadically shifts F0 to align with the register of the target instrument, sacrificing structural consistency. This is reflected in lower structural scores compared to our method.

In summary, the proposed any-to-many strategy proves beneficial: unlike the one-to-one mapping in *P-RAVE*, multi-instrument learning improves generalization for both source and target instruments. Our F0-encoder also surpasses YIN (used in *P-RAVE*) and internal pitch/melody representations (used in *VAE-GAN*) in preserving melodic structure, likely due to its frame-dependence and data-driven design.

VI. REAL-TIME UTILIZATION

We integrate real-time utilization by training both the F0-encoder and the generator with a fully causal configuration. To address streamability, we replace the A-weighting loudness mechanism with a cached equivalent. We evaluate the causal version's timbre transfer abilities by assessing its performance on the same timbre transfer task as used in the URMP ablation study in Table I. As anticipated, Table III shows that the causal version exhibits a slight degradation in FAD performance, nonetheless, without any critical implications. Furthermore, the MMD values indicate that both models produce outputs that are relatively consistent with the target timbre.

Model	**Params**	**RTF**	*FAD*	*MMD*
Non-causal	15.1M	-	**3.18**	0.12
Causal	15.1M	2.5	3.59	0.12

TABLE III
PERFORMANCE OF NON-CAUSAL AND CAUSAL VERSIONS ON THE URMP TIMBRE TRANSFER TASKS.

The model is exported to a widely used torchscript-compatible host-plugin for real-time usage[3]. We urge the reader to listen to audio examples through the link provided in section I.

VII. CONCLUSION

We present a streamlined, any-to-many, multi-instrument timbre transfer model that disentangles timbre, pitch, and loudness. To our knowledge, it is the first model to enable real-time instrument-based timbre morphing with controllable pitch. Experiments show timbre transfer quality comparable with state-of-the-art baselines and specialized any-to-one models. We also find that a learned, instrument-dependent excitation signal improves both output quality and melodic structure. We hope this controllable timbre framework will inspire new directions in audio synthesis. In the future we hope to extend the model to polyphonic input.

[3]https://neutone.ai/fx

References

[1] A. v. d. Oord, S. Dieleman, H. Zen, *et al.*, *WaveNet: A generative model for raw audio*, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499.

[2] J. Engel, C. Resnick, A. Roberts, *et al.*, "Neural audio synthesis of musical notes with WaveNet autoencoders," in *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[3] A. Caillon and P. Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," *ArXiv*, vol. abs/2111.05011, 2021. [Online]. Available: https://arxiv.org/abs/2111.05011.

[4] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, *Stable audio open*, 2024. arXiv: 2407.14358 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2407.14358.

[5] J. Engel, L. (Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," in *International Conference on Learning Representations*, 2020.

[6] J. Copet, F. Kreuk, I. Gat, *et al.*, "Simple and controllable music generation," in *Advances in Neural Information Processing Systems*, 2023, pp. 47 704–47 720.

[7] O. Cífka, A. Ozerov, U. Şimşekli, and G. Richard, "Self-supervised VQ-VAE for one-shot music style transfer," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 96–100.

[8] N. Devis, N. Demerlé, S. Nabi, D. Genova, and P. Esling, "Continuous descriptor-based control for deep audio synthesis," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[9] S. Nercessian, "P-RAVE: Improving RAVE through pitch conditioning and more with application to singing voice conversion," in *Proc. Intl. Conf. on Digital Audio Effects*, 2023, pp. 383–386.

[10] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[11] D.-Y. Wu, W.-Y. Hsiao, F.-R. Yang, *et al.*, "DDSP-based singing vocoders: A new subtractive-based synthesizer and a comprehensive evaluation.," in *ISMIR*, 2022.

[12] J. W. Kim, R. Bittner, A. Kumar, and J. P. Bello, "Neural music synthesis for flexible timbre control," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 176–180.

[13] N. Demerle, P. Esling, G. Doras, and D. Genova, "Combining audio control and style transfer using latent diffusion," in *Proc. of the 25th Int. Society for Music Information Retrieval Conf.*, 2024.

[14] E. Song, K. Byun, and H.-G. Kang, "ExcitNet vocoder: A neural excitation model for parametric speech synthesis systems," in *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[15] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[16] L. Ardaillon and A. Roebel, "Fully-convolutional network for pitch estimation of speech signals," in *Interspeech 2019*, 2019, pp. 2005–2009.

[17] J. W. Kim, J. Salamon, P. Q. Li, and J. P. Bello, "CREPE: A convolutional representation for pitch estimation," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 161–165, 2018.

[18] K. Kim, J. Koo, S. Lee, H. Joung, and K. Lee, *TokenSynth: A token-based neural synthesizer for instrument cloning and text-to-instrument*, 2025. arXiv: 2502.08939 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2502.08939.

[19] P. Esling, A. Chemla–Romeu-Santos, and A. Bitton, "Generative timbre spaces: Regularizing variational auto-encoders with perceptual metrics," in *Proc. Intl. Conf. on Digital Audio Effects*, 2018.

[20] J. M. Grey and J. W. Gordon, "Perceptual effects of spectral modifications on musical timbres," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1493–1500, 1978.

[21] M. Morrison, C. Hsieh, N. Pruyne, and B. Pardo, *Cross-domain neural pitch and periodicity estimation*, 2023. arXiv: arXiv:2301.12258. [Online]. Available: https://arxiv.org/abs/2301.12258.

[22] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. 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, 2019.

[23] Y. Wu, J. Gardner, E. Manilow, I. Simon, C. Hawthorne, and J. Engel, "The chamber ensemble generator: Limitless high-quality MIR data via generative modeling," *arXiv preprint arXiv:2209.14458*, 2022.

[24] R. S. Bonnici, M. Benning, and C. Saitis, "Timbre transfer with variational auto encoding and cycle-consistent adversarial networks," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–8.

[25] R. M. Bittner, J. J. Bosch, D. Rubinstein, G. Meseguer-Brocal, and S. Ewert, "A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Singapore, 2022.

[26] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms," in *Interspeech*, 2019.

[27] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, pp. 723–773, Mar. 2012.