Text to Timbre - a Bibliography with Abstracts

Xavier Davenport (xavierd2@illinois.edu)
Quinn Ouyang (qouyang3@illinois.edu)
University of Illinois at Urbana-Champaign

December 30, 2023

References

[1] P. Altmann, L. Sünkel, J. Stein, T. Muller, C. Roch, and C. Linnhoff-Popien, "Sequent: Towards traceable quantum machine learning using sequential quantum enhanced training," in *International Conference on Agents and Artificial Intelligence*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:255522512

> This work proposes Sequential Quantum Enhanced Training (SEQUENT), an improved architecture and training process for the traceable application of quantum computing methods to hybrid machine learning and provides formal evidence for the disadvantage of current methods and preliminary experimental results as a proof-of-concept for the applicability of SEQUENT.

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

> Deep generative models applied to audio have improved by a large margin the state-of-the-art in many speech and music related tasks. However, as raw waveform modelling remains an inherently difficult task, audio generative models are either computationally intensive, rely on low sampling rates, are complicated to control or restrict the nature of possible signals. Among those models, Variational AutoEncoders (VAE) give control over the generation by exposing latent variables,

although they usually suffer from low synthesis quality. In this paper, we introduce a Realtime Audio Variational autoEncoder (RAVE) allowing both fast and high-quality audio waveform synthesis. We introduce a novel two-stage training procedure, namely representation learning and adversarial fine-tuning. We show that using a post-training analysis of the latent space allows a direct control between the reconstruction fidelity and the representation compactness. By leveraging a multi-band decomposition of the raw waveform, we show that our model is the first able to generate 48kHz audio signals, while simultaneously running 20 times faster than real-time on a standard laptop CPU. We evaluate synthesis quality using both quantitative and qualitative subjective experiments and show the superiority of our approach compared to existing models. Finally, we present applications of our model for timbre transfer and signal compression. All of our source code and audio examples are publicly available.

[3] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Neural audio synthesis of musical notes with wavenet autoencoders," 2017.

NSynth is an audio dataset containing 305,979 musical notes, each with a unique pitch, timbre, and envelope. For 1,006 instruments from commercial sample libraries, we generated four second, monophonic 16kHz audio snippets, referred to as notes, by ranging over every pitch of a standard MIDI pian o (21-108) as well as five different velocities (25, 50, 75, 100, 127). The note was held for the first three seconds and allowed to decay for the final second.

[4] P. Esling, A. Chemla-Romeu-Santos, and A. Bitton, "Bridging audio analysis, perception and synthesis with perceptually-regularized variational timbre spaces," in *International Society for Music Information Retrieval Conference*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:53873046

It is shown that Variational Auto-Encoders (VAE) can bridge the lines of research and alleviate their weaknesses by regularizing the latent spaces to match perceptual distances collected from timbre studies by proposing three types of regularization and showing that these spaces can be used for efficient audio classification.

[5] P. V. Itaboraí and E. R. Miranda, *Quantum Representations of Sound: From Mechanical Waves to Quantum Circuits.* Cham: Springer International Publishing, 2022, pp. 223–274.

This chapter discusses methods for the quantum representation of audio signals. Quantum audio is still a very young area of study, even within the quantum signal processing community. Currently, no quantum representation strategy claims to be the best one for audio applications. Each one presents advantages and disadvantages. It can be argued that quantum audio will make use of multiple representations targeting specific applications. The chapter introduces the state of the art in quantum audio. It also discusses how sound synthesis methods based on quantum audio representation may yield new types of sound synthesizers.

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

The recent success of raw audio waveform synthesis models like WaveNet motivates a new approach for music synthesis, in which the entire process - creating audio samples from a score and instrument information - is modeled using generative neural networks. This paper describes a neural music synthesis model with flexible timbre controls, which consists of a recurrent neural network conditioned on a learned instrument embedding followed by a WaveNet vocoder. The learned embedding space successfully captures the diverse variations in timbres within a large dataset and enables timbre control and morphing by interpolating between instruments in the embedding space. The synthesis quality is evaluated both numerically and perceptually, and an interactive web demo is presented.

[7] A. J. Milne, E. A. Smit, H. S. Sarvasy, and R. T. Dean, "Evidence for a universal association of auditory roughness with musical stability,"

PLOS ONE, vol. 18, no. 9, pp. 1–22, 09 2023. [Online]. Available: https://doi.org/10.1371/journal.pone.0291642

> We provide evidence that the roughness of chords—a psychoacoustic property resulting from unresolved frequency components—is associated with perceived musical stability (operationalized as finishedness) in participants with differing levels and types of exposure to Western or Western-like music. Three groups of participants were tested in a remote cloud forest region of Papua New Guinea (PNG), and two groups in Sydney, Australia (musicians and non-musicians). Unlike prominent prior studies of consonance/dissonance across cultures, we framed the concept of consonance as stability rather than as pleasantness. We find a negative relationship between roughness and musical stability in every group including the PNG community with minimal experience of musical harmony. The effect of roughness is stronger for the Sydney participants, particularly musicians. We find an effect of harmonicity—a psychoacoustic property resulting from chords having a spectral structure resembling a single pitched tone (such as produced by human vowel sounds)—only in the Sydney musician group, which indicates this feature's effect is mediated via a culture-dependent mechanism. In sum, these results underline the importance of both universal and cultural mechanisms in music cognition, and they suggest powerful implications for understanding the origin of pitch structures in Western tonal music as well as on possibilities for new musical forms that align with humans' perceptual and cognitive biases. They also highlight the importance of how consonance/dissonance is operationalized and explained to participants—particularly those with minimal prior exposure to musical harmony.

[8] A. Natsiou, L. Longo, and S. O'Leary, "Interpretable timbre synthesis using variational autoencoders regularized on timbre descriptors," 2023.

> Controllable timbre synthesis has been a subject of research for several decades, and deep neural networks have been the most successful in this area. Deep generative models such as Variational Autoencoders (VAEs) have the ability to generate a high-level representation of audio while providing a

structured latent space. Despite their advantages, the interpretability of these latent spaces in terms of human perception is often limited. To address this limitation and enhance the control over timbre generation, we propose a regularized VAE-based latent space that incorporates timbre descriptors. Moreover, we suggest a more concise representation of sound by utilizing its harmonic content, in order to minimize the dimensionality of the latent space.

[9] C. A. Nicol, "Development and exploration of a timbre space representation of audio," Ph.D. dissertation, University of Glasgow, 2005.

Sound is an important part of the human experience and provides valuable information about the world around us. Auditory human-computer interfaces do not have the same richness of expression and variety as audio in the world, and it has been said that this is primarily due to a lack of reasonable design tools for audio interfaces.There are a number of good guidelines for audio design and a strong psychoacoustic understanding of how sounds are interpreted. There are also a number of sound manipulation techniques developed for computer music. This research takes these ideas as the basis for an audio interface design system. A proof-of-concept of this system has been developed in order to explore the design possibilities allowed by the new system.The core of this novel audio design system is the timbre space. This provides a multi-dimensional representation of a sound. Each sound is represented as a path in the timbre space and this path can be manipulated geometrically. Several timbre spaces are compared to determine which amongst them is the best one for audio interface design. The various transformations available in the timbre space are discussed and the perceptual relevance of two novel transformations are explored by encoding "urgency" as a design parameter.This research demonstrates that the timbre space is a viable option for audio interface design and provides novel features that are not found in current audio design systems. A number of problems with the approach and some suggested solutions are discussed. The timbre space opens up new possibilities for audio designers to explore combinations of sounds and sound design based on perceptual cues rather than synthesiser parameters.

[10] L. Reymore, J. Noble, C. Saitis, C. Traube, and Z. Wallmark, "Timbre Semantic Associations Vary Both Between and Within Instruments: An Empirical Study Incorporating Register and Pitch Height," *Music Perception*, vol. 40, no. 3, pp. 253–274, 02 2023. [Online]. Available: https://doi.org/10.1525/mp.2023.40.3.253

The main objective of this study is to understand how timbre semantic associations—for example, a sound's timbre perceived as bright, rough, or hollow—vary with register and pitch height across instruments. In this experiment, 540 online participants rated single, sustained notes from eight Western orchestral instruments (flute, oboe, bass clarinet, trumpet, trombone, violin, cello, and vibraphone) across three registers (low, medium, and high) on 20 semantic scales derived from Reymore and Huron (2020). The 24 two-second stimuli, equalized in loudness, were produced using the Vienna Symphonic Library.Exploratory modeling examined relationships between mean ratings of each semantic dimension and instrument, register, and participant musician identity ("musician" vs. "nonmusician"). For most semantic descriptors, both register and instrument were significant predictors, though the amount of variance explained differed (marginal R2). Terms that had the strongest positive relationships with register include shrill/harsh/noisy, sparkling/brilliant/bright, ringing/long decay, and percussive. Terms with the strongest negative relationships with register include deep/thick/heavy, raspy/grainy/gravelly, hollow, and woody. Post hoc modeling using only pitch height and only register to predict mean semantic rating suggests that pitch height may explain more variance than does register. Results help clarify the influence of both instrument and relative register (and pitch height) on common timbre semantic associations.

[11] A. Rocchetto, E. Grant, S. Strelchuk, G. Carleo, and S. Severini, "Learning hard quantum distributions with variational autoencoders," *npj Quantum Information*, vol. 4, no. 1, p. 28, Jun 2018. [Online]. Available: https://doi.org/10.1038/s41534-018-0077-z

The exact description of many-body quantum systems represents one of the major challenges in modern physics, because it requires an amount of computational resources that scales exponentially with the size of the system. Simulating the evolution of a state, or even storing its description, rapidly becomes intractable for exact classical algorithms. Recently, machine learning techniques, in the form of restricted Boltzmann machines, have been proposed as a way to efficiently represent certain quantum states with applications in state tomography and ground state estimation. Here, we introduce a practically usable deep architecture for representing and sampling from probability distributions of quantum states. Our representation is based on variational auto-encoders, a type of generative model in the form of a neural network. We show that this model is able to learn efficient representations of states that are easy to simulate classically and can compress states that are not classically tractable. Specifically, we consider the learnability of a class of quantum states introduced by Fefferman and Umans. Such states are provably hard to sample for classical computers, but not for quantum ones, under plausible computational complexity assumptions. The good level of compression achieved for hard states suggests these methods can be suitable for characterizing states of the size expected in first generation quantum hardware.

[12] V. Rosi, A. Ravillion, O. Houix, and P. Susini, "Best-worst scaling, an alternative method to assess perceptual sound qualities," *JASA Express Letters*, vol. 2, no. 6, p. 064404, 06 2022. [Online]. Available: https://doi.org/10.1121/10.0011752

When designing sound evaluation experiments, researchers rely on listening test methods, such as rating scales (RS). This work aims to investigate the suitability of best-worst scaling (BWS) for the perceptual evaluation of sound qualities. To do so, 20 participants rated the "brightness" of a corpus of instrumental sounds (N=100) with RS and BWS methods. The results show that BWS procedure is the fastest and that RS and BWS are equivalent in terms of performance. Interestingly, participants preferred BWS over RS. Therefore, BWS is an alternative method that reliably measures perceptual sound qualities and could be used in many-sounds paradigm.

[13] K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, and R. R. Fay, Eds., *Timbre: Acoustics, Perception, and Cognition.* Springer Cham, 2019.

Outlines the principal perceptual processes that orchestrate timbre processing. Explores timbre as part of specific scenarios, including the perception of the human voice. Details computational acoustic models of timbre.

[14] K. Siedenburg, M. R. Schädler, and D. Hülsmeier, "Modeling the onset advantage in musical instrument recognition," *The Journal of the Acoustical Society of America*, vol. 146, no. 6, pp. EL523–EL529, 12 2019. [Online]. Available: https://doi.org/10.1121/1.5141369

Sound onsets provide particularly valuable cues for musical instrument identification by human listeners. It has remained unclear whether this onset advantage is due to enhanced perceptual encoding or the richness of acoustical information during onsets. Here this issue was approached by modeling a recent study on instrument identification from tone excerpts [Siedenburg. (2019). J. Acoust. Soc. Am. 145(2), 1078–1087]. A simple Hidden Markov Model classifier with separable Gabor filterbank features simulated human performance and replicated the onset advantage observed previously for human listeners. These results provide evidence that the onset advantage may be driven by the distinct acoustic qualities of onsets.

[15] K. Subramani and P. Rao, "Hprnet : Incorporating residual noise modeling for violin in a variational parametric synthesizer," 2020.

Generative Models for Audio Synthesis have been gaining momentum in the last few years. More recently, parametric representations of the audio signal have been incorporated to facilitate better musical control of the synthesized output. In this work, we investigate a parametric model for violin tones, in particular the generative modeling of the residual bow noise to make for more natural tone quality. To aid in our analysis, we introduce a dataset of Carnatic Violin Recordings where bow noise is an integral part of the playing style of higher pitched notes in specific gestural contexts. We obtain insights about each of the harmonic and residual components of the signal, as well as their interdependence, via observations on the latent space derived in the course of variational encoding of the spectral envelopes of the sustained sounds.

[16] K. Tatar, D. Bisig, and P. Pasquier, "Latent timbre synthesis," *Neural Computing and Applications*, vol. 33, no. 1, pp. 67–84, Jan 2021. [Online]. Available: https://doi.org/10.1007/s00521-020-05424-2

We present the Latent Timbre Synthesis, a new audio synthesis method using deep learning. The synthesis method allows composers and sound designers to interpolate and extrapolate between the timbre of multiple sounds using the latent space of audio frames. We provide the details of two Variational Autoencoder architectures for the Latent Timbre Synthesis and compare their advantages and drawbacks. The implementation includes a fully working application with a graphical user interface, called interpolate_two, which enables practitioners to generate timbres between two audio excerpts of their selection using interpolation and extrapolation in the latent space of audio frames. Our implementation is open source, and we aim to improve the accessibility of this technology by providing a guide for users with any technical background. Our study includes a qualitative analysis where nine composers evaluated the Latent Timbre Synthesis and the interpolate_two application within their practices.

[17] D. L. Wessel, "Timbre space as a musical control structure," *Computer Music Journal*, vol. 3, no. 2, pp. 45–52, 1979. [Online]. Available: http://www.jstor.org/stable/3680283