P-RAVE: IMPROVING RAVE THROUGH PITCH CONDITIONING AND MORE WITH APPLICATION TO SINGING VOICE CONVERSION

Shahan Nercessian

iZotope, Inc.
Boston, MA, USA
shahan@izotope.com

ABSTRACT

In this paper, we introduce means of improving fidelity and controllability of the RAVE generative audio model by factorizing pitch and other features. We accomplish this primarily by creating a multi-band excitation signal capturing pitch and/or loudness information, and by using it to FiLM-condition the RAVE generator. To further improve fidelity when applied to a singing voice application explored here, we also consider concatenating a supervised phonetic encoding to its latent representation. An ablation analysis highlights the improved performance of our incremental improvements relative to the baseline RAVE model. As our primary enhancement involves adding a stable pitch conditioning mechanism into the RAVE model, we simply call our method *P-RAVE*.

1. INTRODUCTION

Deep generative audio models aim to reconstruct and/or synthesize novel audio by learning its underlying data distribution. Since the inception of WaveNet [1], models have made considerable gains to improve fidelity, and achieve state-of-the-art realism in many domains. However, they are still largely considered too complex for widespread use, and offer limited controllability to end users.

Recently, the RAVE approach was introduced [2]. This variational autoencoder (VAE) has garnered excitement in the audio community due to its expressive synthesis, stable training procedure, favorable performance, and tractability for streaming applications running on edge devices [3], while modeling audio at sampling rates suitable for music production (i.e. $\geq$ 44.1 kHz). Despite this breakthrough, its baseline formulation can be prone to pitch glitches, especially when applied to out-of-domain input samples. Its latent representation may also still conflate timbral, pitch, and other factors without additional mechanisms to steer their disentanglement, limiting controllability.

Meanwhile, modern voice AI techniques have become emergent in research and pop culture [4]. Singing voice conversion (SVC) is one such application, whose goal is to transform sung material to match the timbre of some target singer while maintaining the source performance. SVC methods such as FastSVC [5] and [6] condition waveform generation on a harmonic excitation signal to counteract the fact that most neural generators lack sufficient pitch stability otherwise [7]. SawSing [8] considered a sawtooth excitation, and in our own work [9], we considered a hybrid end-to-end approach, where a differentiable WORLD synthesizer creates an initial synthesized output from inferred features, which is then further refined via a black-box postnet. Few SVC approaches are amenable to real-time streaming applications [10], so naturally, it is of interest to explore how the RAVE model can be refined for this use case.

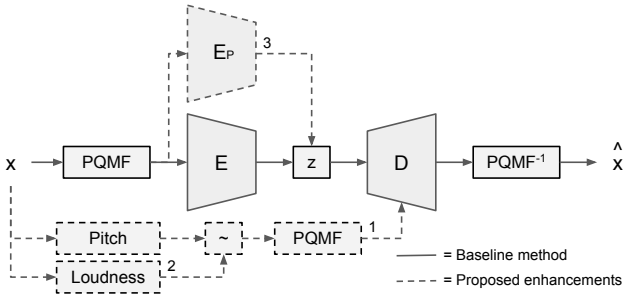

Figure 1: *RAVE model and proposed enhancements in P-RAVE.*

In this work, we offer improvements to the RAVE model. We apply insights from the SVC community to improve tonal signal reconstruction and/or generation in RAVE in a multi-band generator context [11], effectively conditioning its generator on excitation signals capturing pitch and/or loudness information. Accordingly, we call our method *P-RAVE*. Approaches are exemplified for an SVC application, and within this context, we also consider whether the model can benefit from supervised encodings of linguistic content. A byproduct of this work is an efficient phoneme recognizer that learns a feature representation from the time-domain waveform. Our goals are two-fold: we would like to improve the outputs of RAVE while maintaining its advantages, and to disentangle features such as pitch from its latent representation so that they are not conflated in the latent space and/or so that they can be controlled explicitly. In doing so, we are inherently investigating whether and/or how RAVE can be adapted for SVC. We illustrate the benefits of our enhancements relative to a standard RAVE baseline. We organize our paper as follows: Section 2 describes our proposed method, Section 3 reports experimental results, and Section 4 draws conclusions.

2. PROPOSED METHOD

The RAVE model, as well as our proposed additions in P-RAVE, are illustrated in Figure 1. Generally, an input signal x is mapped to a latent encoding z. The decoder aims to invert z, yielding the reconstructed waveform $\hat{x}$. The architecture utilizes a 16-band Pseudo Quadrature Mirror Filterbank (PQMF) [11], which aids model efficiency by allowing the core architecture to operate internally at a fraction (i.e. 1/16th) of the audio system sampling rate f_s. As in [2], we consider f_s = 48 kHz in this work. Signal reconstruction involves generation of an audio waveform from a suit-

able encoding z, along with specifications of any available control signals. Novel signal generation would additionally involve prediction of a relevant latent trajectory (via a second "prior" model). Our focus leans to the former task without restricting the latter.

P-RAVE improves upon RAVE by 1) FiLM conditioning [12] the RAVE generator with a multi-band harmonic excitation signal, 2) incorporating loudness information into said excitation, and 3) appending a supervised phonetic encoding alongside the learned latent representation, considering our particular interest in singing voice applications. We outline the motivation and implementation of each enhancement.

2.1. Harmonic excitation and FiLM conditioning

In order to provide stability and controllability of pitch, we leverage pitch-driven excitations as conditioning signals and adapt them to the multi-band RAVE generator. Combining the excitation generation approaches in [6, 8, 13], we generate the excitation e as

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

$$K = \lfloor \frac{f_s}{2 f_0[n]} \rfloor \quad (2)$$

$$\phi_k[n] = \phi_k[n-1] + 2\pi k \frac{f_0[n]}{f_s} \quad (3)$$

where $\eta \sim \mathcal{N}(0, 1)$ and f_0 is a fundamental frequency contour that can be user-specified or estimated directly from an input signal x. For the latter case, pitch is detected at a specified interval (we arbitrarily use a stride of 128 samples in this work), and upsampled to audio rate using a basic zeroth order interpolation in order to match the input audio waveform dimension. To this end, we leverage the `torchyin` library, which among other conveniences, ensures that the entire pipeline is constructed in PyTorch and can leverage the GPU more effectively during training.

We proceed by creating the multi-band excitation representation $e_{PQMF} = PQMF(e)$. Naturally, these sub-bands still operate at a faster sampling rate than the encoding z (in fact, it would match that of the hypothetical multi-band output signal estimate $\hat{x}_{PQMF}$). Accordingly, we apply FiLM conditioning to layers of the multi-band generator (decoder), as illustrated in Figure 2. We apply successive downsampling layers to e_{PQMF} according to the upsampling factors of each generator layer to ensure that they operate at the same rate. While downsampling could have been accomplished without trainable parameters, we opt to use strided convolutional layers instead, maintaining a constant 16-channel count for each downsampling stage. The outputs of each downsampling stage are subjected to respective 1x1 convolutional layers whose channels equal twice that of the corresponding generator upsampling layer which they are paired with. Output channels are split in half to form the scale and offset terms for FiLM conditioning. For an arbitrary scale γ, offset β, and upsampling layer y, the FiLM-conditioned output is given by

$$y_{FiLM} = \gamma \odot y + \beta \quad (4)$$

When scales and offsets are equal to unity and zero, respectively, FiLM conditioning sites act as pass-throughs. Unlike DDSP [14], note that we are not imposing for the model output to be strictly monophonic per se. The model is still fully capable of generating polyphonic audio, and therefore, we entirely maintain the generality of the RAVE system. We are simply adding a conditioning signal to steer generator upsampling layers, and if the model did not find useful information contained within it, could choose to ignore it. Nonetheless, when applied to intrinsically monophonic applications (e.g. solo voice), we expect that the model would learn to interpret it as an excitation signal (so we continue to refer to it as such), and to non-linearly filter it as a neural source-filter [13].

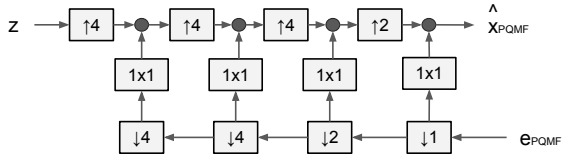

Figure 2: *Proposed multi-band FiLM conditioning applied in the P-RAVE generator. Solid black dots represent conditioning sites.*

2.2. Injection of loudness information into the conditioning

We can also incorporate loudness information into our pitched excitation signal. This is to say that its signal strength can be set to a user-specified value (a constant loudness, amplitude envelope, and/or mapping to MIDI velocity, as in [15]), or to match that of x. In the case of the latter, much like the f_0 computation in Section 2.1, we achieve this by measuring the frame-level root-mean-square (RMS) of x at some notional stride and upsampling it to full resolution, yielding the desired loudness trace L_0. If we similarly extract the RMS of e, yielding L_e, we can embed loudness information into a loudness-adjusted conditioning signal e_L via

$$e_L[n] = \left(\frac{L_0[n] + \epsilon}{L_e[n] + \epsilon} \right) e[n] \quad (5)$$

where $\epsilon = 10^{-5}$ is used for numerical stability. Accordingly, the multi-band excitation signal is then $e_{PQMF} = PQMF(e_L)$.

2.3. Supervised phonetic encoding

When trained for a voice application, the encoder is tasked with capturing not only pitch, loudness, and tone in a global sense, but also timbral changes which vary as a function of the phonetic unit being uttered. As this may prove challenging to accomplish sufficiently in a purely unsupervised manner, the final enhancement considered here concatenates the learned latent representation with a phonetic posteriorgram (PPG) capturing linguistic content.

We train a phonetic encoder on the TIMIT dataset [16] in a supervised manner. One subtlety here is that this dataset is natively sampled at 16 kHz, containing 8 kHz of bandwidth. Therefore, we upsample the data to audio rate (48 kHz in this case), and only consider the lower 5 PQMF sub-bands as input to the phonetic encoder, ideally covering 7.5 kHz with sufficient roll-off by 8 kHz. We use the condensed 40-class phonetic dictionary for the TIMIT dataset (39 phonemes and a silence class) [17], and train the phonetic encoder for 250K training steps before freezing it for the remainder of system training. Maintaining a total encoding size of 128, this leaves 88 latent dimensions to be learned in an unsupervised manner. Input and output sizes aside, the phonetic encoder E_P architecture is identical to that of the unsupervised encoder E. We reduce the number of channels in the two encoders by 50%, such that the number of parameters of their composite is effectively the same as in the encoder of the baseline RAVE model.

3. EXPERIMENTAL RESULTS

For subjective listening, we refer readers to our demo website at https://sites.google.com/izotope.com/prave-demo.

Generally, we observe that the VAE framework offers a sufficient information bottleneck in its latent space [18] for providing the speaker disentanglement needed for SVC. Accordingly, in order to illustrate the effectiveness of our proposed methods, we train models and analyze their robustness in this context. Models were trained on approximately two hours of internal singing voice data of a single target singer. Four different conversion models were considered: a baseline RAVE system, and three P-RAVE systems where we incrementally add our proposed enhancements, as per their enumeration in Section 2. We follow the training strategy outlined in [2], using a batch size of 8, Adam optimizer and its adversarial objective function. In the initial "warm-up" training phase, the encoder and decoder models are optimized jointly for 1M training steps, with the adversarial loss terms omitted. The (latent) encoder is then frozen and the decoder undergoes an additional 2M training steps which attempts to minimize the full objective.

We summarize our quantitative ablation analysis in Table 1, reporting the multi-spectrogram loss (MSL), the number of components needed to summarize 99% of the latent manifold (M_{99}) [2]. We see that our P-RAVE variants considerably improve reconstruction performance relative to the RAVE baseline, as measured by the MSL. Explicit incorporation of loudness information in P-RAVE (1+2) improves upon P-RAVE (1). Meanwhile, P-RAVE (1+2+3) technically sees slightly degraded quantitative performance on the self-reconstruction task relative to P-RAVE (1+2). We attribute this to the fact that the learned latent representation is catered to the target singer (in-domain distribution), while the PPG is a vocalist-independent representation to be leveraged by any source singer (out-of-domain distribution). Therefore, its inclusion still has favorable implications for our ultimate goal of the conversion task. P-RAVE (1) and P-RAVE (1+2) create significantly more compact latent feature representations relative to the baseline RAVE model, as the pitched and/or leveled excitation reduces the burden on the unsupervised encoder to fully model the feature space. The feature space is effectively reduced by one component between P-RAVE (1) and P-RAVE (1+2), hinting that conceivably, there may have indeed been a single latent dimension capturing loudness variations in the data. Interestingly, addition of the PPG in P-RAVE (1+2+3) considerably increases M_{99}. Though seemingly unintuitive, we explain this by noting that the information contained within PPGs arguably reflects the majority of the voice modeling task beyond pitch and loudness. Therefore, by now offloading the unsupervised encoder of its primary modeling "duties", P-RAVE (1+2+3) reduces the posterior collapse effect overall, allowing its unsupervised encoder to concentrate its degrees of freedom to modeling a smaller subspace of the variability in the data in a way that better matches the prior. This can be confirmed by observing that the Kullback-Liebler divergence against the prior for RAVE and P-RAVE (1+2+3) are 12.01 and 1.749, respectively.

Next, we analyze the preservation and controllability of conditioning features across different models for both self-reconstruction (in-domain source vocalist) and conversion (out-of-domain source vocalist), as listed in Table 2. Specifically, we compare conditioning features extracted from source and synthesized performances, measuring their average absolute deviation in fundamental frequency (ΔF in Hz) and loudness (ΔL in dB), and average categorical cross-entropy between source and synthesized PPGs (ΔCE in nats). In the in-domain case, we see that the baseline RAVE model performs decently, though features are better preserved in P-RAVE variants, where P-RAVE (1+2+3) appears to provide the best balance across all features. Within the conversion context, we further discern between whether or not we apply a pitch shift to f_0 so that the converted result is reflective of the register of the vocalist, and report results for both cases. Here, the baseline RAVE model struggles considerably, as it cannot maintain consistent pitch when the source register does not match the target data, and moreover, does not possess an explicit mechanism for applying a pitch shift if it were needed to accomplish a convincing conversion. Again, we see that P-RAVE (1+2+3) is better suited for the application, with outputs roughly achieving their target values across all features.

4. CONCLUSIONS

In this paper, we suggested additions to the RAVE model which improved fidelity and controllability of the generative audio model, and applied it to a singing voice application. In future work, we would like to add further refinements in order to improve fidelity. For example, we may consider a loss term that encourages the model to produce Mel spectrogram-like representations at an intermediate generator layer, or integrate aspects of the very recent developments in [19]. We are also interested to get a better sense of the prominent factors the latent space has learned when it is relieved of modeling pitch, loudness, and phonetic content, and to envision what other forms of transformative audio processing this may be able to unlock. Lastly, we would like to investigate further application of our enhancements in the context of novel tonal content creation.

Table 1: *Quantitative ablation analysis comparing RAVE to our various enhancements in P-RAVE.*

Experiment	MSL	$M_{99\%}$
RAVE	8.568	9
P-RAVE (1)	7.267	4
P-RAVE (1+2)	**7.175**	**3**
P-RAVE (1+2+3)	7.227	14

Table 2: *Comparison of conditioning features between source and synthesized performances.*

	Reconstruction			Conversion (without/with pitch shift)		
Experiment	ΔF (Hz)	ΔL (dB)	ΔCE (nats)	ΔF (Hz)	ΔL (dB)	ΔCE (nats)
RAVE	3.351	3.667	0.0564	61.43 / 103.92	4.918 / 4.923	0.0769 / 0.768
P-RAVE (1)	**0.443**	2.362	0.0556	0.682 / 1.321	4.560 / 2.786	0.0731 / 0.071
P-RAVE (1+2)	0.596	**1.764**	0.0562	**0.237 / 0.791**	**1.922 / 1.958**	0.0706 / 0.0716
P-RAVE (1+2+3)	*0.628*	*1.993*	**0.0515**	*0.277 / 0.823*	*3.3922 / 2.0625*	**0.0696 / 0.0676**

5. REFERENCES

[1] A. van den Oord et al., "WaveNet: A generative model for raw audio," *arXiv:1609.03499*, 2016.

[2] A. Caillon and P. Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," *arXiv:2111.05011*, 2021.

[3] A. Caillon and P. Esling, "Streamable neural audio synthesis with non-causal convolutions," in *Int. Conf. on Dig. Aud. Eff. (DAFx)*, 2022, pp. 320–327.

[4] H.S. Choi, J. Yang, J. Lee, and H. Kim, "NANSY++: Unified voice synthesis with neural analysis and synthesis," in *Int. Conf. on Learn. Rep. (ICLR)*, 2023.

[5] S. Liu et al., "FastSVC: Fast cross-domain singing voice conversion with feature-wise linear modulation," in *IEEE Int. Conf. on Mul. and Ex. (ICME)*, 2021, pp. 1–6.

[6] H. Guo, Z. Zhou, F. Meng, and K. Liu, "Improving adversarial waveform generation based singing voice conversion with harmonic signals," in *IEEE Int. Conf. on Ac., Speech and Sig. Proc. (ICASSP)*, 2022, pp. 6657–6661.

[7] B. Di Giorgi, M. Levy, and R. Sharp, "Mel spectrogram inversion with stable pitch," in *Int. Soc. for Mus. Inf. Ret. Conf. (ISMIR)*, 2022.

[8] D.Y. Wu and Others, "DDSP-based singing vocoders: A new subtractive based synthesizer and a comprehensive evaluation," in *Int. Soc. for Mus. Inf. Ret. Conf. (ISMIR)*, 2022.

[9] S. Nercessian, "Differentiable WORLD synthesizer-based neural vocoder with application to end-to-end audio style transfer," in *154th Aud. Eng. Soc. Conv. (AES)*, 2023.

[10] S. Nercessian et al., "Real-time singing voice conversion plug-in," in *Int. Conf. on Dig. Aud. Eff. (DAFx)*, submitted, 2023.

[11] G. Yang et al., "Multi-band Melgan: Faster waveform generation for high-quality text-to-speech," in *IEEE Work. on Spok. Lang. Tech. (SLT)*, 2021, pp. 492–498.

[12] E. Perez et al., "FiLM: Visual reasoning with a general conditioning layer," in *AAAI Conf. on Art. Int.*, 2018.

[13] X. Wang and J. Yamagishi, "Using Cyclic Noise as the Source Signal for Neural Source-Filter-Based Speech Waveform Model," in *Interspeech*, 2020, pp. 1992–1996.

[14] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," in *Int. Conf. on Learn. Rep. (ICLR)*, 2020, pp. 26–30.

[15] L. Renault, R. Mignot, and A. Roebel, "Differentiable piano model for MIDI-to-audio performance synthesis," in *Int. Conf. on Dig. Aud. Eff. (DAFx)*, 2022, pp. 233–239.

[16] J. S. Garapolo et al., *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*, Linguistic Data Consortium, Philadelphia, 1993.

[17] T.P. Van, H.N. Thanh, and T.M. Thanh, "Improving phonetic recognition with sequence-length standardized MFCC features and deep bi-directional LSTM," in *NAFOSTED Conf. on Inf. and Comp. Sci. (NICS)*, 2018, pp. 322–325.

[18] K. Qian et al., "AutoVC: Zero-shot voice style transfer with only autoencoder loss," in *Int. Conf. on Mach. Learn.*, 2019.

[19] N. Devis et al., "Continuous descriptor-based control for deep audio synthesis," in *IEEE Int. Conf. on Ac., Speech and Sig. Proc. (ICASSP)*, 2023.