RT-PAD-VC – CREATIVE APPLICATIONS OF NEURAL VOICE CONVERSION AS AN AUDIO EFFECT

Paolo Sani, Edgar Andres Suarez Guarnizo, Kishor Kayyar Lakshminarayana, and Christian Dittmar

Fraunhofer Institute for Integrated Circuits, IIS
Erlangen, Germany
{paolo.sani@iis.fraunhofer.de

ABSTRACT

Streaming-enabled voice conversion (VC) bears the potential for many creative applications as an audio effect. This demo paper details our low-latency, real-time implementation of the recently proposed Prosody-aware Decoder Voice Conversion (PAD-VC). Building on this technical foundation, we explore and demonstrate diverse use cases in creative processing of speech and vocal recordings. Enabled by it's voice cloning capabilities and fine-grained controllability, RT-PAD-VC can be used as a low-delay, quasi real-time audio effects processor for gender conversion, timbre and formant-preserving pitch-shifting, vocal harmonization and cross-synthesis from musical instruments. The on-site demo setup will allow participants to interact in a playful way with our technology.

1. INTRODUCTION

Voice conversion (VC) [1, 2] and singing voice conversion (SVC) aims to convert the voice recording of a speaker or singer (source) into a synthetic audio with the voice of another person (target) while preserving the original linguistic content and appropriately transferring the prosody or melody. Obviously, it is required to robustly extract those different speech properties from the source and process them independently for synthesizing the target. Since the desired information is usually not readily available in the source recording, it is often a challenge to disentangle these features. There are already a number of commercial applications[1] that use VC for dubbing of actors' voices into other languages or for anonymization of voice messages [3]. Similarly, several online services[2] already offer SVC capabilities for music production purposes. Although VC and SVC are often considered to be separate tasks, we will summarize both as VC in the remainder of this paper. As we will show, the main difference is the higher requirements for accurate conversion of the pitch trajectory in case of the singing voice, since it carries the melody. Nevertheless, both tasks can be effectively handled by a single VC system, if appropriately implemented. Although recent VC systems synthesize plausible speech signals that exhibit high similarity to the target speaker's voice, the complexity of the neural pipelines is often prohibitive for real-time applications like audio effects processors.

[1]For example: https://elevenlabs.io/
[2]For example: https://www.kits.ai/

1.1. Related Work

In general, significant performance improvements have been achieved in VC through the use of neural architectures and training paradigms. Early approaches to neural VC were based on AutoVC [4] or StarGAN [5]. These rely on a learnable disentanglement of speech features by enforcing information bottlenecks, which are in turn difficult to tune and can lead to leakage of undesired properties into the target speech signal. More recent approaches disentangle the speech features more explicitly through pre-trained feature extractors, either using abstract audio tokens based on self-supervised representation learning (SSL), or human-interpretable feature-representations such phoneme-posteriorgrams (PPGs).

For example, FreeVC [6] is a token-based VC system using the VITS framework. FACodec [7] factorizes speech into multiple attributes, like linguistic content, prosody, timbre and acoustics. One common theme in those approaches is the need for elaborate data augmentation strategies used in training to steer the networks away from undesired information leakage. Moreover, most of these approaches employ a direct reconstruction of the target time-domain signal from the sequence of audio tokens, making it hard to inspect failure cases.

In contrast, interpretable VC methods employ acoustic models to predict time-frequency representations (often mel-spectrograms) of the target speech, and use neural vocoders to synthesize the target time-domain signals. Our recently proposed Prosody-aware Decoder Voice Conversion (PAD-VC) [8] falls into this category. It is conceptually closest to the method presented in [9] whose authors show that a small and interpretable set of phonetically relevant speech features is sufficient to synthesize high quality target speech. PAD-VC also shares similarities with the approach presented in [10], which uses frame-wise formant estimates and other simple spectral envelope descriptors. The authors show that hand-crafted, classic audio features are a suitable representation for reconstructing intelligible and speech, even without using PPGs.

Only few works directly target streaming-enabled VC, where low-latency and real-time applicability are paramount over the voice conversion quality. As an example, the authors of [11] combine conventional pitch-shifting with neural spectral filtering to convert singing voices in real-time. Alternatively, [12, 13] combine audio tokens with light-weight differentiable DSP (DDSP). StreamVC [14] is a recent example of token-based and streaming-enabled VC, which is reported to deliver competitive conversion quality at latencies below 100 ms.

1.2. Contribution

The main purpose of this paper is to present creative audio effects applications of the streaming-enabled variant of PAD-VC, which

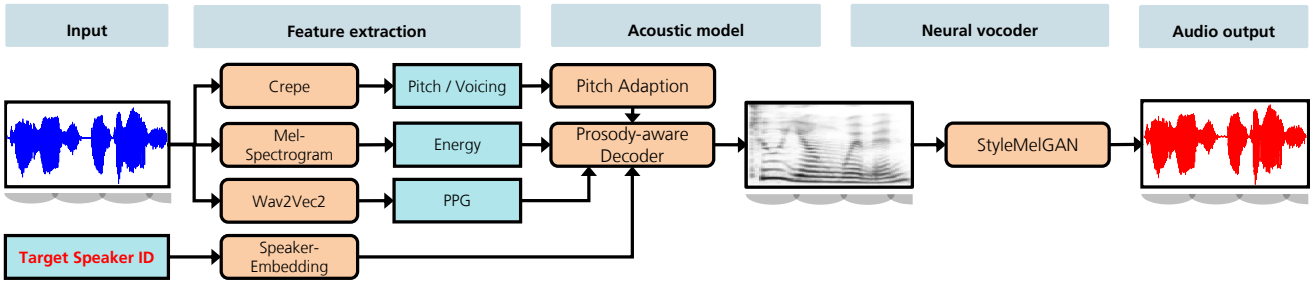

Figure 1: System Overview of PAD-VC and RT-PAD-VC. The main difference is the overlap-add style processing as indicated by the semi-transparent gray windows beneath the waveforms and mel-spectrogram. Blue: Source speaker utterance, Turquoise: Data streams, Apricot: Neural modules, Red: Target speaker utterance

we call RT-PAD-VC. We will recap the details of PAD-VC in Section 2.1 and describe our quasi real-time implementation in Section 2.2. Finally in Section 3, we will lay out expressive and creative voice manipulations that go beyond traditional audio effects and are made possible through RT-PAD-VC. We plan on setting up the demo at the DAFx 2025 venue, so that the participants can interact with RT-PAD-VC in an explorative and playful way.

2. METHOD

In the following paragraphs, we first describe how PAD-VC works and subsequently go into the details of the quasi real-time implementation of RT-PAD-VC.

2.1. Prosody-aware Decoder Voice Conversion

Our recently proposed PAD-VC [8] is a conceptually straightforward neural VC method that is able to disentangle (i.e., independently control) the spoken words or sung lyrics (phonetic content) from the speaking or singing style (prosodic variation) and the speaker's or singer's identity (voice timbre). This is achieved by extracting frame-wise, interpretable features from the source recording and using the decoder to predict the mel-spectrogram of the target speaker's speech or singing. We refer to Figure 1 to provide an overview of the complete processing pipeline. In the following, we will stick to the color-scheme of blue indicating the source speaker's utterance and red to indicate the target speaker's identity and utterance.

The speech features we employ can be roughly categorized into prosodic variation (pitch, energy, voicing confidence), phonetic content (PPG) and voice timbre (speaker embedding). Since they are low-dimensional, they can be understood as a natural information bottleneck. The acoustic model that acts as a decoder, trained to reconstruct the ground-truth mel-spectrograms of speech or singing recordings given their corresponding feature sequences. With the only exception of the speaker embeddings, all input features are either coming from fixed pre-trained models or completely DSP-based. This helps to minimize information leakage between the phonetics, prosody and timbre domains.

We extract prosodic variation features in a frame-rate of approx. 86 Hz from the source recording. They comprise energy (computed as the L2-norm of mel-spectrogram frames), pitch (f_0 in Hertz), and voicing confidence (saliency of the pitch estimate, estimated via CREPE [15]). Since the average pitch varies among individuals due to factors like age and gender, we scale the pitch trajectories to match the statistics of the target speaker. Unlike other works, we perform this step explicitly instead of relying on the speaker embeddings to capture the pitch statistics.

We extract PPGs using a variant of Wav2Vec [16] that has been fine-tuned for recognition of phonemes in American English[3]. The rationale behind using PPGs is to have an interpretable mid-level representation of the spoken words or sung lyrics that is not as rigid as discrete symbolic phoneme sequences but rather a soft activation of phoneme occurrences over time. We illustrate the correspondence between forced-aligned phoneme sequences and PPGs in Figure 2(a).

The speaker embeddings are different from the other features as they are not extracted from the source audio, but instead trained alongside the complete system. We use 64 dimensions to represent the voice timbre of the target speakers. Thus, PAD-VC does not have one-shot capabilities in terms of cloning a target voice. However, we found experimentally that PAD-VC can be fine-tuned with a new target voice using comparatively little training data (e.g., two minutes). Systematic evaluation of this aspect, especially the attainable synthesis quality versus availability of training data is subject to future work.

From the predicted mel-spectogram of the target speaker or singer, we reconstruct the time-domain signal by using StyleMelGAN [17] as a neural vocoder. It is pre-trained with diverse speech corpora and can thus be regarded as universally applicable. We also found in previous works that StyleMelGAN exhibits some degree of robustness to the so-called oversmoothing effect [18]. Furthermore, we already showed in [19] that it can synthesize pitched voices outside the range of the training data, making it suitable for singing voices as well.

2.2. Quasi real-time implementation RT-PAD-VC

Referring to the flow diagram in Figure 1, we want to point out that the PAD-VC pipeline works similar to an audio effects processor. The source recording is provided at the input and transformed into the target voice at the output. Some control parameters allow to change the outcome by switching or adjusting internal data streams. Firstly, there is the target speaker ID and it's corresponding speaker embedding that allow to switch between different voice timbres at the output. Secondly, there is the possibility to manipulate the pitch trajectory, allowing subtle to drastic alterations of the output speech signal. Of course it is also possible to modify energy and voicing confidence, as well as the PPG features

[3] We refer the reader to `https://huggingface.co/vitouphy/wav2vec2-xls-r-300m-timit-phoneme` for details

(see for example [9]), but such manipulations are out-of-scope for this article.

In contrast to conventional audio effects, comparatively compute-intensive modules are involved in our processing chain, rendering it impractical to process the incoming audio sample-by-sample or frame-by-frame. Thus, our main approach to enable real-time processing with PAD-VC is to implement a chunk-wise overlap-add method that is both applied during feature extraction as well as during decoding and synthesis. More precisely, the incoming audio is processed in chunks that are integer multiples of the number of audio samples per analysis frame. The overlap on feature extraction side is realized by having the first frame of the current chunk identical to the last frame of the previous chunk. During synthesis, overlap-add is realized by blending linearly between those overlapping frames. The number of frames in each chunk can be used to trade-off between the latency of the overall processing chain and the voice conversion quality.

To illustrate how shorter chunks negatively impact the PPGs, we depict in Figure 2(a) the correspondence between the force-aligned phoneme sequence of the utterance "Locksmith" and the prosody and PPG features, extracted from the complete recording. To avoid clutter, we only included the phoneme symbols with the highest activations. The PPG was not trained to detect stress- and length-marks, instead giving high activations to the space symbol. Figure 2(b) depicts the result of processing the same recording in small chunks (32 frames, overlap region indicated by the gray vertical bar). While the prosody features stay largely the same, the PPG exhibits wrong phoneme activations and gaps close to the chunk borders. This can be explained by the fact that the Wav2Vec2-based PPG extractor is a sequence-to-sequence transformer architecture that will inevitably produce different outcomes depending on the sequence length and preceding context frames (the same is true for the decoder, which uses recurrent LSTM and CBHG blocks). On the contrary, the pitch, energy and voicing confidence features are extracted per analysis-frame without any dependency of their temporal context and are mostly not affected. A good trade-off between quality and latency can be achieved using a chunk size of 64 frames and an overlap of 1 frame, leading to an algorithmic delay of 0.75 seconds.

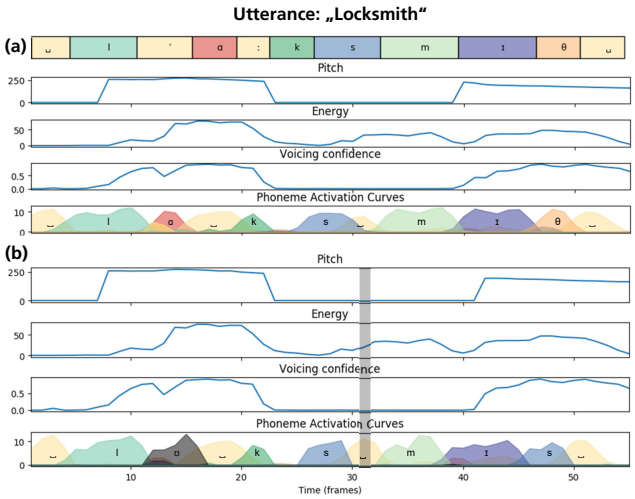

Figure 2: Illustrative example for the impact of **(a)** complete vs. **(b)** chunk-wise feature extraction, see 2.2 for further explanation.

3. CREATIVE APPLICATIONS

In this section, we will briefly describe the different creative applications that are possible through our streaming-enabled implementation of RT-PAD-VC. Two key capabilities of our system are important here. First, the ability to synthesize singing voice even when trained only with speech recordings. This can be attributed to the robust reproduction of desired pitch trajectories. Second, the possibility to fine-tune the decoder towards previously unseen target voices using comparatively little training data. From a practical point of view, we implement a client and server architecture to enable processing of the compute-intensive tasks on remote hardware, while having a light-weight client on device that takes care of audio capturing from a microphone input or other audio stream, sending the audio to the server and playing back or storing the resulting audio into a file. Audio examples for creative applications of RT-PAD-VC can be found on our accompanying website[4].

3.1. Gender conversion

This application is usually demonstrated as a sanity check for any new VC system. As a specialty of RT-PAD-VC, we can process the pitch trajectory and the target voice timbre in a completely decoupled fashion. We can thus gradually morph between the pitch range of the source and the target. In combination with the embedding-based modeling of the target timbre, this can be used to create special effects like an aged voice or a falsetto voice.

3.2. Pitch shifting

If the ID of the source speaker or singer happens to be the same as the target speaker, we can use PAD-VC as a means to perform pitch-shifting in a timbre- and formant-preserving manner. Although this use-case my seem kind of far-fetched, it can still be useful in situations where speech or singing voice recordings are to be re-edited. From the available audio material, we can fine-tune RT-PAD-VC to the target voice and then perform fine-granular edits (e.g., intonation correction).

3.3. Vocal harmonization

By running several instances of RT-PAD-VC, we can generate vocal harmonies to singing voice at the input. Alternatively, we can flatten the pitch trajectory of spoken input and replace it with a musically meaningful melody, thus creating a choir. Especially in the real-time context, engaging call-and-response interactions are possible where the user shouts a short phrase and an ensemble of synthetic voices repeats it with different pitches.

3.4. Cross-synthesis

RT-PAD-VC also synthesizes interesting results when the input audio is neither speech nor singing voice. When processing recordings of monophonic melody instruments, the instrument timbre will be matched with the most similar phoneme sequence and produce some output that resembles scat-singing. In case of percussive instruments, RT-PAD-VC will generate click and pop sounds that occur in speech and are most similar to the percussion. In case of polyphonic music, rather unexpected and chaotic results are synthesized.

[4] https://www.audiolabs-erlangen.de/resources/NLUI/2025-DAFx-RT-PAD-VC

4. CONCLUSIONS

We introduced RT-PAD-VC, our approach to streaming-enabled VC which combines chunk-wise extraction of features capturing the linguistic content and prosody of the input voice recording with an overlap-add application of the acoustic model and neural vocoder. We showed how manipulation of intermediate representations inside the system can be used to realize interesting audio effects that allow to alter speech or singing voice with low-latency. Future work will be directed towards adapting the involved neural architectures to non-causal processing while increasing the conversion quality also for the real-time case.

5. ACKNOWLEDGMENTS

This research was partially supported by the Free State of Bavaria in the DSAI project and by the Fraunhofer-Zukunftsstiftung.

6. REFERENCES

[1] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.

[2] Tomasz Walczyna and Zbigniew Piotrowski, "Overview of voice conversion methods based on deep learning," *Applied Sciences*, vol. 13, no. 5, pp. 3100, 2023.

[3] Ünal Ege Gaznepoglu and Nils Peters, "Deep learning-based F0 synthesis for speaker anonymization," in *Proc. of the European Signal Processing Conf. (EUSIPCO)*, Helsinki, Finland, 2023, pp. 291–295, IEEE.

[4] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, "AutoVC: Zero-shot voice style transfer with only autoencoder loss," in *Proc. of the Int. Conf. on Machine Learning (ICML)*, Long Beach, California, USA, 2019, vol. 97 of *Proceedings of Machine Learning Research*, pp. 5210–5219, PMLR.

[5] Suhita Ghosh, Arnab Das, Yamini Sinha, Ingo Siegert, Tim Polzehl, and Sebastian Stober, "Emo-StarGAN: A semi-supervised any-to-many non-parallel emotion-preserving voice conversion," in *Proc. of the Annual Conf. of the Int. Speech Communication Association (Interspeech)*, Dublin, Ireland, 2023, pp. 2093–2097.

[6] Jingyi Li, Weiping Tu, and Li Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *Proc. of the Int. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5.

[7] Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao, "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," 2024.

[8] Arunava Kr. Kalita, Christian Dittmar, Paolo Sani, Frank Zalkow, Emanuël A. P. Habets, and Rusha Patra, "PAD-VC: A prosody-aware decoder for any-to-few voice conversion," in *Proc. of the Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, 2024, pp. 389–393.

[9] Cameron Churchwell, Max Morrison, and Bryan Pardo, "High-fidelity neural phonetic posteriorgrams," in *Proc. of the Annual Conf. of the Int. Speech Communication Association (Interspeech)*, Seoul, Korea, 2024.

[10] Pablo Pérez Zarazaga, Zofia Malisz, Gustav Eje Henter, and Lauri Juvela, "Speaker-independent neural formant synthesis," in *Proc. of the Annual Conf. of the Int. Speech Communication Association (Interspeech)*, Dublin, Ireland, 2023, pp. 5556–5560.

[11] Shahan Nercessian, Russell McClellan, Cory Goldsmith, Alex M. Fink, , and Nicholas LaPenn, "Real-time singing voice conversion plug-in," in *Proc. of the Int. Conf. on Digital Audio Effects (DAFx)*, Copenhagen, Denmark, 2023.

[12] Shahan Nercessian, "End-to-end zero-shot voice conversion using a DDSP vocoder," in *Proc. of the Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 1–5.

[13] Anders Ridderholm Bargum, Simon Lajboschitz, and Cumhur Erkut, "RAVE for speech: Efficient voice conversion at high sampling rates," in *Proc. of the Int. Conf. on Digital Audio Effects (DAFx)*, Guildford, UK, 2024, pp. 41–48.

[14] Yang Yang, Yury Kartynnik, Yunpeng Li, Jiuqiang Tang, Xing Li, George Sung, and Matthias Grundmann, "Streamvc: Real-time low-latency voice conversion," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, 2024, pp. 11016–11020.

[15] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello, "CREPE: A convolutional representation for pitch estimation," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018, pp. 161–165.

[16] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "Wav2Vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[17] Ahmed Mustafa, Nicola Pia, and Guillaume Fuchs, "StyleMelGAN: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Toronto, Ontario, Canda, 2021, pp. 6034–6038.

[18] Paolo Sani, Judith Bauer, Frank Zalkow, Emanuël A. P. Habets, and Christian Dittmar, "Improving the naturalness of synthesized spectrograms for TTS using GAN-based post-processing," in *Proc. of the ITG Conf. on Speech Communication*, Aachen, Germany, 2023, pp. 270–274.

[19] Judith Bauer, Frank Zalkow, Meinard Müller, and Christian Dittmar, "Evaluating the impact of prosody feature normalization on the controllability of pitch in speech synthesis," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, Regensburg, Germany, 2024, pp. 188–195.