REAL-TIME DRUM-TO-VOCAL PERCUSSION SOUND CONVERSION SYSTEM

Rinka Nobukawa[1,2] **Tomohiko Nakamura**[2]
Shinnosuke Takamichi[3,1,2] **Hiroshi Saruwatari**[1]
[1] Graduate School of Information Science and Technology, University of Tokyo, Tokyo, Japan
[2] National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan
[3] Faculty of Science and Technology, Keio University, Yokohama, Japan
rinka-nobukawa@g.ecc.u-tokyo.ac.jp

ABSTRACT

Vocal percussion (VP) is a vocal technique that emulates drum sounds and plays a crucial rhythmic role in contemporary a cappella music. Despite its importance, synthesizing VP sounds remains difficult due to their noisy, non-linguistic characteristics, which conventional speech and singing voice synthesis methods fail to handle effectively. Our previous work framed VP sound synthesis as a timbre transfer task from drum to VP sounds, leveraging their functional correspondence. It also introduced an offline method using a variational autoencoder-based model called RAVE. In this paper, we propose a real-time drum-to-VP sound conversion system based on this offline method. The system processes input audio in chunks of 46 ms, enabling online operation. We demonstrate that the proposed system operates in real time on the central processing unit of a modern laptop computer.

1. INTRODUCTION

Vocal percussion (VP) is a form of vocalization that emulates the sounds of percussion instruments and plays a crucial rhythmic role in contemporary a cappella music [1]. To computationally handle contemporary a cappella performances, the analysis and synthesis of VP sounds are essential due to their central role in rhythm reproduction, and they have been studied in music-related research communities [2–5].

Unlike speech and singing voices, VP exhibits unique acoustic properties, such as aperiodicity, noisy transients, and a lack of linguistic content [2], which make it unsuitable for conventional speech and singing voice synthesis methods. VP production involves atypical use of articulatory organs and requires extensive practice to master (e.g., see [1]). Given these challenges, VP synthesis has the potential to support vocal practice and enhance musical expression in computer-assisted contemporary a cappella performances.

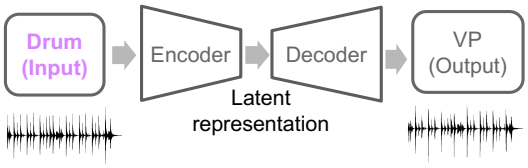


Figure 1: Inference procedure of offline drum-to-VP conversion method, trained only on VP sounds. This figure is adapted from [5].

In our previous work [5], we framed the VP synthesis problem as a timbre transfer task from drum to VP sounds. This formulation reflects the functional correspondence between drums and VP, and enables the application of timbre transfer methods to this task. Following this framework, we constructed a method using RAVE, a real-time audio variational autoencoder [6]. Using this method, we demonstrated its effectiveness in an offline scenario. However, the offline setup is insufficient for scenarios requiring real-time responsiveness, such as live performance or interactive vocal practice.

In this paper, we propose a real-time drum-to-VP sound conversion system based on our previously developed RAVE-based method. While the original implementation operated offline, we extend it to process incoming audio in fixed-length chunks for real-time operation. We demonstrate that the system runs in real time on the central processing unit (CPU) of a modern laptop computer.

2. OFFLINE DRUM-TO-VP SOUND CONVERSION METHOD

Our previously proposed offline method was built upon RAVE, a neural audio synthesizer based on a variational autoencoder (VAE) [6]. RAVE was originally developed for musical instrument sound synthesis. It encodes an input waveform into a latent representation and reconstructs it through a decoder network. To capture temporal characteristics, RAVE predicts an amplitude envelope separately from the waveform and applies it to modulate the decoder output. This architecture allows the model to preserve the

Late-Breaking / Demo Session Extended Abstract, ISMIR 2025 Conference

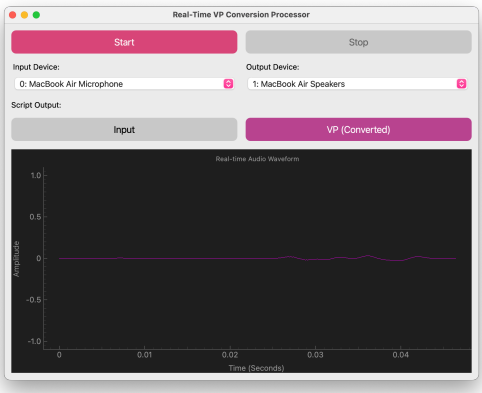
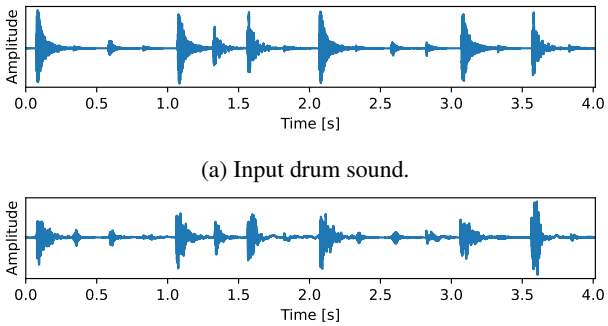


Figure 2: User interface of proposed real-time system.



(a) Input drum sound.

(b) Converted VP sound.

Figure 3: Waveforms of input drum and corresponding converted VP sounds.

temporal structure of the input signal.

RAVE can also be adapted for timbre transfer tasks. In our method, we trained RAVE solely on vocal percussion (VP) recordings, allowing the decoder to learn the target timbral characteristics. During inference, drum sounds are input to the encoder, and the decoder generates VP sounds that acoustically correspond to the input while preserving its rhythmic structure (see Figure 1). This approach does not require paired drum–VP data, which reduces data collection costs and makes it feasible to train the model using independently sourced audio samples. Interestingly, we experimentally found that the trained model could associate each drum instrument (e.g., bass drum and snare drum) with an appropriate VP counterpart, despite the absence of explicit supervision through paired training data.

3. PROPOSED REAL-TIME DRUM-TO-VP SOUND CONVERSION SYSTEM

We develop a real-time drum-to-VP sound conversion system based on the offline method described in Section 2. Specifically, we adopt a block-wise processing strategy and integrate the pretrained offline model into a real-time inference pipeline.

3.1 System Architecture

The system consists of three components: an audio input module, a conversion module, and an audio output module. The audio input module captures streaming drum audio from a microphone or line input and segments it into fixed-length chunks. Each chunk is then passed to the conversion module, which generates the corresponding VP waveform using the RAVE-based model. The audio output module receives the converted waveforms and plays them through a speaker on a per-chunk basis.

Although the core drum-to-VP conversion method remains the same as in the offline setting, we apply an implementation-level modification to support online processing. This modification addresses boundary discontinuities between adjacent chunks, which are typically caused by zero-padding in convolutional layers. Instead of zero-padding, we reuse feature values from the previous chunk to pad the current input. This padding reduces artifacts at chunk boundaries and improves the temporal continuity of the converted waveform. While not described in the original RAVE paper [6], this functionality is provided in the official RAVE codebase [1] .

We confirmed that the entire system operated in real time on the CPU of an Apple MacBook Air (M4). The algorithmic latency equals the chunk length (46 ms), and the average processing time per chunk was approximately 16 ms. Note that the measurement of the processing time was informal, and we leave the formal evaluation of computational performance for future work.

3.2 User Interface

Figure 2 shows the user interface for the proposed real-time conversion system. The control panel includes a central “Start/Stop” button to toggle the conversion process. Drop-down menus allow users to select audio input and output devices. The graph panel displays the converted waveform in real time, helping users verify signal input and monitor the system’s responsiveness during operation. A simple toggle switch enables optional monitoring of the original drum audio.

3.3 Sound Conversion Examples

Figure 3 shows waveform examples of an input drum sound and its converted VP counterpart. These examples illustrate that the proposed system preserves the rhythmic structure of the input while altering its timbral characteristics to resemble VP sounds.

4. CONCLUSION

We proposed a real-time drum-to-VP sound conversion system based on our previously developed RAVE-based offline method. By adopting a block-wise processing strategy and integrating online-aware implementation techniques, the system enables real-time conversion on the CPU of a modern laptop computer.

[1] https://github.com/acids-ircam/RAVE

5. ACKNOWLEDGMENT

This work was supported by JSPS Grants-in-Aid for Scientific Research JP23K18474, JP21H04900, JP23K28108, JST Fusion Oriented REsearch for desruptive Science and Technology (FOREST) JPMJFR226V and Tateisi Science and Technology Foundation.

6. REFERENCES

[1] R. Dietz, *A Cappella 101: A Beginner's Guide to Contemporary A Cappella Singing*. Milwaukee, Wisconsin: Hal Leonard Corporation, 2022.

[2] R. Blaylock, N. Patil, T. Greer, and S. S. Narayanan, "Sounds of the human vocal tract," 2017, pp. 2287–2291.

[3] M. Proctor, E. Bresch, D. Byrd, K. Nayak, and S. Narayanan, "Paralinguistic mechanisms of production in human "beatboxing": A real-time magnetic resonance imaging study," vol. 133, no. 2, 2013.

[4] A. Paroni, N. Henrich Bernardoni, C. Savariaux, H. Lœvenbruck, P. Calabrese, T. Pellegrini, S. Mouysset, and S. Gerber, "Vocal drum sounds in human beatboxing: An acoustic and articulatory exploration using electromagnetic articulography," vol. 149, no. 1, pp. 191–206, 01 2021.

[5] R. Nobukawa, M. Kitamura, T. Nakamura, S. Takamichi, and H. Saruwatari, "Drum-to-vocal percussion sound conversion and its evaluation methodology," *Proceedings of Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, (under review).

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