INTERSPEECH 2020: Show & Tell Contribution
October 25–29, 2020, Shanghai, China

Real-time, full-band, online DNN-based voice conversion system using a single CPU

Takaaki Saeki, Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari

Graduate School of Information Science and Technology, The University of Tokyo, Japan
{takaaki_saeki, shinnosuke_takamichi}@ipc.i.u-tokyo.ac.jp

Abstract

We present a real-time, full-band, online voice conversion (VC) system that uses a single CPU. For practical applications, VC must be high quality and able to perform real-time, online conversion with fewer computational resources. Our system achieves this by combining non-linear conversion with a deep neural network and short-tap, sub-band filtering. We evaluate our system and demonstrate that it 1) achieves the estimated complexity around 2.5 GFLOPS and measures real-time factor (RTF) around 0.5 with a single CPU and 2) can attain converted speech with a 3.4 / 5.0 mean opinion score (MOS) of naturalness.

Index Terms: voice conversion, full-band speech, real-time speech processing, online speech processing

1. Introduction

Voice conversion (VC) is the task of converting source speech into target speech while preserving linguistic information. Since VC has various applications (e.g., live-streaming, speech aid, etc.), many deep neural network (DNN)-based methods have been proposed for high-quality, flexible conversion using GPUs. From a practical standpoint, VC needs to be able to enhance converted-speech quality and achieve real-time, online conversion with fewer computational resources. Furthermore, the available frequency band needs to be extended from a conventional narrow band (16 kHz) [1] to a full band (48 kHz) that covers human audible range.

In this work, we propose a VC system that achieves real-time, online conversion for full-band speech using a single CPU. Our system is based on non-linear conversion with a DNN and direct waveform modification that uses spectral differentials [2]. To establish high-fidelity, computationally efficient VC, our system utilizes sub-band multirate signal processing [3] and filter truncation [4]. Furthermore, our system has an audio I/O for online VC and F0 transformation for cross-gender conversion. When we evaluated our VC system's computational performance and converted-speech quality, we found that 1) it achieves the estimated complexity around 2.5 GFLOPS and measures the real-time factor (RTF) around 0.5 with a single CPU, and 2) it can attain converted speech with a 3.4 / 5.0 mean opinion score (MOS) of naturalness.

In our Show & Tell video, the audience members will be able to experience both high quality and low latency with our system by listening to the original and converted speech in left and right headphones, respectively.

2. System architecture

2.1. Audio I/O for online VC

Figure 1 shows how our system performs online conversion. Our system receives a 5 ms waveform of the source speech and outputs a 5 ms waveform of the converted speech. In the conversion process described in Section 2.2, 25 ms windowed waveform is converted within 5 ms. The final converted waveform is output through the overlap-add process.

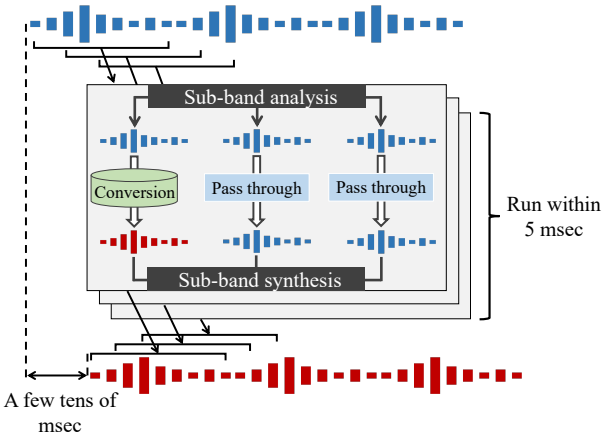

Figure 1: *Overview of our proposed real-time VC system*

2.2. Conversion process

Our method is based on direct waveform modification that uses spectral differentials [2]. We also use a DNN-based acoustic model for smaller-error conversion of vocal tract features. Spectral-differential VC that uses a minimum-phase filter [5] achieves high quality for narrow-band VC, but there are two problems when we extend the method to full-band VC: 1) converted-speech quality degrades due to fluctuations on the high-frequency band (as mentioned in our previous work [6]), and 2) computational cost is heavy (mainly in the filtering operation) due to increased sampling points.

Therefore, we introduce sub-band multirate signal processing [3] that improves converted-speech quality and reduces computational cost. First, we use sub-band multirate processing on the input audio to obtain sub-band signals on three bands, 0–8 kHz, 8–16 kHz, and 16–24 kHz. Then, only the lowest-band (0–8 kHz) signal is converted with a DNN, and the higher-band signals are passed through, as shown in Figure 1. The DNN estimates the real cepstrum of the differential filter from the real cepstrum of the source speech, and we obtain the converted speech by applying the differential filter constructed from the real cepstrum using a minimum-phase filter or a data-driven phase [6]. Finally, the full-band converted speech is synthesized from each sub-band signal. This operation reduces the computational cost and improves the converted-speech quality by modeling only the low-frequency band that contributes to speaker identity and avoiding high-frequency modeling. We can further reduce the computational cost in filtering by truncating the filter to a shorter tap length [4].

Table 1: *Estimated complexity of our proposed system and the conventional method in GFLOPS.*

	Sub-band	Cepstrum analysis	DNN inference	Hilbert trans.	Filtering	Other	Total
Ours (sub-band + truncation)	1.40	0.043	0.33	0.041	0.35	0.30	**2.5**
Ours (sub-band)	1.40	0.043	0.33	0.015	1.40	0.30	**3.5**
Conventional [5]	-	0.20	2.97	0.20	16.78	0.30	**20.5**

Table 2: *Results of experimental evaluations of performance and converted-speech quality.*

	RTF	MOS
Ours (sub-band + truncation)	**0.56**	**3.42**
Ours (sub-band)	**0.86**	**3.41**
Conventional [5]	3.36	2.62

2.3. F0 transformation

Since the method described in Section 2.2 can only convert vocal tract characteristics, we incorporate F0 transformation into our system for cross-gender conversion using direct waveform modification with PICOLA. This method is more computationally efficient and suitable for our purpose than a vocoder-based method.

3. Evaluations

3.1. Evaluation setup

We compared our system with the filter truncated to $1/4$ tap length, our system with the full-tap filter and the conventional method [5]. In the conventional method, the discrete Fourier transform (DFT) length was 2,048 samples, and the number of dimensions of the cepstrum was 120 (0th-through-119th). The DFT length in our system was 512, and the number of dimensions of the cepstrum was 40 (0th-through-39th).

The DNN architecture of the acoustic model was a multilayer perceptron consisting of two hidden layers. The numbers of each hidden unit were 280 and 100 in our system and 840 and 300 in the conventional method. The DNNs consisted of a gated linear unit that included the sigmoid activation layer and the tanh activation layer.

3.2. Complexity

In this section, we estimate the complexity of our system in FLOPS. Our real-time VC consists of sub-band processing, cepstrum analysis, inference with the DNN, the Hilbert transform, and filtering. The complexity of each process can be calculated from the parameters in Section 3.1. Furthermore, we considered around 0.3 GFLOPS complexity for other neglected calculations. Table 1 shows that the estimated complexity of our system is 2.5 GFLOPS, whereas the conventional method is around 20 GFLOPS, demonstrating that our system achieves the lower computational cost than LPCNet [7] for narrow-band (16 kHz) speech synthesis.

3.3. Experimental evaluations

To compare the processing time and converted-speech quality of our system with that of the conventional method [5], we built two intra-gender VC, for female-to-female (f2f) and male-to-male (m2m) conversion. The source and target speakers in the f2f case were stored in the JSUT corpus and Voice Actress Corpus, respectively. Those in the m2m case were stored in the JVS corpus. We used 100 utterances (approx. 12 min.) of each speaker, and the numbers of utterances for training, validation, and test data were 90, 5, 5, respectively. To investigate the effect sub-band processing and filter truncation have on converted-speech quality, we did not implement F0 transformation. However, in a preliminary experiment, we confirmed that the processing time of PICOLA was negligible compared to the total processing time of our system. In the demo video, we demonstrate the cross-gender conversion using our F0 transformation method.

To evaluate performance, we measured the value of the computation time divided by the length of the input waveform (RTF) with an Intel (R) Core i7-6859K CPU. Table 2 shows that our system achieved around 0.5 RTF, whereas the RTF with the conventional method exceeded 1.0, demonstrating that our system operates in real time with a single CPU.

To evaluate the converted-speech quality, we conducted a subjective evaluation of naturalness. Eighty listeners participated in each evaluation through our crowd-sourced evaluation systems, and each listener evaluated ten speech samples. Table 2 shows the average MOS value of m2m and f2f conversion. We confirmed that our VC system can synthesize high-quality converted speech with a 3.4 / 5.0 MOS, whereas the MOS is around 2.6 with the conventional method.

4. Conclusions

In this work, we presented our VC system that achieves real-time, high-quality, online conversion of full-band speech using only a single CPU. The results of our evaluation show that our system can operate in real time on mobile devices and achieve high-quality full-band converted speech. Our future work is increasing our system's flexibility and quality by applying data augmentation and other signal processing methods.

5. Acknowledgements

Part of this work was supported by the MIC/SCOPE #182103104.

6. References

[1] R. Arakawa, S. Takamichi, and H. Saruwatari, "Implementation of DNN-based real-time voice conversion and its improvements by audio data augmentation and mask-shaped device," in *Proc. SSW10*, Vienna, Austria, Sep. 2019, pp. 93–98.

[2] K. Kobayashi, T. Toda, and S. Nakamura, "Intra-gender statistical singing voice conversion with direct waveform modification using log-spectral differential," *Speech Communication*, vol. 99, pp. 211–220, 2018.

[3] R. Crochiere and L. Rabiner, *Multirate digital signal processing*. Englewood Cliffs, N.J. : Prentice-Hall, 1983.

[4] M. Sunohara, C. Haruta, and N. Ono, "Low-latency real-time blind source separation with binaural directional hearing aids," in *Proc. CHAT*, Stockholm, Sweden, Aug. 2017, pp. 9–13.

[5] H. Suda, G. Kotani, S. Takamichi, and D. Saito, "A revisit to feature handling for high-quality voice conversion," in *Proc. APSIPA ASC*, Hawaii, U.S.A., Nov. 2018, pp. 816–822.

[6] T. Saeki, Y. Saito, S. Takamichi, and H. Saruwatari, "Lifter training and sub-band modeling for computationally efficient and high-quality voice conversion using spectral differentials," in *Proc. ICASSP*, Barcelona, Spain, Mar. 2020, pp. 7784–7788.

[7] J. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *Proc. ICASSP*, Brighton, U.K., Feb. 2019, pp. 5891–5895.