10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

Implementation of DNN-based real-time voice conversion and its improvements by audio data augmentation and mask-shaped device

Riku Arakawa[1], Shinnosuke Takamichi[1], and Hiroshi Saruwatari[1]

[1] Graduate School of Information Science and Technology, The University of Tokyo, Japan.
riku.arakawa1996@gmail.com, {shinnosuke_takamichi,hiroshi_saruwatari}@ipc.i.u-tokyo.ac.jp

Abstract

Voice conversion (VC) enables us to change speech while preserving the linguistic information and is expected to play a significant role in augmented human communication. Recently, deep neural network (DNN)-based VC has been attracting attention because it can synthesize high-quality speech. However, existing methods typically assume offline processes (i.e., analysis, conversion, and synthesis) and cannot be directly applied to real-time VC. Therefore, we propose an implementation method of DNN-based VC that works online with low latency. We also propose audio data augmentation to improve the speech quality of real-time VC. Finally, we develop a mask-based real-time VC device to improve robustness against background noise. Experimental results demonstrate that 1) the proposed real-time VC works with 0.50 of the real-time factor, 2) the proposed data augmentation improves speech quality, and 3) the proposed mask-based VC device is more robust to noise than a standard microphone-based VC device.
Index Terms: deep neural network, DNN-based voice conversion

1. Introduction

Voice conversion (VC) [1, 2] is a technique to transform speech while preserving linguistic information and having the desired para-/non-linguistic information. It is expected to play a significant role in elevating human communication beyond their physical constraints. Recent approaches to VC include various machine learning techniques that produce high-quality converted speech. Moving beyond conventional Gaussian mixture model (GMM)-based VC [1, 2], deep neural network (DNN)-based VC [3, 4] is attracting attention these days. Thanks to the non-linear transformation of DNNs and the techniques shared among different research fields, deep architectures can now perform higher-quality sequence-wise conversion. However, since these systems use offline or time processes, (e.g., WORLD [5]-based feature analysis and bi-directional recurrent neural network [4]), they cannot be directly applied to online and low-latency conversion.

In this paper, we propose an implementation method of DNN-based VC that works online with low latency. The fundamental idea is inspired by GMM-based real-time VC [6], and we introduce several techniques for efficient conversion. Towards the use of VC in artificial delayed auditory feedback, we implement VC with a 50 ms latency that is not noticeable by the speaker. To improve the speech quality of the proposed real-time VC, we propose three methods of audio data augmentation that artificially augment training data. Two of these methods, pitch shift and time stretch [7], make the real-time VC robust to perturbations of human speech production, and the third one, time shift, makes it robust to perturbation of the start time of short-time Fourier transform. For the construction of a real-time VC device, we need to ensure that it is robust to background noise if we want it to have practical use in actual environments. One way to make real-time VC robust to noise is the use of audio data augmentation with the artificial injection of a variety of noise [8]. However, making it robust to non-stationary noise (e.g., another person's voice) requires sequential compensation and some delays, which is unsuitable for the proposed real-time voice conversion. Therefore, we have developed a mask-based real-time VC device that physically blocks noise fed into a microphone. Experimental results demonstrate that 1) the proposed real-time VC works with 0.50 of the real-time factor, 2) the proposed data augmentation improves speech quality, and 3) the proposed mask-based VC device is more robust to noise than a standard microphone-based VC device.

2. Implementation of DNN-based real-time voice conversion

In this section, we describe how to implement DNN-based real-time voice conversion that is executable with low latency. The framework consists of analysis, conversion, and synthesis steps, as shown in Fig. 1. The analysis step extracts speech features from the source speaker's speech waveform. The obtained features are then converted into the target speaker's features by using DNNs in the conversion step, and the synthesis step synthesizes the converted speech waveform. All the processes are performed recursively with low latency. The following sections explain the details of the three steps.

2.1. Analysis step

Speech features, mel-cepstral coefficients, power, and F_0 are first extracted from an input speech waveform. As in conventional DNN-based VC, the use of state-of-the-art spectral feature analyzers, such as STRAIGHT [9] and WORLD [5], is required for high-quality feature analysis. However, they come at a high computational cost and require time delays[1]. As an alternative, we use simple fast Fourier transform (FFT)-based mel-cepstral coefficient analysis [10], the same as [6]. We can compensate for the quality disadvantage of FFT-based analysis by using our proposed audio data augmentation, as we describe in **Section 2**. We refer to this feature as *FFT mel-cepstral coefficients* in the following sections. Note that state-of-the-art feature analyzers can be used for extracting the target speaker's features, as the process is not included at run-time.

Next, trajectory smoothing [11], which removes the modulation spectrum components [12] of high modulation frequencies, is applied to the extracted FFT mel-cepstral coefficient sequence. The higher modulation frequency components are neg-

[1]For example, WORLD [5] uses a window function with the length of $3/F_0$ ms. If $F_0 = 70$ Hz, the window length is approximately 43 ms.

10.21437/SSW.2019-17

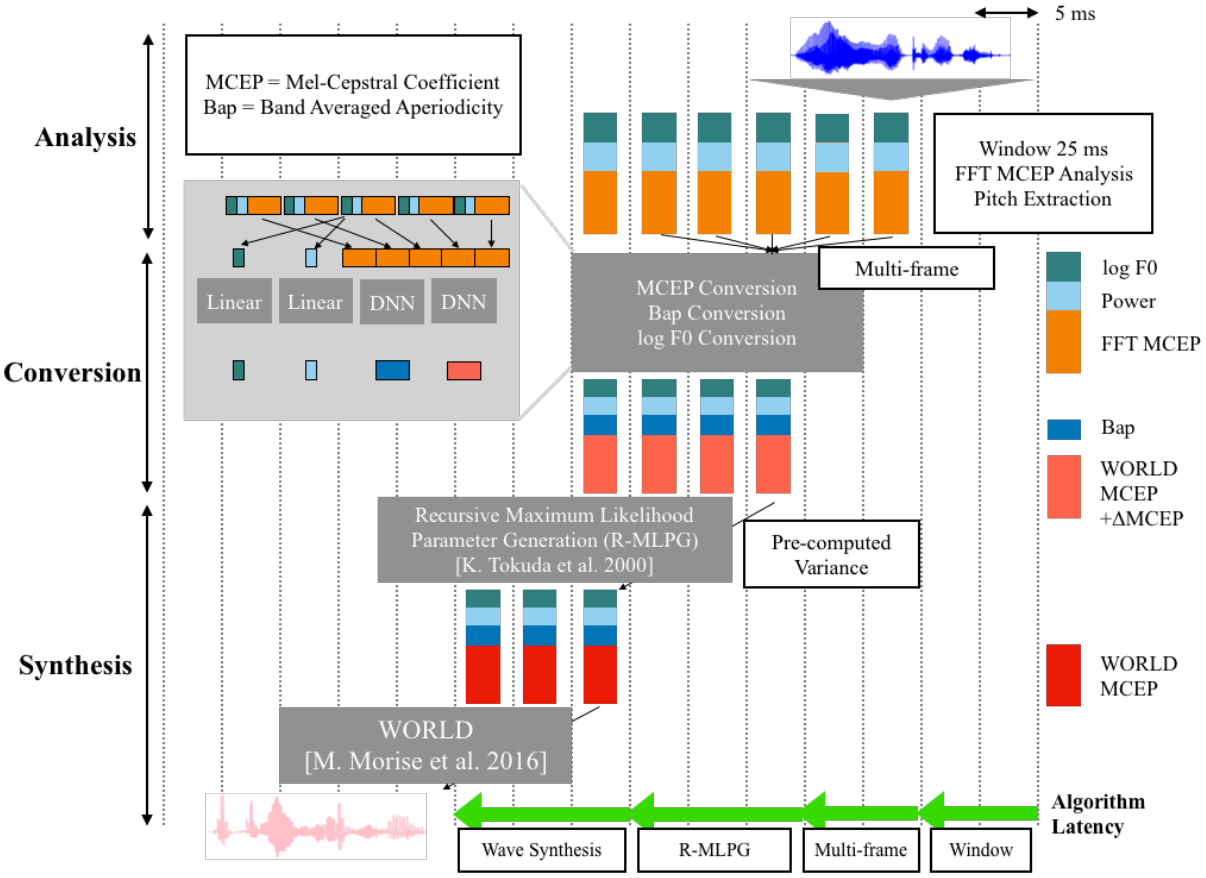

Figure 1: *Proposed real-time DNN-based VC.*

ligible for speech perception and difficult to be modeled with DNNs, and the trajectory smoothing can improve the prediction accuracy of speech features. In [11], a non-causal low pass filter (LPF) is used for the removal, but we use a two-tap finite impulse response (FIR) LPF for the purpose of online and low-delay conversion. The FIR filter has values at frames -1 and 0 (i.e., previous and current frame).

2.2. Conversion step

Speech features of the target speaker are estimated from speech features of the source speaker. We conduct the following four transformations.

1. From source speaker's FFT mel-cepstral coefficients to target speaker's WORLD mel-cepstral coefficients
2. From source speaker's FFT mel-cepstral coefficients to target speaker band-averaged aperiodicity
3. From source speaker's log-scaled F_0 to target speaker's log-scaled F_0
4. From source speaker's power to target speaker's power

The third and fourth transformations are linear, as used in the conventional GMM-based VC [2], and the first and second use DNNs. In the first transformation, we use multi-frame (current $\pm C$ frames) FFT mel-cepstral coefficients as input and one-frame WORLD mel-cepstral coefficients as output. Since the input and output are similar types of features, a DNN with input-to-output residual architecture [13] efficiently works for the feature conversion. Similarly, we use multi-frame FFT mel-cepstral coefficients as input and one-frame band-aperiodicity as output in the second transformation.

To train DNNs, we use the mean squared error (MSE) for the loss function. Given the target speaker's speech feature sequence $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top]^\top$ and the predicted feature sequence $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$, the loss function is

$$L(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{T}(\hat{\boldsymbol{y}} - \boldsymbol{y})^\top (\hat{\boldsymbol{y}} - \boldsymbol{y}), \quad (1)$$

where t represents the frame index and T is the total number of frames. Before the training, trajectory smoothing [11] is performed on the target speaker's WORLD mel-cepstral coefficients in order to improve prediction accuracy. Note that, theoretically, the proposed architecture runs regardless of the DNN's training method, which allows us to improve the model independently. For example, although we did not adopt in the experiment, generative adversarial network (GAN) [14]-based training [15] could be utilized for improving quality in converted speech without increasing computation costs at run-time.

2.3. Synthesis step

A waveform is synthesized from the converted speech features. As in the conventional method [6], recursive maximum likelihood parameter generation (R-MLPG) [16] is applied to the converted WORLD mel-cepstral coefficients. The covariance matrix used for the R-MLPG is calculated in advance [17]. A speech waveform is generated using the WORLD's recursive

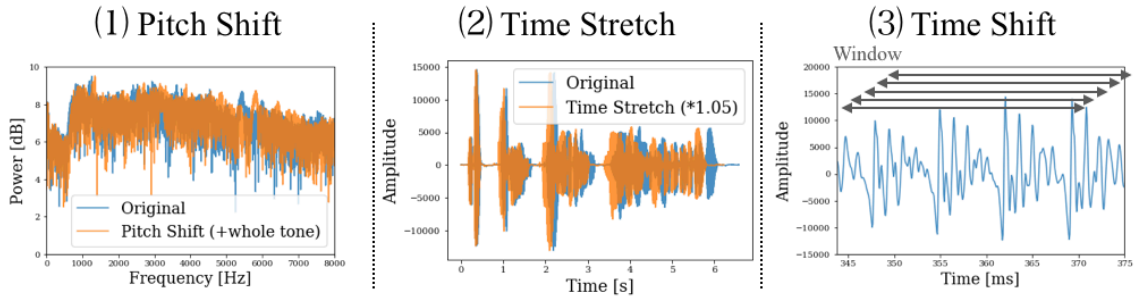

Figure 2: *Visualization of audio data augmentation.*

waveform generation algorithm [5]. Hereinafter, let M be the number of buffer frames in the R-MLPG algorithm and W be that in the WORLD's generation algorithm.

2.4. Algorithm latency

Here, we calculate the algorithm latency of the proposed DNN-based real-time VC. Supposing that the frame shift is 5 ms, frame length is 25 ms, $C = 2$, $M = 3$, $W = 3$ ($W = 3$ covers the length of a one-pitch waveform of 70 Hz of F_0.), and the algorithm latency of the analysis, conversion, and synthesis steps are 10 ms, 10 ms, and 30 ms, respectively. The total latency is 50 ms (see Fig. 1).

3. Audio data augmentation for DNN-based real-time VC

Data augmentation is a way to improve the performance of DNNs. It works by padding artificial data and utilizing the fact that a DNN's performance changes sensitively according to the amount of training data [18]. It is widely used for speech recognition [7] and acoustic event classification [19, 20, 21]. In this paper we apply two methods (pitch shift and time stretch) [7] to make the proposed VC framework robust to perturbation of human speech production. Furthermore, we propose another method (time shift) to tackle a drawback of the FFT-based mel-cepstral analysis we used in **Section 2**. Figure 2 shows the original waveform and the waveforms processed with each data augmentation method.

3.1. Pitch Shift (P-Sh)

Even when a single speaker speaks the same text, the pitch trajectory will be different each time. For training data augmentation, we use the WSOLA algorithm [22] and waveform resampling [23] to slightly change the pitch of the source and target speaker's waveforms. The left of Fig. 2 shows the original power spectrum (blue) and the one shifted a whole tone.

3.2. Time Stretch (T-St)

As discussed in Section 3.1, we introduce data augmentation by time stretch to ensure robustness to speech speed perturbation. The WSOLA algorithm [22] is used for the stretch. The middle image in Fig. 2 shows a case in which a waveform (red) is sped up by 1.05 and thus is shifted to the right of the original one (blue).

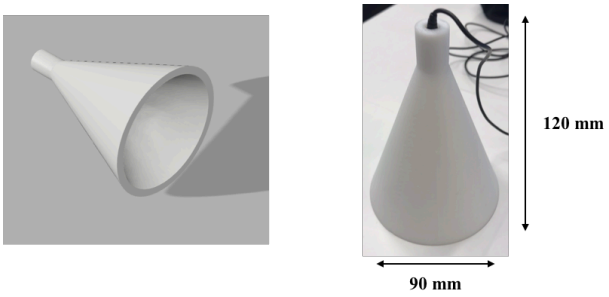

Figure 3: *The proposed device. A small microphone is included in the mask and connected to a laptop PC.*

3.3. Time Shift (T-Sh)

The start time of the frame analysis influences the power spectra (and also mel-cepstral coefficients). Feature analysis in VC must be robust to the start time. For example, WORLD [5] designs a F_0-adaptive window function such that the resulting spectrum is theoretically independent of the start time of analysis, but this framework is unsuitable for real-time VC, as explained in **Section 2**. Instead, we propose a data augmentation method where the start time of FFT-based mel-cepstral analysis is shifted within the frame shift length. For example, setting the frame shift length to 5 ms, we set an offset of the start time with a range of $\pm$ 2.5 ms.

4. Noise-robust DNN-based real-time VC with mask-shaped device

To achieve a real-time VC system for augmented human communication, the system should be usable in noisy environments. There are two possible approaches to this: adding noise suppression before VC or training acoustic models using noisy speech. While the former approach (e.g., monaural speech enhancement [24, 25]) performs noise suppression with low latency, the noise suppression works only in limited cases. Also, while the latter approach (e.g., data augmentation using the artificial injection of a variety of noise [8]) performs a highly robust noise suppression, it requires heavy computation due to deeper architectures for sequential compensation (e.g., suppressing another person's voice).

Therefore, we have developed a real-time VC device that is robust against noise by physically preventing noise contamination. Figure 3 shows the proposed device. A monaural microphone is embedded at the top of the device and connected to a laptop PC (for real-time VC) via a cable. The VC user

Table 1: *Real-time factor (RTF) of our DNN-based real-time VC (mean and standard deviation). The processing time indicates a computation time within one frame (5 ms).*

	Analysis	Conversion	Synthesis
Processing time [ms]	0.85 ± 0.092	1.12 ± 0.51	0.57 ± 0.10
RTF	0.17	0.22	0.11

fits the device over his/her mouth to prevent background noise from contaminating the input speech, thus keeping the conversion quality high. Holding the device very tightly would impede free movement of articulators, so the user holds the device such that a tiny gap is present between the device and his/her face.

5. Experimental evaluation

5.1. Experiment settings

We used 100 utterances (approx. 12 min.) of two Japanese male speakers, the original transcript of which is included in the Voice Actress Corpus [26]. The sampling frequency was 16 kHz. We used 90 % of the corpus for DNN training and the remainder for evaluation. The frame shift was 5 ms . The lengths of the frame and the FFT were 400 (25 ms) and 512, respectively. The number of dimensions of FFT/WORLD mel-cepstral coefficients was 40 (0th-through-39th). The 0th component was used for the power component. C, M, W in **Section 2** were set to 2, 3, 3, respectively. The DNN for FFT mel-cepstral coefficients to WORLD mel-cepstral coefficients was multi-layer perceptron consisting of a 195 (39×5)-unit input layer, 2×500-unit leaky ReLU hidden layers, and a 78 (39×2, static & delta features)-unit linear output layer. The DNN also includes an input-to-output residual network [13] that connects FFT mel-cepstral coefficients at the current frame of the input layer and static features at the output layer. The DNN for FFT mel-cepstral coefficients to band-aperiodicity is a single-layer perceptron consisting of a 195-unit input layer and a 1-unit sigmoid output layer. The FFT/WORLD mel-cepstral coefficients were normalized to have zero-mean and unit-variance. Adam [27] was used as the optimization method. An Intel (R) Core (TM) i7-3770K CPU @ 3.50 GHz was used in the evaluation of processing delay, with the aim of showing the capability of this system in a CPU environment. For the augmentation of the pitch shift, a total of four patterns of {semitone, whole tone} {up, down} were applied. As for time stretch, the speech speed was multiplied by $\{0.95, 1.05\}$. The offsets of time shift were set to $-2.50, -1.25, 0.0, 1.25$, and 2.50 ms.

5.2. Computation cost of proposed real-time VC

We calculated a real-time factor (RTF) to determine whether the proposed VC operates in real time on a single CPU. The RTF is a value of the computation time divided by the length of the input waveform. Table 1 summarizes RTFs of the analysis, conversion, and synthesis steps. While the conversion step using DNNs took more computation time than the others, the total RTF was smaller than 1.0. Therefore, we can confirm that our VC system operates in real time (and with 50 ms of algorithm latency).

5.3. Effectiveness of proposed data augmentation

We evaluated the effectiveness of the three proposed data augmentation methods in terms of prediction accuracy (**Section 5.3.1**) and speech quality (**Section 5.3.2**).

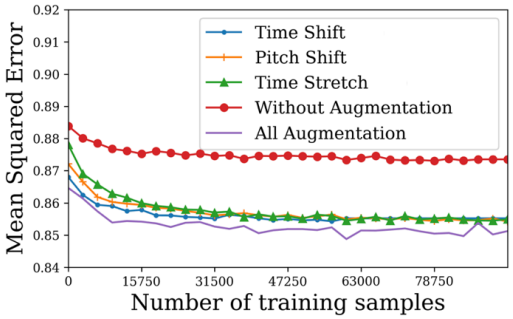

Figure 4: *MSEs at each training step.*

Table 2: *MSEs of trained DNNs.*

Without Augmentation	Pitch Shift	Time Stretch	Time Shift	All Augmentation
0.878	0.859	0.858	0.859	**0.853**

5.3.1. Prediction accuracy

We calculated MSE between the predicted and target WORLD mel-cepstral coefficients of the evaluation data. The compared methods are 1) without augmentation (conventional method), 2) pitch shift, 3) time stretch, 4) time shift, and 5) using all of 2-through-4. Figure 4 shows the MSE at each step of DNN training. The amount of training data was increased by factors of one, five, three, five, and 35, respectively. The x-axis is the number of frames used for iterative training. Table 2 summarizes the MSE values after completing the DNN training. Note that the WORLD mel-cepstral coefficients were normalized to have zero-mean and unit-variance. As indicated in the figure and table, there was no big difference among the different data augmentation methods, but the data augmentation dramatically improved the conversion performance.

5.3.2. Speech quality

We conducted a preference AB test to evaluate the naturalness of the converted speech. The compared methods are real-time VC without and with data augmentation (using all of 2-through-4 in **Section 5.3.1**). We presented a pair of converted speech in random order and had listeners select the speech sample that sounded most natural. Similarly, a preference XAB test on the speaker individuality was conducted using the natural speech as a reference "X." These tests were done using our crowdsourcing evaluation system. Thirty-five listeners participated in each evaluation and ten utterances were used per listener.

The results of the subjective evaluation are shown in Fig. 5. As we can see, both speech quality and speaker individuality were significantly improved by the proposed data augmentation.

5.4. Robustness against noise

Finally, we evaluated the noise robustness of the DNN-based real-time VC with the proposed mask-shaped device, using voices recorded in actual noisy environments. Compared methods are VC without (i.e., monaural microphone only) and with the mask-shaped device. We recorded evaluation data in two environments: an anechoic room (clean environment) and a train station (noisy environment). The recording in the station was

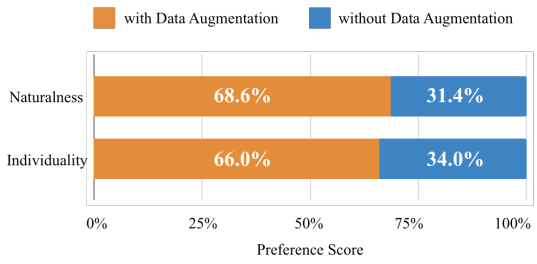

Figure 5: *Results of preference AB test on naturalness and preference XAB test on speaker individuality (evaluation of audio data augmentation). Their p-values were smaller than* 0.001.

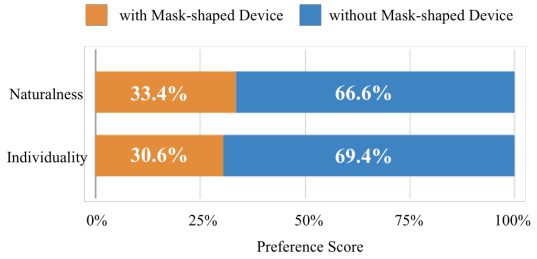

Figure 6: *Results of preference AB test on naturalness and preference XAB test on speaker individuality (evaluation of mask-shaped device in clean environment). Their p-values were smaller than* 0.001.

done at Shibuya station in Tokyo, Japan, which is an extremely busy location where crowds of people mill and chatter.

The preference AB test on speech quality and preference XAB test on speaker individuality were conducted in the same manner as in **Section 5.3.2**. Results for the clean and noisy environments are summarized in Figs. 6 and 7, respectively. Unfortunately, our device degraded speech quality and speaker individuality for the clean-environment recording. This is because the shape of our device slightly impeded the movements of articulators. In addition, since the mask-shaped device was used only for conversion and not for building the conversion models, the mismatch between the types of speech in training and inference could have a negative influence on the quality. On the other hand, the device improved the naturalness for noisy-environment recording due to the physical prevention of noise contamination.

6. Conclusion

We proposed techniques and devices for real-time voice conversion (VC). We first showed how to implement deep neural network (DNN)-based real-time VC, which operates in 0.50 of the real-time factor and with 50 ms of algorithm latency. Second, we proposed three data augmentation methods to achieve robustness against perturbation of human speech production and Fourier transform. Finally, we developed a DNN-based real-time VC system with a mask-shaped device for use in noisy environments. Our future work will focus on further evaluation, including a comparison of quality with conventional DNN-based VC and noise suppression, an evaluation with more speaker pairs, and MOS test. Furthermore, to ameliorate the quality when the mask is used, applying a filter and developing the model using the speech recorded with the mask are desired

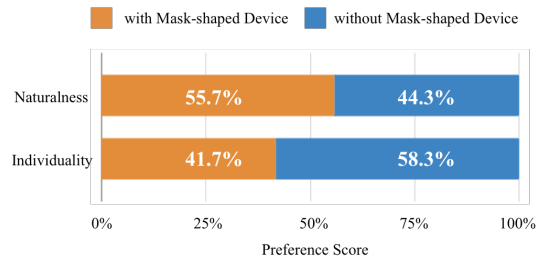

Figure 7: *Results of preference AB test on naturalness and preference XAB test on speaker individuality (evaluation of mask-shaped device in noisy environment). Their p-values were smaller than* 0.01 *and* 0.001*, respectively.*

to be conducted.

Acknowledgements: Part of this research and development work was supported by the Ministry of Internal Affairs and Communications. Also, part of this research was executed in the Social Cooperation Programs "Applications of Spatio-temporal Analysis" with DMM.com LLC.

7. References

[1] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[2] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[3] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order Eigen space using deep belief nets," in *Proc. INTERSPEECH*, Lyon, France, Aug. 2013, pp. 369–372.

[4] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4869–4873.

[5] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, Jul. 2016.

[6] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion," in *Proc. INTERSPEECH*, Portland, U.S.A., Sep. 2012.

[7] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, 2015, pp. 3586–3589.

[8] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep Speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[9] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[10] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, San Francisco, U. S. A., Mar 1992, pp. 137–140.

[11] S. Takamichi, K. Kobayashi, K. Tanaka, T. Toda, and S. Nakamura, "The NAIST text-to-speech system for the Blizzard Challenge 2015," in *Proc. Blizzard Challenge workshop*, Berlin, Germany, Sep. 2015.

[12] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 755–767, Apr. 2016.

[13] Y. Saito, S. Takamichi, and H. Saruwatari, "Voice conversion using input-to-output highway networks," *IEICE Transactions on Information and Systems*, vol. E100.D, no. 8, pp. 1925–1928, 2017.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, Montreal, Canada, Dec. 2014, pp. 2672–2680.

[15] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 755–767, Jun. 2018.

[16] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory," in *Proc. INTERSPEECH*, Brisbane, Australia, Sep. 2008, pp. 1076–1079.

[17] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, Vancouver, Canada, May 2013, pp. 7962–7966.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[19] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Letter*, vol. 24, no. 3, pp. 279–283, 2017.

[20] B. McFee, E. J. Humphrey, and J. P. Bello, "A software framework for musical data augmentation," in *Proc. of ISMIR*, 2015, pp. 248–254.

[21] N. Takahashi, M. Gygli, B. Pfister, and L. V. Gool, "Deep convolutional neural networks and data augmentation for acoustic event recognition," in *Proc. INTERSPEECH*, San Fransisco, U.S.A., 2016, pp. 2982–2986.

[22] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech," in *Proc. ICASSP*, vol. 2, April 1993, pp. 554–557 vol.2.

[23] K. Kobayashi, S. Takamichi, S. Nakamura, and T. Toda, "The NU-NAIST voice conversion system for the Voice Conversion Challenge 2016," in *Proc. INTERSPEECH*, San Francisco, U.S.A., Sep. 2016, pp. 1667–1671.

[24] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation," *IEICE Transactions on Information and Systems*, vol. E97-D, no. 6, pp. 1429–1437, Jun. 2014.

[25] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[26] y_benjo and MagnesiumRibbon, "Voice-actress corpus," http://voice-statistics.github.io/.

[27] D. Kingma and B. Jimmy, "Adam: A method for stochastic optimization," in *arXiv preprint arXiv:1412.6980*, 2014.