LETTER

Voice Conversion Using Input-to-Output Highway Networks

Yuki SAITO[†a)], **Shinnosuke TAKAMICHI**[†b)], ***Nonmembers***, ***and*** **Hiroshi SARUWATARI**[†c)], ***Member***

SUMMARY This paper proposes Deep Neural Network (DNN)-based Voice Conversion (VC) using input-to-output highway networks. VC is a speech synthesis technique that converts input features into output speech parameters, and DNN-based acoustic models for VC are used to estimate the output speech parameters from the input speech parameters. Given that the input and output are often in the same domain (e.g., cepstrum) in VC, this paper proposes a VC using highway networks connected from the input to output. The acoustic models predict the weighted spectral differentials between the input and output spectral parameters. The architecture not only alleviates over-smoothing effects that degrade speech quality, but also effectively represents the characteristics of spectral parameters. The experimental results demonstrate that the proposed architecture outperforms Feed-Forward neural networks in terms of the speech quality and speaker individuality of the converted speech.

key words: *statistical parametric speech synthesis, DNN-based voice conversion, highway networks, over-smoothing*

1. Introduction

Statistical Voice Conversion (VC) is an effective technique that converts input speech into the desired output speech while keeping its linguistic information unchanged. Deep Neural Networks (DNNs) [1] have been used as acoustic models for VC because they can represent the relationship between the input and output speech parameters more accurately than conventional Gaussian mixture models [2]. These acoustic models are trained with training algorithms such as the maximum likelihood criterion [3] and Minimum Generation Error (MGE) criterion [4], [5]. However, the converted speech parameters tend to be over-smoothed, and this phenomenon degrades the quality of the converted speech. The over-smoothing effect is an issue in not only VC but also other speech synthesis techniques, such as text-to-speech synthesis. Hence, several approaches have been devised to reproduce the characteristics of natural speech [2], [6], [7]. On the other hand, VC can utilize not only those approaches, but also input speech information since the input and output parameters are often in the same domain (e.g., cepstrum).

This paper proposes DNN-based VC using input-to-output highway networks. Although the typical DNN-based VC directly estimates a converted spectral parameter sequence, our architecture estimates it as the sum of input spectral parameters and weighted spectral differentials estimated through DNNs. The use of input speech parameters effectively alleviates the over-smoothing effect, and the weights of the spectral differentials effectively represent the characteristics of the spectral parameters. Experimental results demonstrate that the proposed architecture yields significant improvements in terms of converted speech quality and speaker individuality.

Manuscript received February 15, 2017.
Manuscript revised April 20, 2017.
Manuscript publicized April 28, 2017.

[†]The authors are with the Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, 113–8656 Japan.
a) E-mail: yuuki_saito@ipc.i.u-tokyo.ac.jp
b) E-mail: shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
c) E-mail: hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp
DOI: 10.1587/transinf.2017EDL8034

2. Conventional DNN-Based VC

The acoustic models are trained using MGE training algorithm [5]. Let $\boldsymbol{y}$ be a natural speech parameter sequence $[\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_T^\top]^\top$, and $\hat{\boldsymbol{y}}$ be a converted speech parameter sequence $[\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$, where T denotes the total frame length. The loss function $L(\boldsymbol{y}, \hat{\boldsymbol{y}})$ is the mean squared error between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$, as follows:

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

After applying a weighting matrix $\boldsymbol{W}$ [3] to an input speech parameter sequence $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_T^\top]^\top$ for calculating its static-dynamic speech feature sequence, the DNNs predict a static-dynamic speech feature sequence of the converted speech. $\hat{\boldsymbol{y}}$ is generated from the static-dynamic features by using the maximum likelihood-based parameter generation algorithm [2]. We define the above speech parameter conversion as $\hat{\boldsymbol{y}} = \boldsymbol{G}(\boldsymbol{x})$. Figure 1 illustrates typical voice conversion with the DNN-based acoustic models.

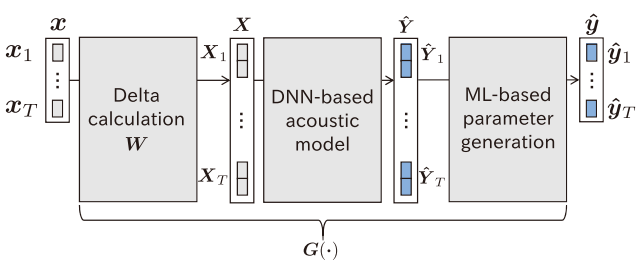

Fig. 1 Typical voice conversion based on speech parameter conversion. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are static-delta input and output speech features, respectively.

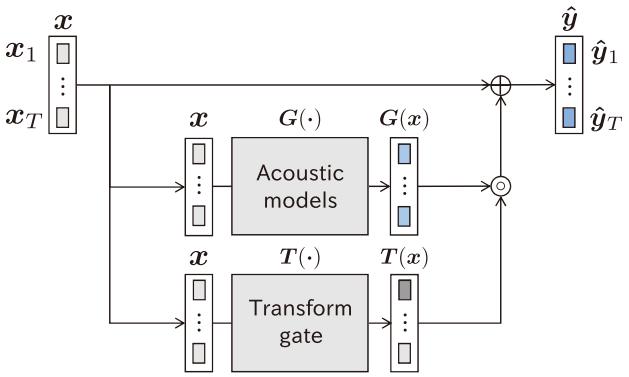

Fig. 2 Voice conversion using input-to-output highway networks.

3. Proposed Architecture

3.1 VC Using Input-to-Output Highway Networks

Highway networks [8], [9] are weighted skip-connections between layers, and they often connect hidden layers. Given that the input and output are often in the same domain (e.g., cepstrum) in VC, we propose a VC using highway networks connected from the input to output as follows:

$$\hat{\boldsymbol{y}} = \boldsymbol{x} + \boldsymbol{T}(\boldsymbol{x}) \circ \boldsymbol{G}(\boldsymbol{x}) \tag{2}$$

where $\circ$ is the Hadamard product. $\boldsymbol{T}(\cdot)$ is the transform gate of highway networks described as Feed-Forward neural networks. Each value of $\boldsymbol{T}(\boldsymbol{x})$ ranges from 0.0 to 1.0, and they represent time- and feature-varying weights of $\boldsymbol{G}(\boldsymbol{x})$. When $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{0}$, input speech parameters are directly used as converted speech parameters, and when $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{1}$, the architecture is equivalent to residual networks [10]. Therefore, the input speech parameters are strongly transformed by $\boldsymbol{G}(\cdot)$ when the value of the transform gate becomes close to 1.0. Figure 2 shows the proposed architecture. The loss function for training is equal to Eq. (1), and all model parameters of $\boldsymbol{T}(\cdot)$ and $\boldsymbol{G}(\cdot)$ are simultaneously estimated to minimize the loss function.

3.2 Discussion

Since our architecture utilizes both input speech parameters and spectral differentials weighted by the transform gate, it efficiently alleviates over-smoothing of the converted speech parameters. Figure 3 shows scatter plots of the speech parameters. This figure plots pairs of mel-cepstral coefficients whose corresponding value of the transform gate is large (i.e., it is close to residual networks) or small (i.e., it is close to direct use of input speech parameters). We can see that our architecture alleviates distribution shrinkage better than Feed-forward neural networks in both cases.

The variation in spectral parameters between speakers strongly depends on not only the speaker pair but also the frequency band and phonetic environments. For instance, formant structures change more in the inter-gender case than

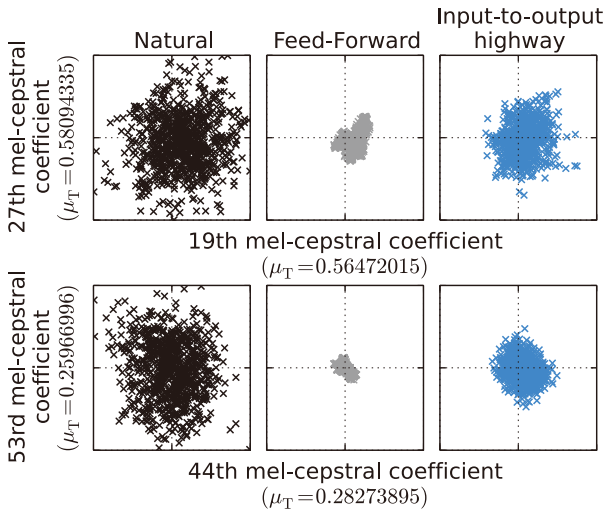

Fig. 3 Scatter plots of speech parameters. μ_T denotes the value of the transform gate averaged over one utterance.

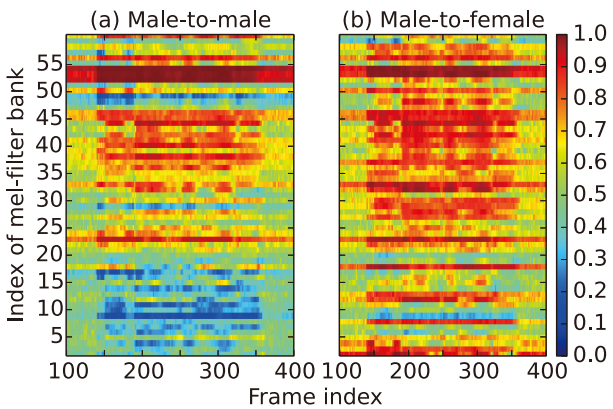

Fig. 4 Examples of activation of transform gates using mel-filter banks.

in the intra-gender case, but the inter-speaker variation is small in the lower frequency bands within the same gender. On the other hand, inter-phoneme variation (intra-speaker variation) is large in the lower frequency bands [11]. Therefore, *golden* VC should avoid over-transformation (e.g., frequency warping [12]) when the input feature is close to the output feature and should apply a flexible transformation when the input is far from the output feature. The transform gates in Fig. 2 can be interpreted as the variation and characteristics of the spectral parameters. Figure 4 shows examples of activation of the transform gates using mel-filter banks. This figure shows that $\boldsymbol{G}(\cdot)$ greatly transforms the spectral parameters in the high frequency band, since they strongly represent the characteristics of the speaker. Meanwhile, in male-to-male speaker conversion (Fig. 4 (a)), $\boldsymbol{G}(\cdot)$ does not transform the spectral parameters in the low frequency band as much as in the case of male-to-female speaker conversion (Fig. 4 (b)).

From another perspective, our architecture can be regarded as *soft* selection of features. The dimensionalities of the speech features (e.g., the numbers of mel-cepstral coef-

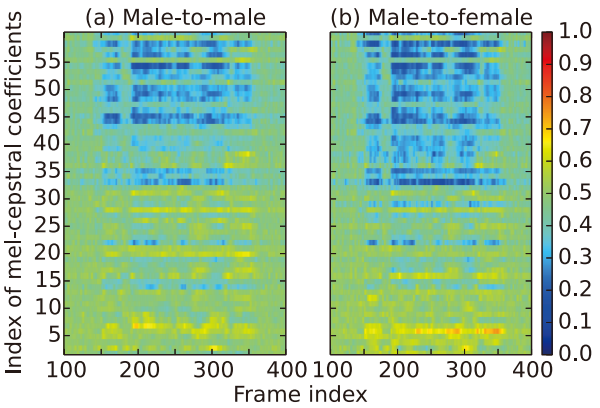

Fig. 5 Examples of activation of transform gates using mel-cepstral coefficients.

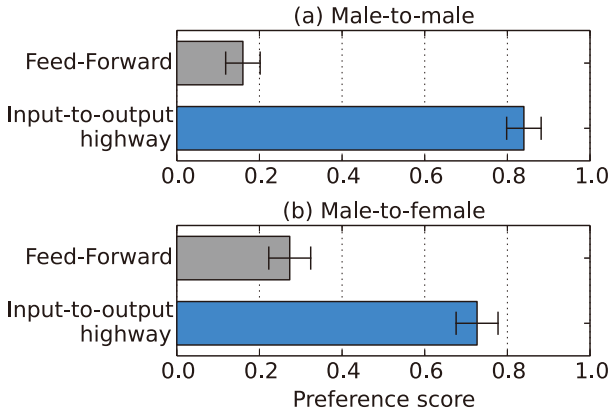

Fig. 6 Preference scores of speech quality of converted speech with 95% confidence intervals.

ficients) are hyper-parameters for VC. For instance, the use of only the lower order of the mel-cepstrum makes the training robust while it degrades speech quality. On the other hand, the use of the rich orders improves speech quality but suffers from the randomness of the higher order of the mel-cepstrum. The former case corresponds to $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{1}$ for the lower order and $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{0}$ for the higher order. The latter case corresponds to $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{1}$ for all orders. Whereas such a *hard* selection is often used, our architecture can utilize *soft* selection; i.e., each activation of $\boldsymbol{T}(\boldsymbol{x})$ varies from 0.0 to 1.0 depending on $\boldsymbol{x}$. Figure 5 shows examples of activation of the transform gates using mel-cepstral coefficients. We can see that the lower orders of the mel-cepstral coefficients, which are dominant in speaker conversion, tend to be strongly transformed by $\boldsymbol{G}(\cdot)$. On the other hand, the higher orders of the mel-cepstral coefficients tend to be not completely ignored, but weakly converted.

Finally, the transform gate of our architecture is similar to adaptive soft-masking filtering [13] in speech enhancement. Hence, it is expected that knowledge can be shared between voice conversion and speech enhancement.

4. Experimental Evaluation

4.1 Experimental Conditions

We used speech data of two male speakers and one female speaker taken from the ATR Japanese speech database [14]. The speakers uttered 503 phonetically balanced sentences. We used 450 sentences (subsets A to I) for training and 53 sentences (subset J) for evaluation. Speech signals were sampled at a rate of 16 kHz, and the shift length was set to 5 ms. The 0th-through-59th mel-cepstral coefficients were used as the spectral parameter and F_0 and 5 band-aperiodicity [15], [16] were used as excitation parameters. The STRAIGHT analysis-synthesis system [17] was used for the parameter extraction and waveform synthesis. The 0th mel-cepstral coefficients of the input speech were directly used as those of the converted speech. To improve training accuracy, speech parameter trajectory smoothing [18] with a 50 Hz cutoff modulation frequency was applied to the spectral parameters in the training data. In the training phase, the spectral features were normalized to have zero-mean unit-variance, and the MGE training [5] was performed. We built DNNs for male-to-male and male-to-female conversion. The DNN architectures were Feed-Forward networks. The architecture included 3×512-unit Rectified Linear Unit (ReLU) [19] hidden layers and a 118-unit linear output layer. The acoustic models output static and dynamic mel-cepstral coefficients (118-dim.) frame by frame. The transform gate only had a 59-unit input and 59-unit sigmoid output layers. We used AdaGrad [20] as the optimization algorithm, setting the learning rate to 0.01. F_0 was linearly transformed, and band-aperiodicity was not transformed.

To evaluate our architecture, we conducted a subjective evaluation of the converted speech quality and speaker individuality.

4.2 Subjective Evaluation

In the subjective evaluation, we compared the proposed architecture (input-to-output highway) with the conventional one (Feed-Forward). A preference test (AB test) was conducted to evaluate the speech quality. We presented every pair of converted speech of the two architectures in random order, and we forced listeners to select speech samples that sounded like they had better quality. Similarly, an XAB test on the speaker individuality was conducted using natural speech as the reference, i.e., "X". Thirty listeners participated in each assessment of our crowd-sourced evaluation systems.

The results of the preference tests on speech quality and speaker individuality are shown in Fig. 6 and Fig. 7, respectively. We found that our architecture scored higher in both speech quality and speaker individuality than the conventional Feed-Forward neural network-based VC. Therefore, we demonstrated the effectiveness of our architecture.

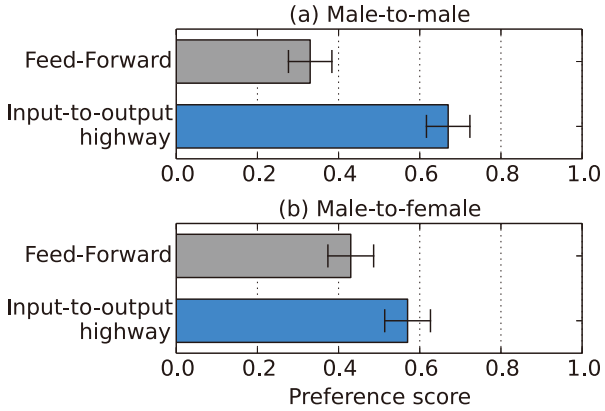

Fig. 7 Preference scores of speaker individuality of converted speech with 95% confidence intervals.

5. Conclusion

This paper proposed a deep neural network-based voice conversion using input-to-output highway networks. Since the input speech parameters are directly used through the input-to-output highway networks for estimating the converted speech parameters, the proposed architecture effectively alleviates the over-smoothing effect that degrades speech quality. Moreover, our architecture can represent the characteristics of spectral parameters with the transform gates of the highway networks. The experimental results showed significant improvements in terms of speech quality and speaker individuality. In the future, we will investigate the phonetic environment in the proposed architecture.

Acknowledgements

Part of this work was supported by ImPACT Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan), SECOM Science and Technology Foundation, and JSPS KAKENHI Grant Number 16H06681.

References

[1] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," IEEE Signal Process. Mag., vol.32, no.3, pp.35–52, May 2015.

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

[3] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," Proc. IEEE, vol.101, no.5, pp.1234–1252, April 2013.

[4] Y.-J. Wu and R.-H. Wang, "Minimum generation error training for HMM-based speech synthesis," Proc. ICASSP, Toulouse, France, pp.89–92, May 2006.

[5] Z. Wu and S. King, "Improving trajectory modeling for DNN-based speech synthesis by using stacked bottleneck features and minimum trajectory error training," IEEE Trans. Audio Speech Lang. Process., vol.24, no.7, pp.1255–1265, July 2016.

[6] 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 Trans. Audio Speech Lang. Process., vol.24, no.4, pp.755–767, April 2016.

[7] Y. Saito, S. Takamichi, and H. Saruwatari, "Training algorithm to deceive anti-spoofing verification for DNN-based speech synthesis," Proc. ICASSP, New Orleans, U.S.A., pp.4900–4904, March 2017.

[8] R.K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," Proc. ICML Deep Learning Workshop, Lille, France, July 2015.

[9] X. Wang, S. Takaki, and J. Yamagishi, "Investigating very deep highway networks for parametric speech synthesis," Proc. 9th ISCA Speech Synthesis Workshop, California, U.S.A., pp.166–171, Sept. 2016.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. CVPR, Las Vegas, U.S.A., pp.770–778, June 2016.

[11] T. Kitamura and M. Akagi, "Speaker individualities in speech spectral envelopes," J. Acoust. Soc. Jpn (E), vol.16, no.5, pp.283–289, Sept. 1995.

[12] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," IEEE Trans. Audio Speech Lang. Process., vol.18, no.5, pp.922–931, July 2010.

[13] J. van Hout and A. Alwan, "A novel approach to soft-mask estmation and log-spectral enhancement for robust speech recognition," Proc. ICASSP, Kyoto, Japan, pp.4105–4108, March 2012.

[14] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda, and H. Kuawhara, "A large-scale Japanese speech database," ICSLP90, pp.1089–1092, Kobe, Japan, Nov. 1990.

[15] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," MAVEBA 2001, Firentze, Italy, pp.1–6, Sept. 2001.

[16] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," Proc. INTERSPEECH, Pittsburgh, U.S.A., pp.2266–2269, Sept. 2006.

[17] H. Kawahara, I. Masuda-Katsuse, and A.D. Cheveigné, "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, April 1999.

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

[19] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," Proc. AISTATS, Lauderdale, U.S.A., pp.315–323, April 2011.

[20] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," Journal of Machine Learning Research, vol.12, pp.2121–2159, July 2011.