2018 IEEE International Symposium on Multimedia (ISM)

Spectrum Enhancement of Singing Voice Using Deep Learning

Ryuka Nanzaka
Graduate School of System Informatics
Kobe University
Kobe, Japan
ryuka.nanzaka@stu.kobe-u.ac.jp

Tsuyoshi Kitamura
Graduate School of System Informatics
Kobe University
Kobe, Japan
kitamura@stu.kobe-u.ac.jp

Yuji Adachi
MEC COMPANY LTD.
Amagasaki, Japan

Kiyoto Tai
MEC COMPANY LTD.
Amagasaki, Japan

Tetsuya Takiguchi
Graduate School of System Informatics
Kobe University
Kobe, Japan
takigu@kobe-u.ac.jp

***Abstract*—In this paper, we propose a novel singing-voice enhancement system that makes the singing voice of amateurs similar to that of professional opera singers, where the singing voice of amateurs is emphasized by using a singing voice of a professional opera singer on a frequency band that represents the remarkable characteristic of the professional singer. Moreover, our proposed singing-voice enhancement based on highway networks is able to convert any song (that a professional opera singer does not sing). As a result of our experiments, the singing voice of the amateur singer at the middle-high frequency range which contains a lot of frequency components that affect glossiness was emphasized while maintaining speaker characteristics.**

***Index Terms*—deep learning, speech enhancement, voice conversion, highway networks.**

I. Introduction

In this paper, we propose a novel singing-voice enhancement system that converts the frequency band of a singing voice of a person to that of another person. Many people usually sing for entertainment, such as when they sing karaoke and so on. Also, in recent years, the number of opportunities for singing has been increasing as people are selling personally-made CDs and posting their music videos on the Internet, regardless of whether they are professional or amateur singers. A level of singing proficiency may be classified according to differences in vocalization, correctness of pitch and tempo, and strength and emotion in a song.

In this paper, to make the singing voice of amateurs similar to that of professional opera singers, the singing voice of amateurs is emphasized by using a singing voice of a professional opera singer on a frequency band that represents the remarkable characteristic of the professional singer. Unlike playing music with a band, when a CD is created or a music video is posted on the Internet, the sound recording is performed for every part individually, such as vocals and musical instruments. After correction and processing for each recorded signal, one music waveform is obtained by synchronizing all the individual tracks. When processing vocal sounds, a commercial software or other tools to modify the fundamental frequency or amplitude may be used.

However, since these processes need to be carried out manually by the user, much time and labor are required. Therefore, much research related to how to automate the correction and processing of singing voices has been carried out. S. Bock [1] *et al.* proposed a method to remove vibrato by using smoothing of the spectrum envelope. S. Yong [2] *et al.* proposed a method to correct the tempo, pitch and amplitude of singing voices using professional singing voices. This is carried out by taking alignment using DTW (Dynamic Time Warping) in order to match the singing voice length of the target, and then PSOLA (Pitch Synchronous Overlap and Add) is employed to correct the pitch, and so on.

In this paper, we aim to convert the singing voice of an amateur to the glossier and well-projected voice of a professional by giving the frequency band of the professional opera singer that contributes its glossiness to the voice of the amateur. Doing this makes it possible to use the converted voice for voice training, or to convert a singing voice recorded using conversion software into the singing voice of a person with a higher level of proficiency. Moreover, our proposed singing-voice enhancement based on neural networks is able to convert any song (that a professional opera singer does not sing).

II. Singing voices of an amateur singer and opera singer

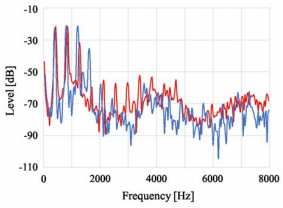
The singing voices of a female professional opera singer and a female amateur singer, who has not learned opera, are used in this paper. Fig. 1 shows the spectrum of the phonemes of “a”, “u”, and “i” that the opera singer and the amateur singer voiced while singing.

Fig. 1 (a) shows the spectrum of the “a” phoneme sung by two amateur singers during singing. Fig. 1 (b) shows the phoneme of “a” sung by three opera singers. Figs. 1 (c), (d), and (e) show the average spectra of two amateur singers and

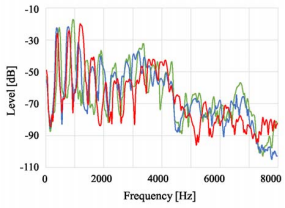
978-1-5386-6857-3/18/$31.00 ©2018 IEEE
DOI 10.1109/ISM.2018.00-18

the average spectrum of three opera singers for phonemes of "a", "u" and "i". When analyzing the spectra, a section of 0.5 seconds is extracted from the singing voices for the same part of the musical score.

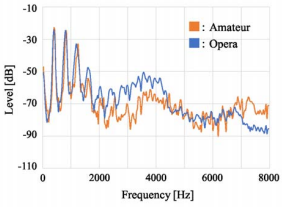
As shown in Fig. 1, there is no difference between the voice of the opera singers and the amateur singers around the low-frequency band. At the middle-high frequency range from 3,000 Hz to 4,000 Hz band, the professional voice is stronger compared to the voice of the amateur, and it indicates that the middle-high frequency range includes many glossy components of the voice (of the professional opera singers).



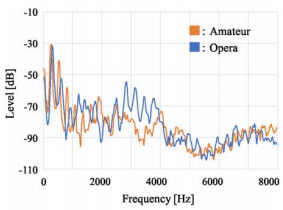
(a) "a" of two amateur singers.



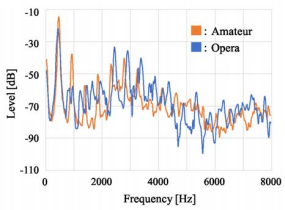
(b) "a" of three professional opera singers.



(c) "a" of amateur and opera singers.



(d) "u" of amateur and opera singers.



(e) "i" of amateur and opera singers.

Fig. 1. Sample spectra of amateur and opera singers.

III. Enhancement of Singing Voice Using Deep Learning

As mentioned above, in order to convert the amateur voice into the glossy and well-projected voice of an opera singer, we considered adding or emphasizing the energy component of the middle-high frequency range from 3,000 Hz to 4,000 Hz of the voice of the opera singer into the voice of the amateur. In order to convert any song that is not included in the training data and that the opera singer does not sing, a framework of neural networks is proposed in this paper.

In this work, the speech features of the amateur singer are used as the input of neural networks, and those of the opera singer are used as supervised data. In order to estimate the difference between the bands from 3,000 Hz to 4,000 Hz, highway networks [3] is used, which have a structure that enables us to calculate the difference in spectra. Fig. 2 shows the outline of the proposed method. It is divided into two processes to create the parallel data used for learning and to emphasize the singing voice of the amateur singer. After voice enhancement using highway networks, the singing voice is resynthesized using a vocoder, where the fundamental frequency (F0) and the aperiodicity index of the amateur ' s voices are used.

A. Creation of parallel data

Regarding the singing voice used in this work, the song tempo is different among all singers (even if all singers sing the same song), and it is necessary to create parallel data. Sangeon *et al.* [2] shows that alignment errors occur at places where vibrato or pitch fluctuation is large when creating parallel data for singing voices. For this reason, in this work, we compare the singing voice of amateurs and that of professionals with the lyric information (consisting of 44 phonemes) and singing voices, and then carry out forced alignment to obtain the time length of the phoneme. Then, DTW (Dynamic Time Warping) [4] is applied for each phoneme to obtain parallel data. The path of the DTW is shown in Fig. 3. It shows the spectrograms from 0 Hz to 5,000 Hz. As shown in Fig. 3, the obtained alignment avoids any influence from vibrato.

B. Spectrum enhancement

Highway networks are widely used for voice signal processing, such as voice synthesis [5] and voice conversion [6]. In this work, using highway networks, the amateur voice is converted to the professional opera voice at the frequency band that represents the remarkable characteristic of the professional singer. For the learning of highway networks, the spectrum obtained by analyzing the amateur singing voice is used as the network input and the spectrum obtained from the voice of the opera singer is used as supervised data. Fig. 4 shows the outline of the forward propagation of the highway networks. Propagation of the highway networks is calculated as follows.

$$y = H(x) \circ T(x) + x \tag{1}$$

Here, $\circ$ represents the element-wise multiplication. $H(\cdot)$ is the non-linear function consisting of the delta calculation, spectral differences estimation described as feed-forward neural networks, and parameter generation using MLPG algorithm [12]. $T(\cdot)$ is the highway networks gate function described as feed-forward neural networks. Each value of the gate output $T(x)$ ranges [0.0-1.0] and weights the spectral differences $H(x)$. When each value of the gate output $T(x)$ is zero, input features

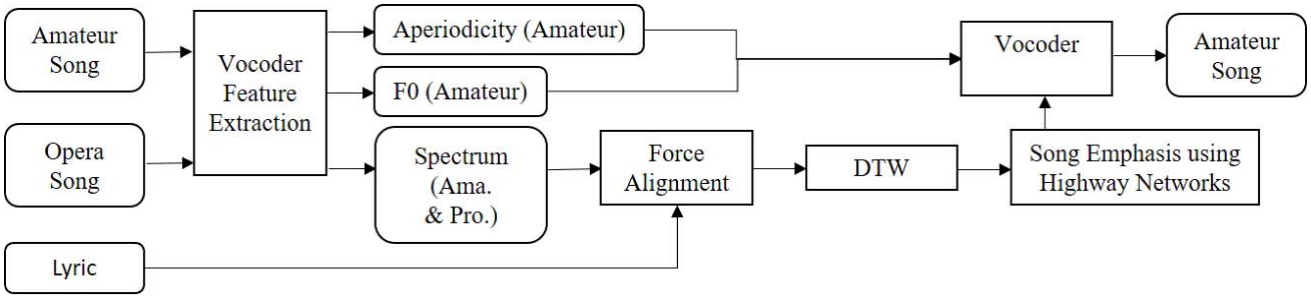


Fig. 2. Proposed method.

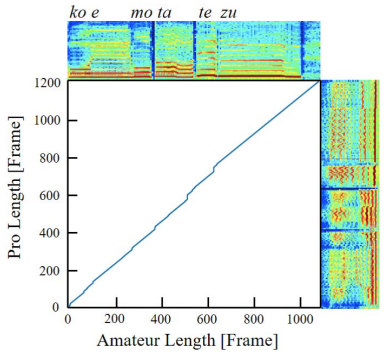


Fig. 3. Example of DTW path between two singing voices.

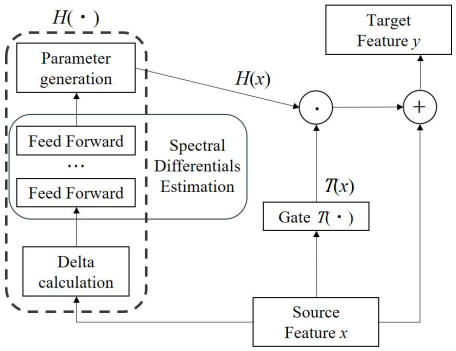


Fig. 4. Singing voice emphasis using input-to-output highway networks.

x become the network prediction. When $T(x)$ is one, these networks become residual networks.

IV. Experimental Evaluation

A. Experimental conditions

As experimental data, singing voices of one soprano opera singer and one amateur without opera experience were used. We used 15 songs of about one and half minutes in length (including the silent section) centered on nursery rhymes as learning data. The audio data were sampled at 16 kHz with a bit depth of 16 bits and shifted every 5ms. For the training data of the singing voice, the silent section including the rest, is removed from the parallel data. WORLD [13] was used to extract the spectra, fundamental frequency, and aperiodicity indices. For the highway network features, 59-dimensional mel-cepstrum [14] obtained by applying a mel-filter bank to a spectrum extracted by WORLD was used. Both input and output are normalized to a mean of 0 and a variance of 1.

For the highway networks that estimate the difference spectrum, a feedforward network, which has two hidden layers, was used. Adagrad [15] is used for the learning algorithm of the network, and its learning rate is set to 0.01. Each hidden layer has 256 units, and the activation function of the hidden layer is ReLU. As the activation function of the output layer, a sigmoid function was used. The network of input layer, which has 59 units, is used and the output layer, which has 59 units for the gate function, is used.

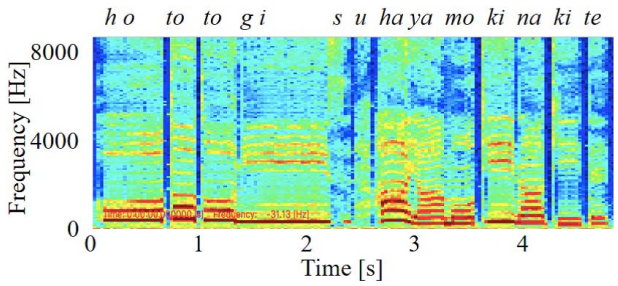
B. Experimental results

The spectrograms of an amateur voice, a converted voice by using the proposed method, and a professional voice (that is not included in the training data) are shown in Fig. 5. Fig. 5 shows that the energy of the band of 3,000 Hz to 4,000 Hz of the vowel of the voice of the amateur is emphasized by the proposed method. In addition, for the low-frequency band including the speaker's characteristics, the spectrum of the amateur singer is maintained though the professional voice is used as the teacher of highway networks.

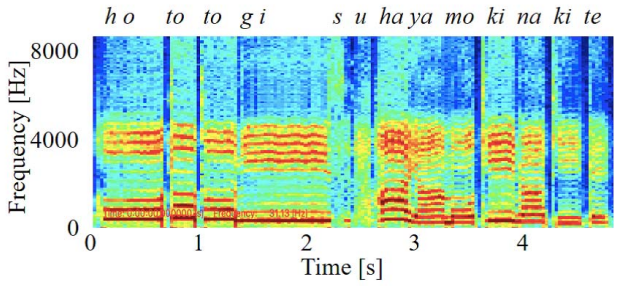
The spectrum for the phonemes "o" and "i" before and after conversion are shown in Fig. 6. It shows that the components in the frequency band from 3,000 Hz to 4,000 Hz are emphasized. However, as a result of smoothing due to the neural networks, small fluctuations are ignored, and it is considered that a decrease of the energy of high-frequency components causes the deterioration of the sound quality.

V. Conclusion

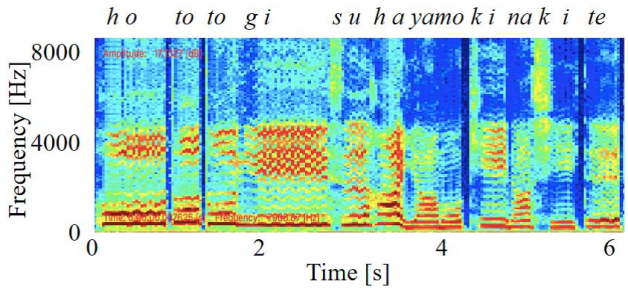
In this paper, a novel singing-voice enhancement system was proposed, where a singing voice of an amateur singer is converted to that of a professional opera singer at the frequency band that represents the remarkable characteristic of the professional singer. Comparing the professional singing voice with that of an amateur, the singing voice of the professional singer had strong energy in the frequency band



(a) An amateur singer.



(b) Proposed method.



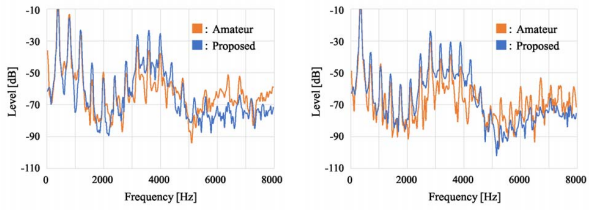
(c) A professional opera singer.

Fig. 5. Examples of spectrograms.

from 3,000 Hz to 4,000 Hz, which contained a lot of frequency components that affect glossiness.

For preprocessing, forced alignment was carried out using musical score information for training data, and DTW was applied to each phoneme to obtain parallel data. When training a singing-voice enhancement system, highway networks are used to estimate the difference between the amateur voice and the professional opera voice. As a result of our experiments, the singing voice of the amateur singer at the middle-high frequency range was emphasized while maintaining speaker characteristics.

However, since it is considered that the sound quality has deteriorated due to smoothing by the neural networks, in the future, in order to prevent deterioration of the sound quality, we will consider the band emphasis and style conversion of the singing voice signal.



(a) spectrum of “o”. (b) spectrum of “i”.

Fig. 6. Spectrum enhancement using highway networks.

REFERENCES

[1] S. Bock and G. Widmer, “Maximum filter vibrato suppression for onset detection,” in *Conference on Digital Audio Effects*, 2013.

[2] S. Yong and J. Nam, “SINGING EXPRESSION TRANSFER FROM ONE VOICE TO ANOTHER FOR A GIVEN SONG,” in *IEEE ICASSP*, pp. 151-155, 2018.

[3] S. Rupesh Kumar, G. Klaus, and S. Jürgen, “Highway networks,” *arXiv preprint arXiv:1505.00387*, 2015.

[4] P. Senin, “Dynamic time warping algorithm review,” *Technical reports, Information and Computer Science Department, University of Hawaii, USA*, pp. 1-23, vol. 855, 2008.

[5] W. Xin, S. Takaki, and J. Yamagishi, “Investigating very deep highway networks for parametric speech synthesis,” *SSW-9*, 2016.

[6] Y. Saito, S. Takamichi, and H. Saruwatari, “Voice Conversion Using Input-to-Output Highway Networks,” in *IEICE Transactions on Information and Systems*, vol. 100, No. 8, pp. 1925-1928, 2017.

[7] Y. Saito, S. Takamichi, and H. Saruwatari, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *IEEE ICASSP*, pp. 1315-1318, 2000.

[8] H. Kawahara, M. Ikuyo, and D. Alain, “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, pp. 187-207, 1999.

[9] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” in *IEEE ICASSP*, pp. 137-140, 1992.

[10] H. Kawahara, E. Jo, 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,” in *MAVEBA*, pp. 59-64, 2001.

[11] “Methods for subjective determination of transmission quality,” in *ITU-T Recommendation*, p. 800, 1996.

[12] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *IEEE ICASSP*, pp. 1315-1318, 2000.

[13] 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. 99, No. 7, pp. 1877-1884, 2016.

[14] T. Fukada, K. Tokuda, T. Kobayashi, and T. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” in *IEEE ICASSP*, pp. 137-140, 1992.

[15] D. John, H. Elad, and S. Yoram, “Adaptive subgradient methods for online learning and stochastic optimization,” *Journal of Machine Learning Research*, vol. 12, pp. 2121-2159, 2011.