Opera Singing Voice Synthesis Considering Vowel Variations

Kenta Katahira* Yuji Adachi† Kiyoto Tai† Ryoichi Takashima* Tetsuya Takiguchi*

* *Graduate School of System Informatics, Kobe University, Japan*
† *MEC Company Ltd., Japan*

***Abstract*—In order to improve a synthesized opera singing voice that is obtained using deep neural networks (DNN), we focus on the vowel variations of the opera singing voice. Variations in the F0 of the vowels of the opera singing voice are analyzed using k-means clustering. Then, from the analysis results, a new score feature for a DNN-based voice synthesis system is introduced as a one-hot vector. The objective evaluation results show that the use of the new score feature improves the quality of the synthesized opera singing voice.**

***Index Terms*—opera singing voice synthesis, deep neural network, vowel variations**

I. Introduction

Singing voice synthesis systems generate singing voices based on the pitch, length, and other information of a given musical score or lyrics. These systems are currently based on statistical parametric singing voice synthesis using deep learning [1].

Generally, researchers have made use of popular music when carrying out their studies. In this paper, because we want to achieve singing voices with rich expressiveness, we focus on opera music as the object of our study because it is noted for the varied expressiveness of the singing voice. Opera music has some characteristics that differ from those of general popular music [2]. We analyze vowel variations of the opera singing voice and improve its voice synthesis based on deep neural networks, taking these variations into cosideration.

II. Analysis of Vowel Variations

A. Vowel Variations

Vowels are determined by the tongue position, the lip shape, and the adjustment of the jaw opening. For example, in the case of Japanese vowels, “a” and “o” (as in “army” and “open”), the jaw is widely opened, and in the case of “i”, “e”, and “u”, the jaw is not widely opened.

For the opera singing voice, the frequency of the first formant becomes higher by increasing the jaw opening when pronouncing vowels, and it makes the singing voice easier for the listener to hear. Therefore, the pronunciations of “i”, “e”, and “u” may be close to those of “a” and “o” when opera singers try to keep the jaws open to produce a high-quality opera singing voice.

Fig. 1 shows the spectra of “u”, “o”, and a pronunciation variation of “u”, where a professional opera singer sings in C5.

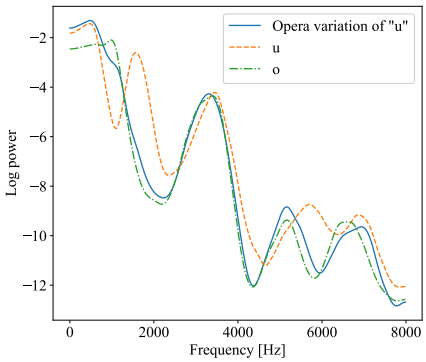

Fig. 1. An example of a variation of an opera vowel “u”.

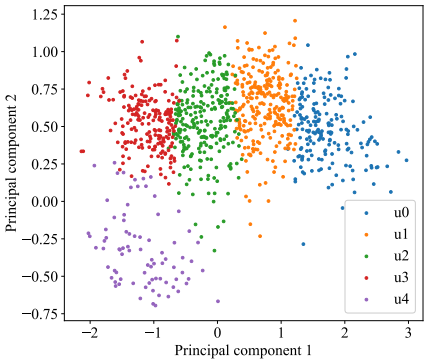

Fig. 2. Clustering results of “u”.

As shown in this figure, there appears to be a strong frequency component of “u” around 1,700 Hz, but there does not appear to be a strong part of “o” and a pronunciation variation of “u” around 1,700 Hz. It is considered that the spectrum shape of the pronunciation variation of “u” seems to be similar to that of “o”.

B. Variation Tendencies

We analyze the vowel distribution of an opera singing voice database consisting of 93 minutes of singing voice (48 songs) by a Japanese professional female opera singer.

All vowel utterances are extracted from the database and the 60-dimension mel-cepstrum averaged for each vowel utterance is used for the vowel variation analysis. Then, the principal

978-1-7281-9802-6/20/$31.00 ©2020 IEEE

TABLE I
AVERAGE F0 OF CLUSTERS OF "U"

Cluster	u0	u1	u2	u3	u4
Average F0 [Hz]	304.78	355.74	437.19	513.87	608.92

component analysis is applied to the mel-cepstrum, where the number of principal components is 10.

Fig. 2 shows the clustering result of "u", where the k-means clustering is used. We set $k = 5$ which is the best score in the experiments. For example, the feature vectors whose first principal component is less than 0 and the second one is less than 0.25 are classified into the cluster "u4". In comparison with the features of "o", we found that the spectra in the cluster "u4" are close to those of the pronunciation of "o".

Table I shows the average F0 for each cluster. From Fig. 2 and Table I, we see that, as the first principal component score decreases, the F0 of the vowel "u" becomes higher. A similar tendency is observed for other vowels as well. Also, as the F0 of the vowel "u" becomes high, it is thought that the pronunciation becomes similar to that produced by "jaw opening".

C. Score Feature

In this work, a DNN-based acoustic model is employed for opera singing voice synthesis, where the input is a score feature. To consider the vowel variation described in the previous subsection, a new score feature is introduced as a one-hot vector using the vowel cluster ID that is obtained in II-B, where the number of dimensions of the one-hot vector is 25 (= 5 vowels $\times$ 5 clusters).

III. EXPERIMENT AND RESULTS

A. Experiment Conditions

We used an opera singing voice database consisting of 93 minutes of singing (48 songs) by a Japanese professional female opera singer. Forty-three of the songs are used for training a DNN, and 5 songs are used for the test.

A 534-dimension score feature is used as the input feature of a DNN-based baseline system. In our proposed method, a novel score feature, which includes the information of the vowel cluster, is added to the 534-dimension score feature.

For acoustic features, WORLD [3] is used to extract the 60-dimensional mel-cepstrum, logarithmic fundamental frequency, and band aperiodicity, with their 1st- and 2nd-order derivatives, and a voiced/unvoiced binary value is also used.

A bi-directional gated recurrent unit is used as an acoustic model of opera singing voice synthesis, where global variance is employed for training the acoustic model [4]. The GRU network consists of three layers with 1,024 units.

B. Evaluation

Table II shows objective evaluation results. The new score feature, which considers vowel variations, improves the mel-cepstral distortion and the global variance distance in comparison with conventional musical score features. In our approach,

TABLE II
OBJECTIVE EVALUATION RESULTS.

	MGC (dB)	GVD
Conventional	5.378	1.955×10^{-1}
Proposed	5.258	1.567×10^{-1}

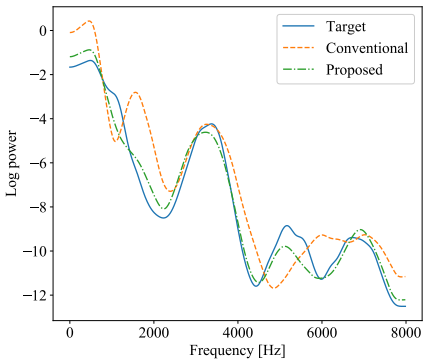

Fig. 3. Comparison of synthesized "u" spectral envelopes.

to consider vowel variations of an opera singing voice, a vowel is classified into five types of utterances using k-means clustering. By using the new feature obtained from the vowel cluster, high quality opera singing voice synthesis is achieved. The clustering seems to help the model learning to classify vowels on the similarity of mel-cepstrum.

Fig. 3 shows the spectral envelope of the target utterance "u" and the synthesized spectral envelopes. The spectral envelope synthesized without using vowel variation information seems to be different from that of the target utterance. On the other hand, the spectral envelope synthesized by considering vowel variations is similar to that of the target utterance. Therefore, it can be seen that the use of the vowel clustering is effective for dealing with the opera vowel variations.

IV. CONCLUSION

In this paper, a new score feature for a DNN-based opera singing voice synthesis was introduced, considering the vowel variations of the opera singing voice. To deal with the vowel variations, each vowel was classified into five clusters according to the spectrum shape and the F0 frequency by using k-means clustering. Then, by adding the cluster information to the musical score feature, we were able to synthesize an opera singing voice of higher sound quality.

REFERENCES

[1] Y. Hono et al., "Recent Development of the DNN-based Singing Voice Synthesis System — Sinsy," In Proc. APSIPA, pp. 1003-1009, 2018.
[2] R. Nanzaka, T. Kitamura, Y. Adachi, K. Tai, T. Takiguchi, "Spectrum Enhancement of Singing Voice Using Deep Learning," in Proc. IEEE International Symposium on Multimedia, pp. 167-170, 2018.
[3] M. Morise, F. Yokomori, K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," IEICE Trans. on Information and Systems, 99(7), 1877-1884, 2016.
[4] K. Hashimoto, K. Oura, Y. Nankaku and K. Tokuda, "Trajectory training considering global variance for speech synthesis based on neural networks," In Proc. ICASSP, pp. 5600-5604, 2016.

978-1-7281-9802-6/20/$31.00 ©2020 IEEE