Advanced Science and Technology Letters
Vol.30 (ICCA 2013), pp.169-173
http://dx.doi.org/10.14257/astl.2013.30.36

Vector Quantization-Based Parameter Control for Speech Conversion with Improved Naturalness

*Kwang Myung Jeon, *Woo Kyung Seong, *Hong Kook Kim, and **Sung Dong Jo

*School of Information and Communications
Gwangju Institute of Science and Technology (GIST)
{kmjeon, wkseong, hongkook}@gist.ac.kr
**Info-Communications Development Team
Hyundai Motor Company
sdjo@hyundai.com

Abstract. In this paper, a parameter control method is proposed to improve naturalness of converted speech. The proposed method first extracts feature parameters from both input and reference speeches per each frame, and then concatenates them into a feature parameter vector. Next, all the feature vectors are clustered into a certain number of representative vectors by using vector quantization. Finally, the vectors are used to control parameters for speech conversion. It is shown from the performance evaluation that the proposed method can convert speech with more naturalness than a linear parameter control method.

Keywords: Voice conversion, parameter control, naturalness, vector quantization

1 Introduction

Speech conversion has been applied in various fields, such as security, multimedia, and front-end processing for speech recognition [1][2]. In general, speech conversion consists of speech parameterization, parameter control, and speech re-synthesis from the controlled parameters [1][2]. In particular, the parameter control directly affects the naturalness of converted speech because it changes fundamental factors for speech characteristics such as level of pitch, tone, and speaking rate [1].

Conventional parameter control methods for speech conversion can be classified into two types: artificial parameter control and target speaker-based parameter control [2]. The artificial parameter control method heuristically sets parameters, thus it is suitable for obtaining various conversion results. However, the artificial parameter control method cannot guarantee the naturalness of converted speech as it does not consider correlation of each parameter while controlling [2]. For this reason, speech converted by the artificial parameter control method is applicable only for security or entertainment, not for databases (DB) of speech synthesis and recognition, which require high naturalness.

ISSN: 2287-1233 ASTL

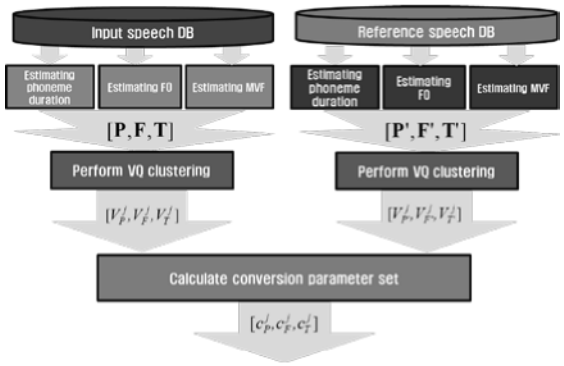

Fig. 1. Procedure of the proposed vector quantization-based parameter control method.

On the other hand, the target speaker based parameter control method changes parameters by morphing characteristics of input speech into those of a target speaker. To this end, feature mapping methods using a vector quantization (VQ)-based codebook and Gaussian mixture model (GMM) have been proposed [2][3]. Thus, since the target speaker based parameter control method well reflects speaking characteristics of the target speaker, it can provide high naturalness of converted speech compared to the artificial parameter control method. However, its dependency on the target speaker makes it hard to convert different speech types exhibiting various characteristics.

In order to mitigate such limitations, this paper proposes a parameter control method that converts speech with various characteristics while maintaining naturalness. First, the proposed method extracts feature parameters from the input and the reference speeches per frame, and then concatenates them into a feature parameter vector. Next, all the vectors are clustered into a certain number of representative vectors by using VQ. After that, conversion ratios between the input speeches and the reference speeches are obtained through the clustered centroids. Note that these conversion ratios are regarded as a set of the conversion parameters. Finally, multiple sets of the conversion parameters are applied to speech conversion. The overall procedure of the proposed method is similar to that of the conventional target speaker-based parameter control method due to the presence of the reference speeches. However, the major difference between the two methods is that while the conventional method tries to mimic characteristics of the only one speaker, the proposed method can convert different speech sources with various characteristics without losing naturalness, owing to the reference speeches composed of multiple speakers.

Following this introduction, Section 2 proposes a VQ-based parameter control method. Section 3 evaluates the performance of the proposed method. Finally, Section 4 concludes the paper.

2 Proposed VQ-Based Parameter Control Method

Fig. 1 shows an overall procedure of the proposed VQ-based parameter control method for speech conversion. As shown in the figure, the proposed method requires a

certain volume of input speech data and reference speech data, where input and reference speech DB consist of utterances of one speaker and multiple speakers, respectively. Details of the proposed method are as follows.

First, the proposed method calculates level of tone and speaking rate per frame of the speeches which belong to the input and reference DBs, respectively. Specifically, the level of tone is measured by calculating fundamental frequency (F0), *P*, and maximum voiced frequency (MVF), *F* [4]. In addition, the level of speaking rate is measured by estimating duration of each frame, *T*. Each set of *P*, *F*, and *T* of input speech DB and target speech DB is stacked up to the total number of frames. From now on, [**P**, **F**, **T**] and [$\mathbf{P}'$,$\mathbf{F}'$,$\mathbf{T}'$] denote the set of vectors for input speech and reference speech DB, respectively. In this paper, *P*, *F*, and *T* are calculated by the F0 estimation method based on instantaneous frequency [5], the MVF estimation method based on subband energy difference [4], and phoneme duration estimation based on a forced alignment using a speech recognizer [6], respectively.

Next, VQ is trained on [**P**,**F**,**T**] and [$\mathbf{P}'$,$\mathbf{F}'$,$\mathbf{T}'$] by the Linde-Buzo-Gray (LBG) algorithm [7] to obtain *N* centroids of the vectors. Note that the *j*-th centroids of [**P**, **F**, **T**] and [$\mathbf{P}'$, $\mathbf{F}'$, $\mathbf{T}'$] are represented as $[v_P^j, v_F^j, v_T^j]$ and $[v_{P'}^j, v_{F'}^j, v_{T'}^j]$, respectively. Subsequently, centroids between the input speech DB and reference speech DB are paired. In other word, a centroid of [$\mathbf{P}'$,$\mathbf{F}'$,$\mathbf{T}'$], which is the nearest one to the *j*-th $[v_P^j, v_F^j, v_T^j]$, is found as

$$j^* = \arg\min_{1\le k\le N} \sqrt{w_P (v_P^j - v_{P'}^k)^2 + w_F (v_F^j - v_{F'}^k)^2 + w_T (v_T^j - v_{T'}^k)^2} \quad (1)$$

where w_P, w_F, and w_T correspond to weighting values for *P*, *F*, and *T*, respectively, all of which are set to 1 in this paper. From the result of Eq. (1), the *j*-th conversion parameter, $\{c_P^j, c_F^j, c_T^j\}$, is obtained by the ratio between the *j*-th centroid of the input speech DB and the j^*-th centroid of the reference speech DB, given as

$$[c_P^j, c_F^j, c_T^j] = \left[\frac{v_{P'}^{j^*}}{v_P^j}, \frac{v_{F'}^{j^*}}{v_F^j}, \frac{v_{T'}^{j^*}}{v_T^j}\right], \quad 1 \le j \le N. \quad (2)$$

Finally, $[c_P^j, c_F^j, c_T^j]$ $(1 \le j \le N)$ is applied to speech conversion based on the pitch synchronous harmonic and non-harmonic model (PS-HNH) [9], resulting in an *N* times larger-sized converted speech DB compared to the input speech DB.

3 Performance Evaluation

In order to evaluate the performance of the proposed VQ-based parameter control method, subjective quality comparison was conducted on converted speech DBs that were obtained by both the conventional linear parameter control method and the proposed method. To this end, the input speech DB consisted of 50 utterances of an

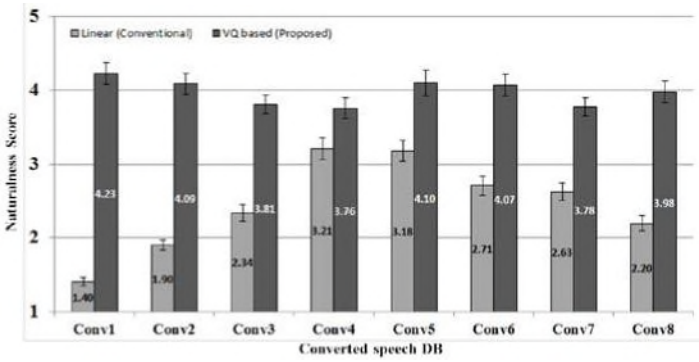

Fig. 2. Comparison of naturalness scores between the different parameter control methods applied to eight different speech conversions.

American male, while the reference speech DB consisted of 50 utterances of three American males and two American females (250 utterances in total). The conventional method set $[c_P, c_F, c_T]$ to eight different values ranging from 0.3–1.7 in increments of 0.2, whereas the proposed method obtained $[c_P, c_F, c_T]$ from Eq. (2) with N=8. Next, speech conversion was performed using the PS-HNH speech analysis and synthesis technique [9]. Finally, naturalness of each converted speech DB was evaluated by nine listeners who were all males aged 20–30 years old. All listeners were asked to give a score with scale of 1 to 5 for each converted speech.

Fig. 2 compares naturalness scores of the converted speeches by the conventional and proposed method. In this figure, numbers on the bars indicate naturalness scores averaged over all the listeners and converted speech files for a given $[c_P, c_F, c_T]$. It was shown from the figure that the converted speeches obtained by the proposed method achieved a higher (by 1.53) average score than those obtained by the conventional method.

4 Conclusion

In this paper, a VQ-based parameter control method for speech conversion was proposed to provide better naturalness in converted speech. The proposed method used VQ to cluster feature parameters of input speeches and reference speeches, and then obtained conversion parameters from the centroids. Subjective listening testing indicated that the proposed method could generate converted speech with higher naturalness than the conventional linear parameter control method.

Acknowledgments. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the government of Korea (MSIP) (No. 2012-010636), and by the MSIP, Korea, under the Information Technology Research

Center (ITRC) support program supervised by the National IT Industry Promotion Agency (NIPA) (NIPA-2013-H0301-13-4005).

References

1. Abe, M., Nakamura, S., Shikano, K., Kuwabara, H.: Voice conversion through vector quantization. In: Proceedings of ICASSP, (1988) pp. 655-658.
2. Arslan, L. M., Talkin, D.: Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum. In: Proceedings of Eurospeech, (1997) pp. 1347-1350.
3. Toda, T., Saruwatari, H., Shikano, K.: Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. In: Proceedings of ICASSP, (2001) pp. 841-844.
4. Jeon, K. M.: Harmonic and Non-harmonic Modeling of Speech for Statistical Parametric Speech Synthesis. Master Thesis, School of Information and Communications, Gwangju Institute of Science and Technology, (2012).
5. Kawahara, H., Estill, J., Fujimura, O.: Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. In: Proceedings of 2nd MAVEBA, (2001) pp. 13-15.
6. Zen, H., Tokuda, K., Black, A. W.: Statistical parametric speech synthesis. Speech Communication, 51(11), (2009) pp. 1039-1064.
7. Linde, Y., Buzo, A., Gray, R. M.: An algorithm for vector quantizer design. IEEE Transactions on Communications, 20(1), (1980) pp. 84-95.
8. Kawahara, H., Masuda-Katsuse, I., Cheveigné A.: 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, 27(3), (1999) pp. 187-207.
9. Jeon, K. M., Kim, H. K.: High-quality speech modification based on pitch-synchronous harmonic and non-harmonic modeling of speech. Advanced Science and Technology Letters, 14(1), (2012) pp. 176-179.

