IMPROVING VOICE QUALITY OF HMM-BASED SPEECH SYNTHESIS USING VOICE CONVERSION METHOD

Yishan Jiao, Xiang Xie, Xingyu Na, Ming Tu

School of Information and Electronics
Beijing Institute of Technology
Beijing, China

ABSTRACT

HMM-based speech synthesis system (HTS) often generates buzzy and muffled speech. Such degradation of voice quality makes synthetic speech sound robotically rather than naturally. From this point, we suppose that synthetic speech is in a different speaker space apart from the original. We propose to use voice conversion method to transform synthetic speech toward the original so as to improve its quality. Local linear transformation (LLT) combined with temporal decomposition (TD) is proposed as the conversion method. It can not only ensure smooth spectral conversion but also avoid over-smoothing problem. Moreover, we design a robust spectral selection and modification strategy to make the modified spectra stable. Preference test shows that the proposed method can improve the quality of HMM-based speech synthesis.

Index Terms— HMM-based speech synthesis, voice conversion, local linear transformation, temporal decomposition

1. INTRODUCTION

HMM-based speech synthesis is popular in recent decades due to its flexibility and adaptability [1]. However, the voice quality is its biggest drawback compared with unit-selection synthesis [2]. According to [3], the degradation of voice quality is mainly caused by three factors: the vocoder, the accuracy of acoustic models, and over-smoothness. There are many attempts to alleviate these problems, such as high-quality vocoders like STRAIGHT [4], better acoustic models like the trajectory HMMs [5], and some methods to enhance details of over-smoothed spectra like the speech parameter generation algorithm considering global variance [6]. All these previous studies tried to solve the problem from the causes, while in this paper, we view it from the consequence. In fact, no matter which part of HTS goes wrong, the only result is the low quality speech.

Based on this consideration, we assume that synthetic speech and the original speech are from different speaker spaces, and we propose to use voice conversion method to improve voice quality. Among various voice conversion methods, GMM-based [7] and linear transformation-based [8] methods are two major ones. Considering the over-smoothing problem, linear transformation seems a good choice. [9] describes an effective method called local linear transformation (LLT) which can preserve spectral details after conversion. We propose to use LLT combined with temporal decomposition (TD) [10] as the conversion method because temporal decomposition (TD), which decomposes spectral parameters into a set of events, could avoid discontinuity between consecutive frames. Moreover, a robust selection and modification strategy is also proposed to keep the converted spectra stable. Preference test proves that our proposed method is an effective way to improve the quality of synthetic speech.

Relation to prior work. Our work is related to voice conversion studies [8] [9]. Their purpose was to change a source speaker's style to another speaker's, while our aim is to modify synthetic speech so that its quality can be improved. Our work is also related to the studies that used real speech data to alleviate over-smoothness of synthetic speech [11] [12]. These previous studies, however, used training data directly to control parameter generation process or to replace synthetic speech. In contrast, our study tries to find out the mapping between synthetic and original speech by building up a parallel synthetic speech dataset against the training corpus.

This paper is organized as follows. Section 2 shows the general conversion framework we propose and the methods of temporal decomposition and local linear transformation. Section 3 describes the proposed robust selection and modification strategy. Experimental setup and the result of subjective test are presented in Section 4. The conclusion and acknowledgement are in Section 5 and 6, respectively.

2. PROPOSED CONVERSION FRAMEWORK

In layman terms, the purpose of voice conversion is to transform a source speaker's voice to another target speaker's. In our work, we assume that synthetic and original speech are in different speaker spaces–we can say that one is from a machinery speaker and the other is from a human speaker. Our aim is to find out the mapping between them and convert the artificial voice to the natural human voice. The general frame-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

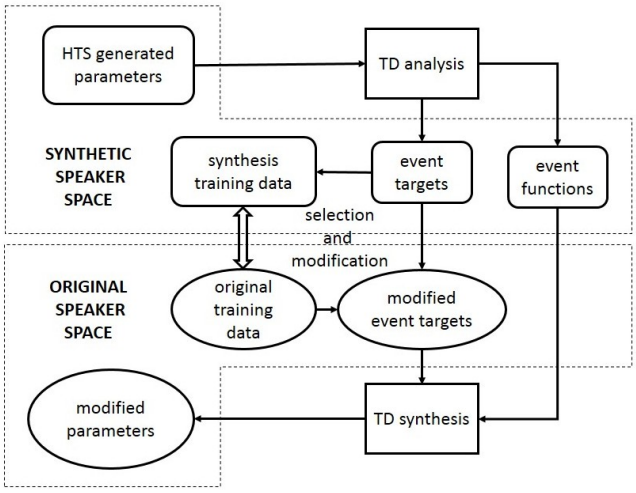

Fig. 1. Proposed conversion framework.

work of our proposed system is shown in figure 1.

As in voice conversion, a parallel dataset should be first built up. In our work, we construct this dataset by re-synthesizing training speech using HTS [13]. This synthetic speech is aligned to the original training data with the guide of labels. Note that there is no need to synthesize speech waveforms because we only need a parallel parameter database.

Before conversion, spectral parameters are first decomposed into a sequence of overlapping event functions and the corresponding event targets with TD as in (1).

$$\hat{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}_k \phi_k(n),\ 1 \leqslant n \leqslant N \tag{1}$$

where $\mathbf{a}_k$ is the kth event target. $\phi_k(n)$, the kth event function, describes the temporal evolution from the kth target to the next. The approximation of the nth spectral parameter vector, $\mathbf{y}(n)$, is represented by $\hat{\mathbf{y}}(n)$. N is the number of frames in the analyzed speech segment.

Event functions are related to the content or intelligibility of speech, while event targets, which are context-independent, relate to voice quality or the speaker's style [14]. Since speech synthesized by HTS usually has good intelligibility, event functions are preserved to keep the transformation continuous. On the other hand, event targets are transformed from synthetic speaker space to the original speaker space with voice conversion method.

To avoid over-smoothing problem, we adopt local linear transformation (LLT) method. The main idea of LLT is to select a set of neighbors for each source vector (synthetic target vector) and compute the transformation between these vectors and their aligned target vectors (original training parameters) as in (2).

$$\mathbf{N}^s \mathbf{W} = \mathbf{N}^o \tag{2}$$

where $\mathbf{N}^s$ and $\mathbf{N}^o$ are the aligned synthetic and original training data. The local regression or the transformation $\mathbf{W}$ is obtained by solving (3) using least square method:

$$\mathbf{W} = ((\mathbf{N}^s)^{\mathrm{T}} \mathbf{N}^s)^{-1} (\mathbf{N}^s)^{\mathrm{T}} \mathbf{N}^o \tag{3}$$

After that, the transformation $\mathbf{W}$ can be applied to convert the synthetic event target as in (4).

$$(\mathbf{a}_{\text{conv}})^{\mathrm{T}} = (\mathbf{a}^s)^{\mathrm{T}} \mathbf{W} \tag{4}$$

In TD synthesis part, the modified event targets are combined with the preserved event functions to get new spectral parameters, which are in the original speaker space.

3. SELECTION AND MODIFICATION STRATEGY

In our method, linear spectral frequency (LSF) is used as spectral parameter due to its close relation to the formant structure. One of the properties of LSF is that all the parameters in an LSF vector must be ordered in sequence so that the spectrum can be stable. However, transformations may destroy its order if the selection is not precise. Thus we use the labels to confine the scope of selection. Besides, it can also reduce searching time.

Nonetheless, it still cannot guarantee the converted LSF vectors are all stable. It also concerns with the accuracy of the selected neighbor set. In [9], the size of neighbor set is fixed to 40 (female to male) and 130 (male to female). In our work, we adapt this number according to the range of distances between alternative neighbors and the source vector. Only the one whose distance to the source vector is under a threshold oughts to be put into the neighbor set. In this way, a number of variable-length neighbor sets are constructed as in (5).

$$\mathbf{N}(\mathbf{a}^s) = \{\mu_1^s, \mu_2^s, ..., \mu_k^s\} \tag{5}$$

where $\mathbf{N}(\mathbf{a}^s)$ is the neighbor set of the source vector $\mathbf{a}^s$. The distances to $\mathbf{a}^s$ of all the neighbors in $\mathbf{N}(\mathbf{a}^s)$ have an ascending order, which are all smaller than a threshold we set. It means that μ_1^s is nearest to the source vector.

Moreover, if the selected neighbor set has few alternatives or the converted event target $(\mathbf{a}_{\text{conv}})^{\mathrm{T}}$ is still in disorder, we turn to the replacement method. It means that the synthetic event target will be replaced by the original data whose aligned synthetic training parameter is nearest to the source vector. It can be represented by (6)

$$(\mathbf{a}_{\text{modified}})^{\mathrm{T}} = \begin{cases} (\mathbf{a}_{\text{conv}})^{\mathrm{T}}, \text{ if } (\mathbf{a}_{\text{conv}})^{\mathrm{T}} \text{is stable} \\ \mu_1^o, \text{ otherwise} \end{cases} \tag{6}$$

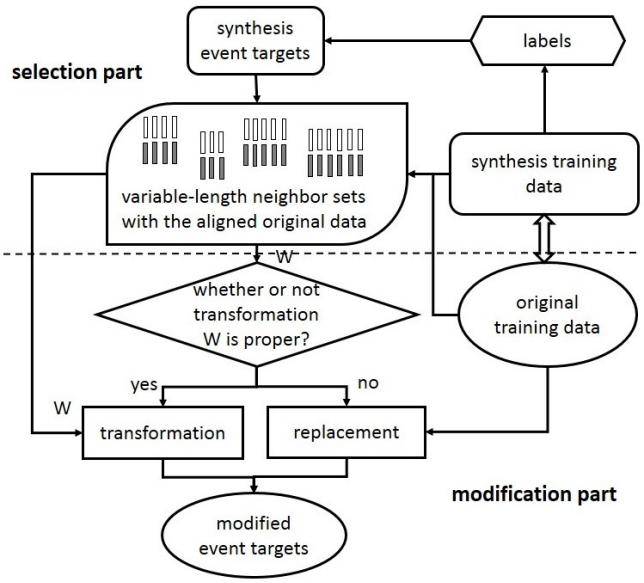

Fig. 2. Selection and modification strategy.

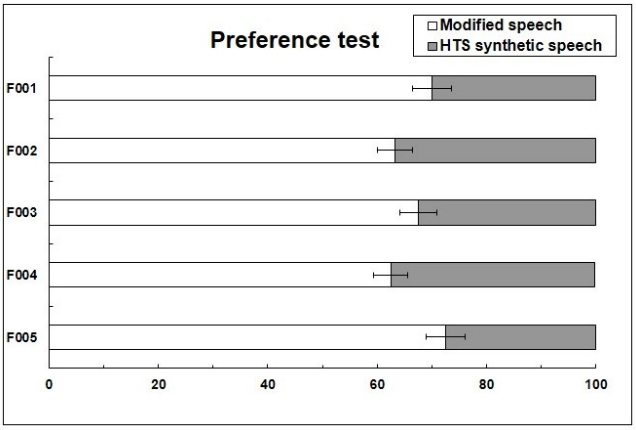

Fig. 3. Preference test between modified speech and HTS synthetic speech for 5 speakers with 95% confidence interval.

where μ_1^o is the aligned original vector of μ_1^s. The proposed selection and modification strategy is illustrated in figure 2.

Note that although we also make some replacement like [12], the distance we measure is between vectors from the same space, the synthetic speaker space, rather than different spaces. This is because we believe that the nearest distance between vectors from different spaces cannot reflect their real mapping relationship. However, if we can find out a neighbor for the source vector in its own space, the mapping between this neighbor and its counterpart in the target space could be reasonably applied to convert this source vector.

4. EXPERIMENTS

4.1. Experimental setup

We used ASCCD Mandarin speech corpus [15], including 5 speakers (F001-F005) and each speaker 300 utterances, for HTS training and synthesis. Each speaker has approximately 15 minutes speech. All speech was sampled at 16 kHz and windowed by a 25-ms hamming window with a 5-ms frame shift. STRAIGHT was used as the vocoder to analyze speech and synthesize waveforms. First we obtained fundamental frequency (F0), aperiodicity (AP), and spectral envelope (SP) with STRAIGHT, and then SPTK-3.5 [16] was used to generate linear spectral frequency (LSF) from spectral envelopes. The order of LSF was set to 16. In the uniform framework of HMM [17], feature vectors were composed of spectral parameters, log F0 and their delta and delta-delta coefficients. 10-state (including the start and the end states) context-dependent HMMs were used for training.

To build up a parallel synthetic dataset against the original one, we used the generated HMMs from HTS to re-synthesize training speech with the guide of training labels. Only LSF parameters with their aligned original data were saved.

50 sentences (10 for each speaker) which were not contained in the training data were also synthesized with HMMs as the test speech. Pitch and aperiodic parameters were preserved, while LSF parameters were decomposed with an up-to-date TD called modified restricted temporal decomposition (MRTD) [18]. The density of events was about 40 in one second. Each event target was regarded as a source vector, and its neighbor set was selected from the synthetic training data by K nearest neighbor (KNN) method. Here, K was set to 20. All the selection was under the guide of labels as in figure 2 . According to the selection strategy mentioned above, unqualified vectors were excluded from the neighbor set if their distances to the source vector were larger than a threshold. Empirically, it was set to 0.3. Transformation $\mathbf{W}$ was computed by solving (3) using least square method. $\mathbf{W}$ would then be determined whether or not suitable to convert the source vector by (6).

After modification, TD synthesis generated the modified LSF parameters which were then transformed to spectral envelops. STRAIGHT output speech waveforms with these modified spectral parameters and the preserved F0 and AP.

4.2. Experimental results

Preference test was conducted to evaluate the proposed method. Six native Mandarin speakers who had normal hearings were asked to listen to 50 pairs (10 for each speaker) of synthesis and modified speech then gave their preference on the quality. The results were shown in figure 3. From the figure we can see that our proposed method is an effective way to improve voice quality of synthetic speech.

5. CONCLUSION

In this paper, we have proposed to regard synthetic speech from HTS in a different speaker space from the original speech. So we proposed to use voice conversion method to build up the mapping between these two spaces. LLT combined with TD is proposed as our conversion method. It can keep the conversion smooth but not over-smooth. We have also designed a robust selection and modification strategy to ensure the conversed LSF parameters stable. Subjective test has shown that our proposed method can improve the quality of HMM-based speech synthesis.

In the future, we will investigate more conversion methods to improve both voice quality and prosody of the synthetic speech.

6. ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (Grant No. 91120015, No. 90920304, No. 11161140319, No. 61001188).

We are particularly grateful to Prof. Masato Akagi and Dr. Trung-Nghia Phung from JAIST for their warm discussion and helpful suggestion.

7. REFERENCES

[1] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proceedings of IEEE Speech Synthesis Workshop*, September 2002, pp. 227–230.

[2] A. J. Hunt and A. W. Black, "Unit selection in concatenative speech synthesis system using a large speech database," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996, pp. 373–376.

[3] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, November 2009.

[4] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch–adaptive time–frequency smoothing and an instantaneous–frequency–based F_0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, April 1999.

[5] H. Zen, *Reformulating HMM as a trajectory model by imposing explicit relationships between static and dynamic features*, Ph.D. thesis, Nagoya Institute of Technology, Nagoya, Japan, January 2006.

[6] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. E90–D, no. 5, pp. 816–824, 2007.

[7] A. Kain and M. W. Macon, "Spectral voice convesion for text-to-speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, USA, 1998, pp. 258–288.

[8] H. Ye and S. Young, "Perceptually weighted linear transformation for voice conversion," in *Proceedings of EUROSPEECH*, Geneva, Switzerland, 2003.

[9] V. Popa, H. Silen, and J. Nurminen, "Local linear transformation for voice conversion," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012.

[10] B. S. Atal, "Efficient coding of LPC parameters by temporal decomposition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1983, pp. 81–84.

[11] Y. J. Wu and R. H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006, pp. 89–92.

[12] T. N. Phung, T. S. Phan, T. T. Vu, M. C. Luong, and M. Akagi, "Improving naturalness of HMM-based TTS trained with limited data by temporal decomposition," *IEICE Transactions on Information and Systems*, vol. E96–D, no. 11, pp. 2417–2426, November 2013.

[13] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proceedings of Sixth ISCA Workshop on speech synthesis*, 2007, pp. 294–299.

[14] P. N. Binh and M. Akagi, "Efficient modelling of temporal structure of speech for applications in voice transformation," in *Proceedings of Interspeech*, 2009, pp. 1631–1634.

[15] Phonetics Lab, Institute of Linguistics, CASS, "ASCCD: Read discourse corpus with prosodic, segmental and syntactic annotation," http://ling.cass.cn/yuyin/english/resc6.htm, 2006.

[16] S. Imai, T. Kobayashi, K. Tokuda, T. Masuko, K. Koishida, S. Sako, and H.Zen, "Speech signal processing toolkit (SPTK), Version 3.5," http://sptk.sourceforge.net, 2009.

[17] K. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proceedings of EUROSPEECH*, September 1999, pp. 2347–2350.

[18] P. C. Nguyen, T. Ochi, and M. Akagi, "Modified restricted temporal decomposition and its application to low rate speech coding," *IEICE Transactions on Information and Systems*, vol. E86–D, no. 3, 2003.