Nguyen and Phung *EURASIP Journal on Audio, Speech, and Music Processing* (2017) 2017:14
DOI 10.1186/s13636-017-0113-5

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Reducing over-smoothness in HMM-based speech synthesis using exemplar-based voice conversion

Gia-Nhu Nguyen[1] and Trung-Nghia Phung[2*]

Abstract

Speech synthesis has been applied in many kinds of practical applications. Currently, state-of-the-art speech synthesis uses statistical methods based on hidden Markov model (HMM). Speech synthesized by statistical methods can be considered over-smooth caused by the averaging in statistical processing. In the literature, there have been many studies attempting to solve over-smoothness in speech synthesized by an HMM. However, they are still limited. In this paper, a hybrid synthesis between HMM and exemplar-based voice conversion has been proposed. The experimental results show that the proposed method outperforms state-of-the-art HMM synthesis using global variance.

Keywords: Speech synthesis, HMM-based, Voice conversion, Exemplar-based, Non-negative matrix factorization, Over-smooth

1 Introduction

Speech synthesis (SS) is the artificial production of human speech, which is the core part of text-to-speech (TTS) systems that convert text content in a specific language into a speech waveform [1]. SS can be used in several applications, including educational applications, speech to speech translators, and applications for telecommunications and multimedia.

Among many kinds of SS that have been proposed, the state-of-the-art one is HMM-based SS (HMMSS) [1, 2]. In this approach, spectral and prosodic features of speech are modeled and generated in an unified statistical framework using HMMs. HMMSS has many advantages that have been shown in the literature, such as the high intelligibility of synthesized speech, a small footprint, a low computational load, and the flexibility to change the voice characteristics [1]. Although HMMSS has many advantages, the quality of its synthesized speech is still far from natural, which is mainly due to two reasons: buzziness and over-smoothness in synthesized speech. The former is a common issue with speech coding, which has recently been significantly improved, while the latter is caused by averaging in the statistical processing in HMMSS, which is still a remaining problem of HMMSS at present. In HMMSS, the structure of the estimated spectrum corresponds to the average of different speech spectra in the training database due to the use of the mean vector. In this case, the spectrum estimated by HMMs is an average approximation of all corresponding speech spectra in the training database. The detailed structure in the original speech may be missing in this kind of approximation. This characteristic in speech synthesized by HMMSS can be considered too medial or over-smooth in synthesized speech. When synthesized speech is over-smooth, it sounds "muffled" and far from natural [3–5].

The degree of articulation (DoA) provides information on the style/personality [6]. DoA is characterized by modifications of the phonetic context, of the speech rate, and of the spectral dynamics (vocal tract rate of change). Over-smooth speech with too-slow transitions may affect the production of the appropriate DoA, and important information on style/personality may be loss. Over-smoothness may be acceptable for reading speech or neutral speech but not suitable for expressive speech in general. Mainly, the range and the velocity of the tongue tip movements are the primary modulation parameters associated with emotional expression [7].

* Correspondence: ptnghia@ictu.edu.vn
[2]Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam
Full list of author information is available at the end of the article

Therefore, too smooth or too stable speech with slow movements cannot be efficient to represent some kinds of emotional speech with high movements of the tongue tip. Since "over-smoothness" causes a reduction in identification of emotions/expressions/styles in speech that can also affect to the perception of the naturalness, it is one of the main remaining factors reducing the naturalness of HMMSS.

Although both the spectral and prosodic trajectories generated by HMMSS are over-smooth, the effect of over-smoothness in a spectral sequence is more serious since the structures of spectral features are more complex, as shown in [5]. In the literature, there have been many studies attempting to solve over-smoothness in HMMSS. Using multiple mixtures for modeling state output probability density can reduce over-smoothness in synthesized speech [3]. However, these methods cause another problem with over-training due to the increased number of model parameters. A method of combining continuous HMMs with discrete HMMs and a method of increasing the number of HMM states have also reduced the over-smoothness in HMMSS [5]. However, these methods increase the complexity of HMMs and are not convenient in practical synthesis systems.

The state-of-the-art statistical method to reduce over-smoothness is the parameter generation algorithm that takes into consideration global variance (GV) [4]. In this approach, a global variance model is trained to model the variation of temporal parameter trajectories at an utterance level. The generated parameter sequence maximizes a likelihood based on not only an HMM likelihood but also a global variance likelihood, resulting in an increase of the global variance of the generated parameter trajectories. The experimental results with these methods revealed that the naturalness of synthetic speech was significantly improved. However, the average parameter generation and the variance estimations in HMMSS require a large amount of training data, and adaption methods in HMMSS also require a large amount of target data [8]. Therefore, state-of-the-art HMMSS with parameter generation considering global variance still requires a large-scale database with a size in gigabytes for training each voice.

Hybrid approaches between HMMSS and unit selection, such as the HMM trajectory tiling (HTT) system [9], have recently been studied as another solution to reduce over-smoothness in HMMSS. The smooth HMM trajectory is used to guide the selection of each short frame to concatenate the waveforms in HTT. This method is based on the concept of transforming/replacing short frames of speech synthesized by HMMSS to the physically closest frames of the original speech. The naturalness of HTT is comparable to that of unit selection, and its intelligibility is comparable to that of HMMSS and the original speech. However, HTT still has drawbacks. The major one is the use of short frames, which requires a perfect selection process. If the selection process is imperfect due to a limited data corpus, it may be easy to perceive discontinuities between frames. As a result, this synthesizer still requires a huge amount of data for each voice.

Building a large-scale speech corpus is a costly task that takes a long time and a great deal effort by engineers, acousticians, and linguists. The difficulties of building large-scale speech corpora are more serious in under-resourced languages. Synthesizing speech with a limited amount of data is also critical for customizing synthesized speech with multiple individual and emotional voices or in movies and games applications when we need to synthesize the voices of famous people who are not alive or have lost their voice characteristics. Therefore, the general motivation of this research is to reduce over-smoothness in HMMSS under limited data conditions.

Voice conversion (VC) is a task to convert some features of the source speech to those of the target speech, while other features of the source speech are kept unchanged [10]. VCs usually require a small amount of target data to synthesize multiple target voices. The most popular VCs are statistical ones [10]. Since statistical VCs (and statistical SS in general) have to cope with the over-smoothness problem, some non-statistical methods such as exemplar-based [11], neural-network-based [12], and frequency-wrapping-based [8, 13] have been proposed recently. Neural-network-based VC can reduce over-smoothness but the conversion process still degrades speech quality [8], which can be attributed to the unconstrained nature of neural networks. Frequency-wrapping-based VC can also reduce over-smoothness but frequency wrapping cannot cope with voicing or nasality [13]. Up to now, the most successful VCs in overcoming over-smoothness appear to be methods using bilinear models such as the exemplar-based method of Wu [11]. These methods can synthesize target voices without over-smoothness with a small amount of the original target data.

Borrowing from the idea of using bilinear models in VC, a few hybrid TTSs using bilinear models have recently been proposed to improve the naturalness of HMMSS [14, 15].

In [14], the authors use a bilinear model called modified restricted temporal decomposition (MRTD) [16] to decompose speech synthesized by HMMSS into event targets closely related to the naturalness of speech and event functions closely related to speech intelligibility. Event targets of spectral features are then modified to close with those of the original speech. The experimental results show that this method can improve the

naturalness of synthesized speech. However, this method still has the following drawbacks. The first one is the run-time searching algorithm in the whole target space for target selection (target modification) that may cause a high computation cost for real-time synthesizing. The second one is the large reconstruction errors of bilinear model MRTD [16] may cause artifacts in synthesized speech.

In [15], the authors used another bilinear model using non-negative matrix factorization (NMF), the most successful bilinear model in VC as shown in [11], instead of MRTD to decompose speech synthesized by HMMSS into two component factors, one closely related to the naturalness of speech and the other closely related to the intelligibility of speech. While the intelligibility factors are kept, the spectral naturalness factor of speech synthesized by HMMSS is modified by minimizing the total squared error over reference actual speech data. This modification is based on the supposition that we have the reference actual speech data for every target utterance. This supposition is usually impractical.

The method proposed in [15] is speaker independent. Therefore, data from multiple speakers was used for training. The ideal of proposing a speaker-independent method is interesting. However, the quality of speaker-independent methods is usually lower than that of speaker-dependent methods where the focus is on improving the naturalness of synthesized speech of each specific speaker.

The authors in [15] also proposed to use the modulation spectrum of spectral features to process over-smoothness of spectral features in both time and frequency domains. However, over-smoothness in spectral features mostly occurs in the time domain [4]. In addition, using the modulation spectrum increases the reconstruction errors in synthesized speech.

In this paper, a hybrid TTS between HMMSS and bilinear model NMF has been proposed to reduce temporal over-smoothness of HMMSS with significant improvements compared with the method in [15].

This paper is organized as follows: the proposed hybrid between HMMSS and exemplar-based VC using NMF is shown in Section 2. The experimental evaluations are given in Section 3. Finally, conclusions are in Section 4.

2 Proposed hybrid TTS between HMMSS and exemplar-based voice conversion

2.1 Using non-negative matrix factorization for voice conversion

NMF is used to implement exemplar-based VC [11] and is also used in the hybrid TTS proposed in this paper between HMMSS and exemplar-based VC. The core idea of the NMF method is to represent a magnitude spectrum as a linear combination of a set of basis spectra (called as "speech atoms") as follows:

$$x = \sum_{t=1}^{T} a_t^{(X)}.h_t = A^{(X)}.h \tag{1}$$

where $x \in R^{p \times 1}$ represents the spectrum of one frame, T is the total number of speech atoms, $A^{(X)} = \left[a_1^{(X)}, a_2^{(X)}, ..., a_T^{(X)}\right] \in R^{p \times T}$ is the dictionary of speech atoms built from training source speech, $a_t^{(X)}$ is the t^{th} speech atom which has the same dimension as x, $h = [h_1, h_2, ..., h_T] \in R^{T \times 1}$ is the non-negative weight or activation vector and h_t is the activation of the t^{th} speech atom.

Therefore, the spectrogram of each source utterance can be represented as:

$$X = A^{(X)}.H \tag{2}$$

where $X \in R^{p \times M}$ is the source spectrogram, and $H \in R^{T \times M}$ is the activation matrix.

In order to generate converted speech spectrogram, the aligned source and target dictionaries are assumed to share the same activation matrix. Finally, the converted spectrogram is represented as:

$$Y = A^{(Y)}.H \tag{3}$$

where $Y \in R^{q \times M}$ is the converted spectrogram, and $A^{(Y)} \in R^{q \times T}$ is the dictionary of the target speech atoms from target training data.

2.2 Exemplar-based voice conversion

STRAIGHT [17] is used as a tool to extract speech features and to synthesize speech. Mel Frequency Cepstral Coefficients (MFCC) are obtained by using Mel-cepstral analysis on the STRAIGHT spectrum that is used to align two parallel utterances by dynamic time warping (DTW).

The VC has two separate stages: training stage and conversion stage.

In the training stage, the parallel source and target dictionaries are constructed as shown in Fig. 1. Given one pair of parallel utterances from source and target, the following process is employed to construct the dictionaries:

1) Extract STRAIGHT spectrum from both source and target speech signals;
2) Apply Mel-cepstral analysis to obtain MFCCs;
3) Perform dynamic time warping on the source and target MFCC sequences to align the speech to obtain the source-target frame pairs;
4) Apply the alignment information to the source and the target spectrums.

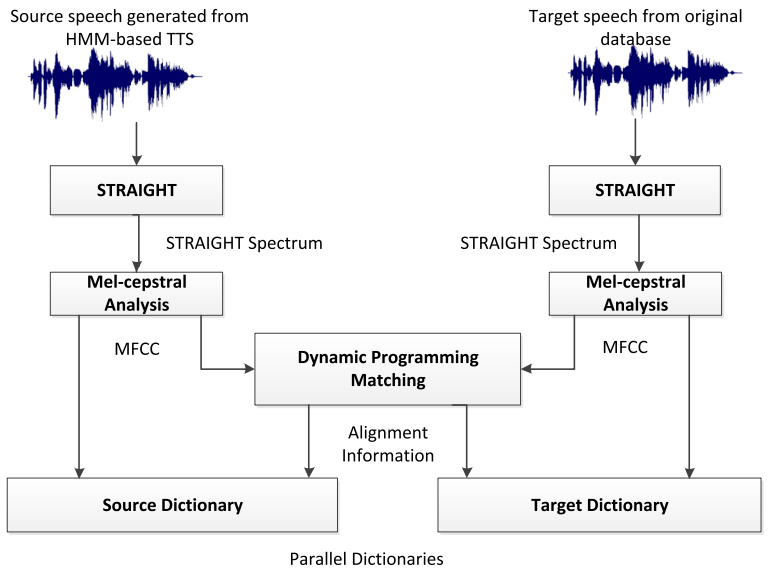

Fig. 1 Construction of source and target dictionaries for each utterance

The above four steps are applied for all the parallel training utterances. All the spectrum pairs (column vectors in source and target dictionaries) are used as speech atoms.

The conversion stage includes three tasks:

1) Extract source spectrum using STRAIGHT;
2) Estimate the activation matrix from Eq. (2);
3) Utilize the activation matrix and the target dictionary to generate the converted spectrum using Eq. (3), as shown in Fig. 2.

For each testing source speech atom, for each frame, the closest $a^{(X)}$ is searched in $A^{(X)}$, and then the correspondent target $a^{(Y)}$ is found by looking up the parallel dictionary $(A^{(X)}, A^{(Y)})$ built in the training stage.

2.3 Hybrid TTS between HMMSS and exemplar-based voice conversion

Figure 3 shows the flowchart of the proposed hybrid TTS between HMMSS and exemplar-based VC. All training sentences used in building the voice are synthesized where phoneme durations are also generated in the form of output label files. The pairs of HMMSS outputs and the corresponding original speech database are used in VC training to construct the source and target dictionaries for each utterance. In the conversion stage, any given sentence is first synthesized using the HMMSS. Then, exemplar-based VC is applied using the parallel dictionaries to generate the improved synthesized waveforms.

Comparing with the method in [14], the proposed method removes the runtime target searching in the whole target space by using a parallel dictionary previously built at the training stage. The proposed method

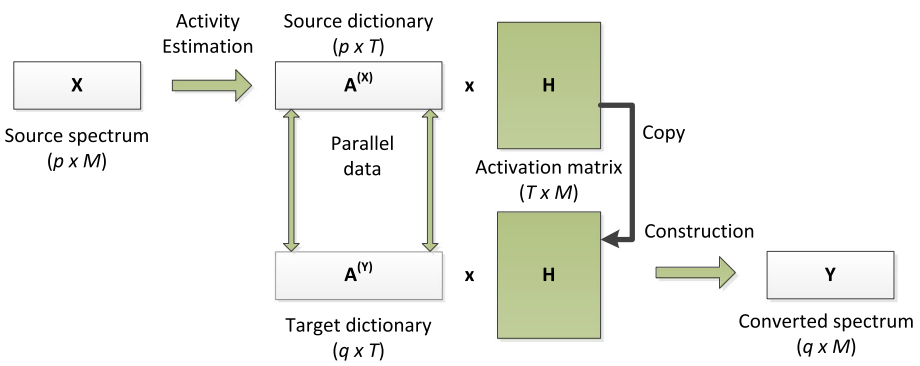

Fig. 2 Conversion stage

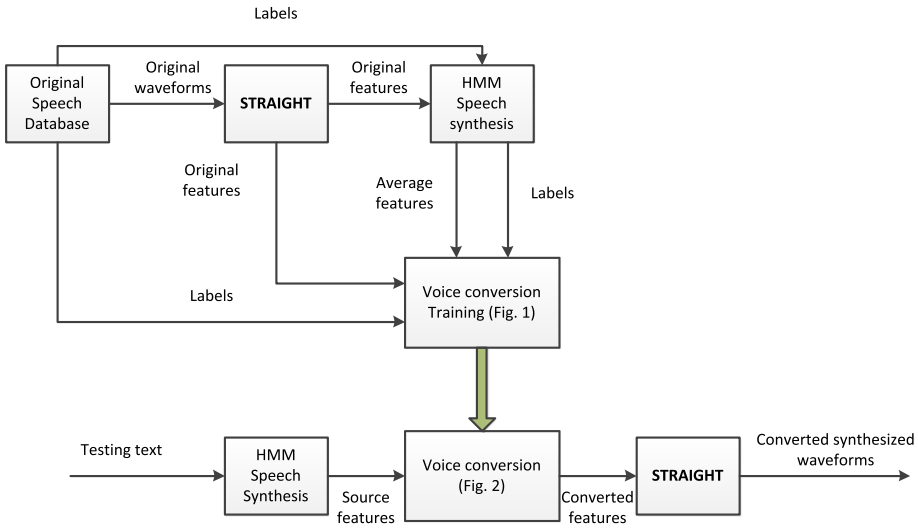

Fig. 3 Hybrid TTS between HMMSS and exemplar-based voice conversion

also removes the artifacts caused by large reconstruction errors of MRTD.

Compared with the method in [15], the proposed method in this paper has some significant differences. Firstly, in our proposed method, speech synthesized by HMMSS is converted to the natural speech by using a large parallel dictionary previously built in the training stage without using any actual reference speech in the conversion stage. Secondly, this parallel dictionary is built for each specific speaker in order to improve the over-smoothness in synthesized speech for each specific speaker only. Thirdly, MFCCs are used at the spectral feature instead of using the modulation spectrum to focus on improving only temporal over-smoothness in synthesized speech.

3 Experimental evaluations

3.1 Experimental conditions

The data corpus CSTR VCTK [18] was used for experimental evaluations. The corpus contains 108 English speakers with various accents. Each speaker spoke approximately 400 sentences recorded at 96 kHz sampling rate. For these experiments, the speech data was down-sampled to 16 kHz.

In this paper, we chose speech data of 400 sentences from a single female speaker for training HMMSS. In the VC stage, all 400 sentences were used for training and 15 sentences that are part of these 400 sentences were used for testing.

To implement a HMMSS, we chose Festival TTS framework http://www.cstr.ed.ac.uk/projects/festival/.

Acoustic features including 513 dimensional STRAIGHT spectrum, 24 coefficient MFCC, F0 and aperiodicity band energies were extracted at a 5 ms shift using STRAIGHT [14]. A hidden semi-Markov model was used containing static, delta and delta-delta values, with one stream for the spectrum, three streams for F0, and one for the band-limited aperiodicity.

3.2 Objective measures

Two objective measures were used. Those are the average standard deviation of spectral magnitude and the Mel-cepstral distortion.

The average standard deviation of spectral magnitude (in dB) was used to objectively measure the smoothness of synthesized speech. A lower average standard deviation indicates more over-smoothness.

The Mel-cepstral distortion is calculated as follows [10].

$$\mathrm{MCD} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}\left(mfcc_d^t - \hat{mfcc}_d^t\right)^2} \tag{4}$$

where $mfcc_d^t$, $\hat{mfcc}_d^t$ are the $d^{\text{-th}}$ coefficients of the source and target Mel-cepstral coefficients, respectively.

MCD is calculated between a synthesized frame and the corresponding original frame. The frame alignment is obtained by using dynamic time wrapping between parallel source and target sentences. A lower MCD indicates smaller distortion (usually less over-smooth).

For objective tests, fifteen test sentences were synthesized using four methods:

(1) HMMSS: Baseline HMM-based synthesis without GV,
(2) HMMSS + GV: HMMSS with GV,
(3) HMMSS + VC: HMMSS combined with VC,
(4) HMMSS + GV + VC: HMMSS + GV combined with VC (*on both VC training and testing stages*).

The objective evaluation results are shown in Table 1. These results indicate that:

- Speech synthesized by HMMSS is most over-smooth;
- Both HMMSS + GV and HMMSS + VC can efficiently reduce the over-smoothness compared with HMMSS. However, HMMSS + VC is less over-smooth (more rough) and closer to the original speech;
- HMMSS + GV + VC reduce the over-smoothness compared with HMMSS + GV but do not reduce the over-smoothness compared with HMMSS + VC.

These generally numerical results are consistent with the generally visual results shown in Fig. 4.

Table 1 Objective evaluation results

	Average standard deviation of spectral magnitude (dB)	Mel-cepstral distortion (MCD) (dB)
HMMSS	4.26	6.38
HMMSS + GV	7.43	4.76
HMMSS + VC	7.96	4.31
HMMSS + GV + VC	7.55	4.62

3.3 Subjective measures

In our subjective evaluations, we want to focus on evaluating the smoothness instead of the general quality because the motivation of this research is to reduce the over-smoothness. Notice that the smoothness is closely related to the general quality but they are not the same. Reducing the over-smoothness does not always improve the general quality. However, since over-smoothness has been the biggest issue over several years in HMMSS, we think the work to reduce over-smoothness will lead to significant improvements of the general quality in the near future.

To help the subjects clearly perceive what the smoothness is, we trained the subjects by listening to the original speech with ideal smoothness and speech synthesized by HMMSS with over-smoothness many times to distinguish the ideal smoothness and the over-smoothness.

The subjective measure is an AB preference test in the form of pairwise comparison between speech sentence pairs synthesized by the two systems: v1 and v2. The preference test demands a forced choice what kind of synthesized speech closer to ideal smoothness (and the other will be closer to over-smoothness) between each paired test stimuli. The order of presentation of the sentences is randomized.

Fifteen test sentences were synthesized using four methods HMMSS, HMMSS + GV, HMMSS + VC, and HMMSS + GV + VC.

Ten native speakers participated in the subjective tests with pairs of the following combinations:

- HMMSS (v1) and HMMSS + GV (v2);
- HMMSS (v1) and HMMSS + VC (v2);
- HMMSS + GV (v1) and HMMSS + GV + VC (v2);
- HMMSS + VC (v1) and HMMSS + GV + VC (v2);
- HMMSS + GV (v1) and HMMSS + VC (v2).

Table 2 shows the results of the subjective test for four pair choice. These results indicate that:

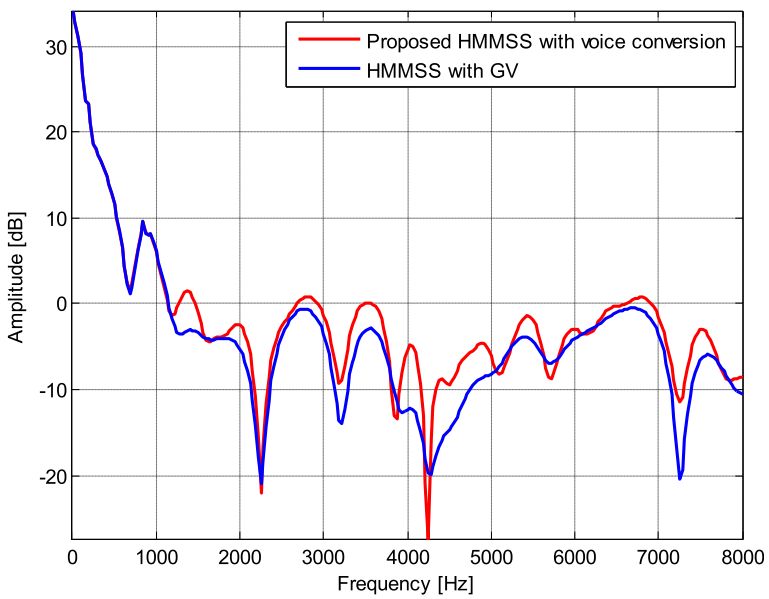

Fig. 4 Over-smoothness reduction for spectral envelope

Table 2 Average preference for pairwise comparison

Pairs: v1 and v2	v1%	v2%
1. HMMSS and HMMSS + GV	36.67	63.33
2. HMMSS and HMMSS + VC	31.33	68.67
3. HMMSS + GV and HMMSS + GV + VC	50.67	49.33
4. HMMSS + VC and HMMSS + GV + VC	51.33	48.67
5 HMMSS + GV and HMMSS + VC	46.67	53.33

- In pair 1, the preference for HMMSS + GV over HMMSS is clear with 63.33%.
- In pair 2, the preference for HMMSS + VC over HMMSS is clear with 68.67%.
- In pairs 3 and 4, there is no clear preference of HMMSS + GV + VC over HMM + GV or HMM + VC. Therefore, there are no clear reasons to use GV and the VC together.
- In pair 5, the preference for HMMSS + VC over HMMSS + GV is 53.33%.

Similar to the objective evaluation results presented in the previous section, the subjective evaluation results show that HMMSS + VC is the most similar to the original voice in term of smoothness. However, while all objective results are very clear, the subjective preference for HMMSS + VC over HMMSS + GV is not very clear.

4 Conclusions

Reducing over-smoothness in HMMSS is important in practical uses. In this paper, a hybrid synthesis between HMMSS and the exemplar-based voice conversion using NMF has been proposed to reduce over-smoothness of HMMSS. Results of both objective and subjective evaluations show that speech synthesized by the proposed method (HMMSS + VC) is the most similar to the original voice in terms of smoothness. Although the subjective preference for the proposed method over HMMSS + GV is not very clear, these results contribute to the research on removing over-smoothness in HMMSS, which is the biggest issue over several years in HMMSS.

Acknowledgements
This work was supported by the Ministry of Education and Training of Vietnam (project B2016-TNA-27).

5 Publisher's Note

Author details
[1]Duy Tan University, Da Nang, Vietnam. [2]Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam.

Received: 26 December 2016 Accepted: 19 June 2017
Published online: 24 June 2017

References

1. J Yamagishi et al., *A training method of average voice model for HMM-based speech synthesis. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* 86.8, 2003, pp. 1956–1963
2. K Tokuda, Z Heiga, AW Black, *An HMM-based speech synthesis system applied to English. IEEE Speech Synthesis Workshop*, 2002
3. K Tokuda et al., *Speech parameter generation algorithms for HMM-based speech synthesis. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP'00,* Vol. 3. IEEE, 2000
4. T Tomoki, T Keiichi, *A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. IEICE Transactions on Information and Systems* 90.5, 2007, pp. 816–824
5. M Zhang et al., *Improving HMM Based speech synthesis by reducing over-smoothing problems. 6th International Symposium on Chinese Spoken Language Processing, ISCSLP'08*, IEEE, 2008
6. G Beller, O Nicolas, R Xavier, *Articulation degree as a prosodic dimension of expressive speech. Fourth International Conference on Speech Prosody*, 2008
7. S Lee, B Erik, N Shrikanth, *An exploratory study of emotional speech production using functional data analysis techniques. 7th International Seminar on Speech Production*, 2006
8. Y Agiomyrgiannakis, R Zoi, *Voice morphing that improves TTS quality using an optimal dynamic frequency warping-and-weighting transform. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016
9. Y Qian, FK Soong, Y Zhi-Jie, *A unified trajectory tiling approach to high quality speech rendering. IEEE Transactions on Audio, Speech, and Language Processing* 21.2, 2013, pp. 280–290
10. T Toda, AW Black, T Keiichi, *Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Transactions on Audio, Speech, and Language Processing* 15.8, 2007, pp. 2222–2235
11. W Zhizheng, C Eng Siong, L Haizhou, Exemplar-based voice conversion using joint nonnegative matrix factorization. Multimedia Tools and Applications **74**(22), 9943–9958 (2015). Springer
12. L-H Chen et al., *DNN-based stochastic postfilter for HMM-based speech synthesis. INTERSPEECH*, 2014
13. E Godoy, R Olivier, C Thierry, *Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora. IEEE Transactions on Audio, Speech, and Language Processing* 20.4, 2012, pp. 1313–1323
14. Y Jiao et al., *Improving voice quality of HMM-based speech synthesis using voice conversion method*, 2014
15. AT Dinh, M Akagi, *Quality improvement of HMM-based synthesized speech based on decomposition of naturalness and intelligibility using non-negative matrix factorization. O-COCOSDA* (, Bali, Indonesia, 2016), pp. 62–67
16. C Nguyen Phu, T Ochi, A Masato, *Modified restricted temporal decomposition and its application to low rate speech coding. IEICE Transactions on Information and Systems* 86.3, 2003, pp. 397–405
17. H Kawahara, *Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP-97,* Vol. 2. IEEE, 1997
18. C Veaux, Y Junichi, MD Kirsten, *CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit*, 2016