HMM-based Speech Synthesis with Multiple Individual Voices using Exemplar-based Voice Conversion

Trung-Nghia Phung,

Thai Nguyen University, Thai Nguyen, Vietnam

Abstract

Traditional text-to-speech (TTS) systems can synthesize only single individual voice. When we need to synthesize other individual voices, we have to train the system again with the new voices. The training process normally requires a huge amount of data that is usually available with a few specific voices existed in the database.

The state of the art TTS using Hidden Markov Model (HMM), called as HMM-based TTS, can synthesize speech with various voice personality characteristics by using speaker adaptation methods. However, both of the voices synthesized and adapted by HMM-based TTS are “over-smooth”. When these voices are over-smooth, the detail structures clearly linked to speaker individuality may be missing. We can also synthesize multiple voices by using some voice conversion (VC) methods combined with HMM-based TTS. However, current voice conversions still cannot synthesize target speech while keeping the detail information related to speaker individuality of the target voice and just using limited amount data of target voices. In this paper, we proposed to use exemplar-based voice conversion combined with HMM-based TTS to synthesize multiple high-quality individual voices with a few amount of target data. The evaluation results using the English data corpus CSTR confirmed the advantages of the proposed method.

Key words:

HMM-based speech synthesis, Speaker adaptation, Exemplar-based voice conversion, Non-negative matrix factorization, Speaker individuality

1. Introduction

Among many kinds of TTS systems that have been proposed, the state-of-the-art TTS is HMM-based [1, 2]. In this approach, spectral and prosodic features of speech are modeled and generated in an unified statistical framework using HMMs. HMM-based TTS has many advantages that have been shown in the literature, such as the high intelligibility of synthesized speech, the small footprint, the low computational load [1]. However, traditional HMM-based TTS can only synthesize single voice that is fully trained already.

Several practical applications require multiple synthesized voices while the requirement of having huge amounts data of original target voices for training is usually not available. To solve this problem, two approaches have been proposed. The first approach is using HMM-based speaker adaptation methods [3]. This approach can adapt synthesized speech to target voices with a few amounts of target data. In both HMM-based synthesis and adaptation methods, the structures of the estimated spectrum correspond 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. This characteristic in speech synthesized and adapted by HMM-based TTS can be considered “too medial", or “over-smooth”. When synthesized and adapted speech is over-smooth, it sounds “muffed" and the detail structure in the original speech clearly linked to speaker individuality may be missing. Moreover, over-smooth speech with too-slow transitions may also affect to produce important information on personality. As a consequence, the personality perception in speech synthesized and adapted by HMM-based TTS is not fully response to practical applications.

Another approach can be used to synthesize multiple target voices is to combine HMM-based TTS with a VC method as a post-processing step. Several VC methods can convert a source voice to various target voices using limited amount data of target voices. State-of-the-art VC methods use Gaussian Mixture Model (GMM) [4]. However, both GMM and HMM approximations are based on the uses of mean vectors. Therefore, state-of-the-art VC still cannot synthesize target speech while keeping the detail information related to speaker individuality of the target voice.

In this paper, we proposed to use the exemplar-based VC using non-negative matrix factorization [5] combined with HMM-based TTS to synthesize multiple voices that can keep the detail information related to speaker individuality. The experimental results with English data corpus CSTR VCTK [6] show that the proposed method outperforms both of HMM-based adaptation method and the method combined from HMM-based TTS and GMM-based VC.

Manuscript received May 5, 2017
Manuscript revised May 20, 2017

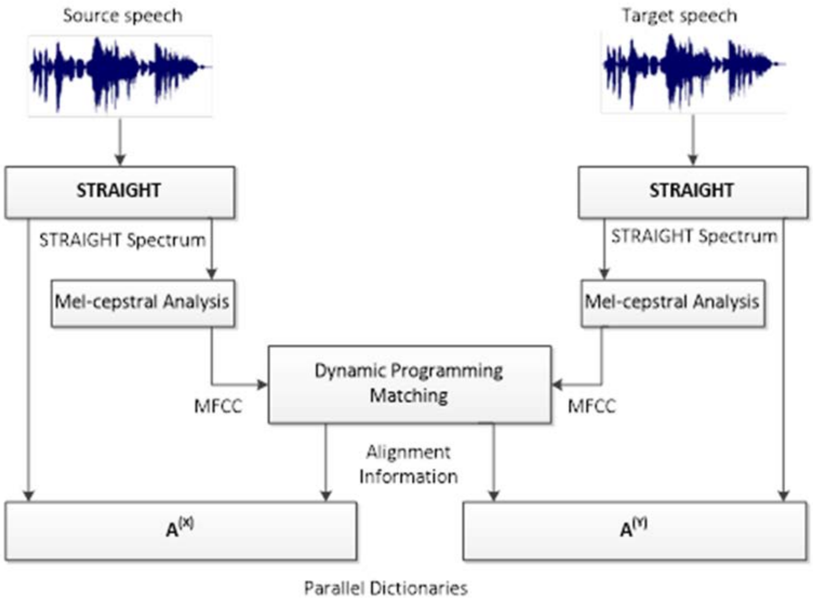

Fig. 1 Construction of source and target dictionaries for each utterance

2. Speaker individual information in speech signal

Speech information can be categorized into linguistic and non-linguistic information. Although information on speaker identity exists on the linguistic level, non-linguistic information is more important to speaker individuality [7]. The non-linguistic information affecting speaker individuality can be divided to sociological and physiological factors. However, effect of physiological factors on the acoustic speech signal is more clearly.

The most important non-linguistic factors strongly affecting to the speaker individuality are closely related to physical characteristics of speaker vocal tract and these characteristics are usually represented by spectral features such as formant or cepstrum [7].

The degree of articulation (DoA) also provides information on the personality [8]. DoA is characterized by modifications of the speech rate and of the spectral dynamics and these dynamics are actually the rate of vocal tract temporal change.

Over-smooth speech spectral features generated by statistical methods such as HMM and GMM with too-slow transitions may affect to produce the appropriate DoA and the important information on personality may be loss.

3. The proposed method

3.1 Using non-negative matrix factorization for individual VC

The core idea of NMF method is to represent a spectral vector as a linear combination of a set of basis vector (called as speech atoms) as follows [5]:

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

where $x \in R^{p\times 1}$ represents the spectrum of one frame, T is the total number of speech atoms, $A^{(X)} = [a_1^{(X)}, a_2^{(X)}, ..., a_T^{(X)}] \in R^{p\times T}$ 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:

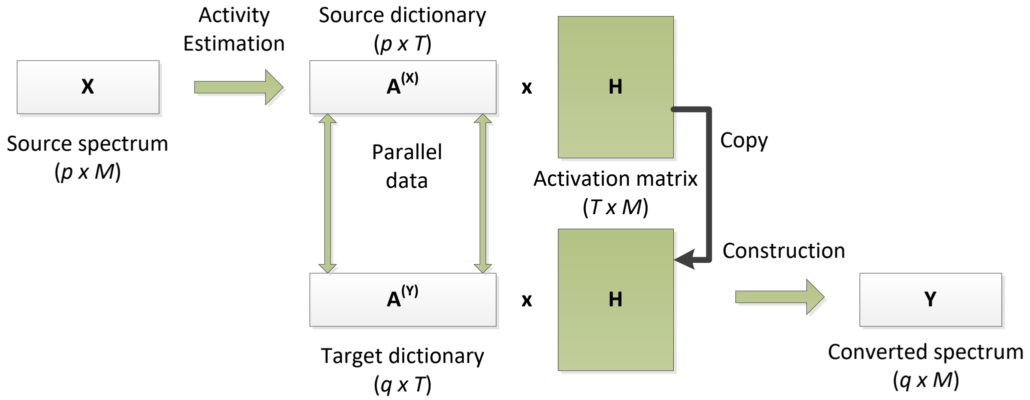

Fig. 2 Conversion Stage

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

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:

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

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.

3.2 Exemplar-based individual VC

STRAIGHT [9] is used as a tool to extract speech features and to synthesize speech while Mel Frequency Cepstral Coefficients (MFCC) obtained by using Mel-cepstral analysis on the STRAIGHT spectrum is used to align two parallel utterances by the dynamic time warping (DTW).

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

In 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 dictionary:

1) Extract STRAIGHT spectrum from both source and target speech signal;

2) Apply Mel-cepstral analysis to obtain MFCCs;

3) Perform dynamic time warping on the source and target MFCC sequence to align the speech to obtain source-target frame pairs;

4) Apply the alignment information to the source and target spectrum. 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: extract source spectrum using STRAIGHT; estimate activation matrix from Eq. (2); 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 in one 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 training stage.

3.3 Combination between HMM-based TTS and exemplar-based VC

Fig 3 shows the flowchart of the proposed system combined from HMM-based TTS and exemplar-based VC. All training sentences used in building the voice are synthesized where realized phoneme durations are also generated in the form of output label files. The pairs of HMM-based TTS 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 HMM-based TTS. Then, exemplar-based VC is applied using the parallel dictionaries to generate the improved synthesized waveforms.

4. Experimental Evaluations

4.1. Experimental conditions

The data corpus CSTR VCTK [6] was used for experimental evaluations. The corpus contains 108 English speakers approximately 400 sentences for each speaker on average recorded at 96 kHz sampling rate.

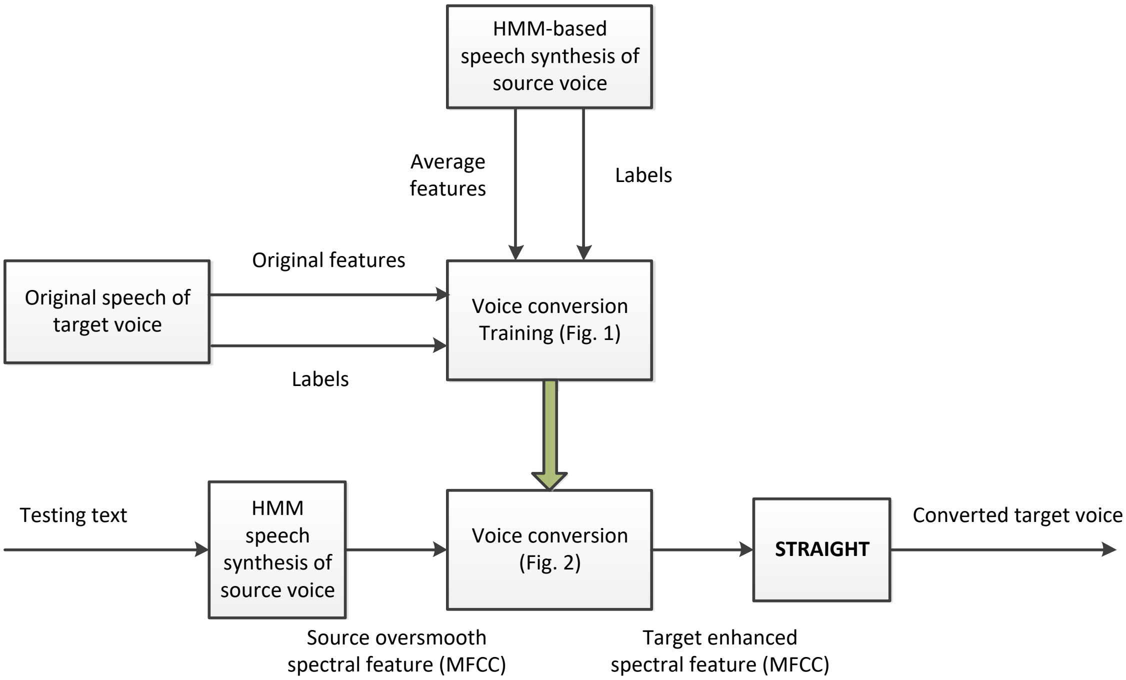
HMM-based speech synthesis of source voice
Average features
Labels
Original speech of target voice
Original features
Labels
Voice conversion Training (Fig. 1)
Testing text
HMM speech synthesis of source voice
Voice conversion (Fig. 2)
STRAIGHT
Converted target voice
Source oversmooth spectral feature (MFCC)
Target enhanced spectral feature (MFCC)

Fig. 3 Combination between HMM-based TTS and exemplar-based VC

The speech data was first downsampled to 16 kHz. Then, we chose speech data of 400 sentences from a female source speaker for training HMM-based TTS and we chose speech data of 100 sentences from both female source and target speakers for HMM-based speaker adaptation [3] and for training the VCs including the GMM-based [4] and the exemplar-based VC.

To implement a HMM-based TTS with speaker adaptation, we chose Festival TTS framework [10].

In evaluations, 10 same sentences of the original target voice, of the voice adapted from HMM-based TTS to the target voice, and of the voice converted to the target voice using the GMM-based VC and the proposed combined system were used for testing.

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

4.2. Objective measures

Mel-cepstral distortion was used as objective measures. The mel-cepstral distortion is calcuate as follows [4].

$$MCD = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}(mfcc_d^t - \hat{m}fcc_d^t)^2} \qquad (4)$$

where $mfcc_d^t$, $\hat{m}fcc_d^t$ are the d-th coefficients of the source and target mel-cepstral coefficients, respectively.

MCD is calculated between an original target frame and the corresponding frame adapted by HMM, converted by GMM and by the proposed combined system. The frame alignment is obtained by using dynamic time wrapping between parallel source and target sentences. A lower of MCD indicates the better adaptation or conversion methods. The objective evaluation results are shown in Table 1. These results indicate that the speech converted by using the exemplar-based VC is closest with the original target speech.

Table 1: Objective evaluations: Mel-cepstral distortion (MCD) between the original target speech and each adapted or converted speech

	MCD (dB)
Speech adapted by HMM	5.67
Speech converted by GMM-based VC	5.09
Speech converted by exemplar-based VC	3.24

4.3. Subjective measures

In the subjective test of speaker individuality, an ABX test [4] was conducted. A means the source speaker, B means the target speaker, and X means the converted or adapted speech. Ten listeners with normal hearing were asked to select if X was closer to A or B, and provide the score from 1 to 5 according to his/her perception of speaker individuality when comparing. The score of 1 means that the adapted / converted speech is very similar to the source speaker, and the score of 5 means that the adapted / converted speech is very similar to the target speaker. Results of the ABX test are shown in Table 2. This result shows that the speech individuality of converted speech of our proposed method is the most similar to the target speaker among the methods.

Table 2: ABX results for HMM-based speaker adaptation (1), GMM-based VC (2), and Exemplar-based VC (3)

ABX scores		
(1)	(2)	(3)
3.2	3.6	4.1

5. Conclusions

Both of the voices synthesized and adapted by HMM-based TTS or converted by GMM-based VC are “over-smooth”. When these voices are over-smooth, the detail structures clearly linked to speaker individuality may be missing. Then, HMM-based and GMM-based methods cannot synthesize target speech while keeping the detail information related to speaker individuality of the target voice. In this paper, we proposed to use exemplar-based VC combined with HMM-based TTS to synthesize multiple high-quality individual voices with a few amount of target data. The subjective and objective evaluation results confirmed the efficiency of the proposed method.

Acknowledgments

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

References

[1] Tomoki, Toda, and Keiichi Tokuda. "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis." IEICE TRANSACTIONS on Information and Systems 90.5 (2007): 816-824.

[2] Tokuda, Keiichi, Heiga Zen, and Alan W. Black. "An HMM-based speech synthesis system applied to English." IEEE Speech Synthesis Workshop. 2002.

[3] Yamagishi, Junichi, and Takao Kobayashi. "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training." IEICE TRANSACTIONS on Information and Systems 90.2 (2007): 533-543.

[4] Toda, Tomoki, Alan W. Black, and Keiichi Tokuda. "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory." IEEE Transactions on Audio, Speech, and Language Processing 15.8 (2007): 2222-2235.

[5] Wu, Zhizheng, et al. "Exemplar-based voice conversion using non-negative spectrogram deconvolution." SSW. 2013.

[6] Veaux, Christophe, Junichi Yamagishi, and Kirsten MacDonald. "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit." (2016).

[7] Lavner, Yizhar, Judith Rosenhouse, and Isak Gath. "The prototype model in speaker identification by human listeners." International Journal of Speech Technology 4.1 (2001): 63-74.

[8] Beller, Grégory, Nicolas Obin, and Xavier Rodet. "Articulation degree as a prosodic dimension of expressive speech." in Fourth International Conference on Speech Prosody. 2008.

[9] Kawahara, Hideki. "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited." IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2. IEEE, 1997.

[10] http://www.cstr.ed.ac.uk/projects/festival/