Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Multiple emotional voice conversion in Vietnamese HMM-based speech synthesis using non-negative matrix factorization

Trung-Nghia Phung *

Thai Nguyen University of Information and Communication Technology, Thai Nguyen 25000, Vietnam

ARTICLE INFO

Article history:
Received 16 May 2017
Received in revised form
23 June 2017
Accepted 23 June 2017

Keywords:
HMM-based speech synthesis
Voice adaption
Exemplar-based voice conversion
Non-negative matrix factorization
Emotional speech synthesis

ABSTRACT

Most of current text-to-speech (TTS) systems can synthesize only single voice with neutral emotion. If different emotional voices are required to be synthesized, the system has to be trained again with the new emotional voices. The training process normally requires a huge amount of emotional speech data that is usually impractical. The state of the art TTS using Hidden Markov Model (HMM), called as HMM-based TTS, can synthesize speech with various emotions by using speaker adaption methods. However, both of the emotional voices synthesized and adapted by HMM-based TTS are "over-smooth". When these voices are over-smooth, the detail structures clearly linked to speaker emotions 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 emotions of the target voice and just using limited amount data of target voices. In this paper, we proposed to use exemplar-based emotional voice conversion combined with HMM-based TTS to synthesize multiple high-quality emotional voices with a few amount of target data. The evaluation results using the Vietnamese emotional speech data corpus confirmed the merits of the proposed method.

1. Introduction

HMM-based is the state -of-the-art TTS up to now in which spectral and prosodic features of speech are modeled and generated by an unified statistical framework using HMMs (Tomoki and Tokuda, 2007; Tokuda et al., 2002). In the literature, HMM-based TTS has been shown several merits such as the high intelligibility of synthesized speech, the small footprint, the low computational load (Tomoki and Tokuda, 2007). However, conventional HMM-based TTSs are able to synthesize only single neutral voice that is fully trained already before instead of synthesizing any required emotional voices.

In many practical applications, TTS with multiple synthesized emotional voices is required while the requirement of having huge amounts data of emotional target voices for training is usually not available. Two approaches have been proposed to solve the above problem. The first approach is using HMM-based voice adaption methods (Takashi et al., 2009). In this approach, synthesized neutral speech is adapted to target emotional voices with a few amounts of emotional target data. However, in both HMM-based synthesis and voice adaption, 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. On the other hand, the spectrum estimated by HMMs is an average approximation of all corresponding speech spectra in the training database. Therefore, speech synthesized or adapted by HMM-based TTS is "too medial", or "over-smooth". When synthesized or adapted speech is over-smooth, it sounds "muffed" and the detail structure in the original speech clearly linked to speaker emotions may be missing. As a result, the emotion perception in speech synthesized and adapted by HMM-based TTS is far from being applied in different kinds of practical applications.

Using a VC method as a post-processing step for HMM-based TTS is another approach to synthesize multiple emotional target voices. Several VC methods can convert a source neutral voice to various target emotional voices using limited amount data of target emotional voices. State-of-the-art emotional VC methods use Gaussian Mixture Model (GMM) (Tomoki and Tokuda, 2007; Aihara et al., 2012). However, both GMM and HMM

* Corresponding Author.
Email Address: ptnghia@ictu.edu.vn (T. N. Phung)
https://doi.org/10.21833/ijaas.2017.08.001
2313-626X/

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 emotions of the target voice.

In this research, we proposed to use the exemplar-based VC using non-negative matrix factorization combined with HMM-based TTS to synthesize multiple emotional voices that can keep the detail information related to speaker emotions. The experimental results with Vietnamese speech corpus show that the proposed method improves the efficiency of emotional speech synthesis compared with using HMM-based adaption and using GMM-based VC combined with HMM-based TTS.

2. Emotions in speech signal

Speaker emotion information exists on both linguistic and non-linguistic levels. However, non-linguistic factors are closer to speaker emotions. The non-linguistic factors including physical characteristics of speaker vocal tract represented by spectral features strongly affect to the speaker emotions. Moreover, prosodic features such as pitch contour or fundamental frequency (F0) also affect to speaker emotions in the speech signals (Lavner et al., 2001; Chappell and Hansen, 1998).

Most of emotional voice adaption and voice conversion methods focus on spectral features only. Some other methods use simple statistical mean and variance scaling of F0 conversions (Tomoki and Tokuda, 2007; Chappell and Hansen, 1998; Gillett and King, 2003; Helander and Nurminen, 2007).

The degree of articulation (DoA) characterized by modifications of the speech rate and of the spectral dynamics also provides information on the emotions (Beller et al., 2008). Over-smoothness and too-slow transitions in both spectral and prosodic features generated by using statistical methods such as HMM and GMM may affect to produce the appropriate DoA to express important information repressing emotions.

3. The proposed method

3.1. Using non-negative matrix factorization for emotional voice conversion

The core idea of NMF method is to represent a speech feature (such as spectral or F0) as a linear combination of a set of basis vector (called as speech atoms) (Wu et al., 2013) as follows (Eq. 1):

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

where $x \in R^{P\times 1}$ represents the speech feature of one frame, T is the total number of speech atoms, $A^{(X)} = \left[a_1^{(X)}, a_2^{(X)}, \dots, a_T^{(X)}\right] \in R^{P\times T}$ 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, $h = [h_1, h_2, \dots, h_T] \in R^{T\times 1}$ is the non-negative weight or activation vector and h_t is the activation of the t^{th} speech atom.

Therefore, the speech feature of each source utterance can be represented as (Eq. 2):

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

where $X \in R^{P\times M}$ is the speech feature, and $H \in R^{T\times M}$ is the activation matrix.

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

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

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

3.2. Exemplar-based emotional voice conversion

STRAIGHT (Kawahara, 1997) 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 and F0 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 MFCC and F0. The above four steps are applied for all the parallel training utterances. All MFCC and F0 pairs (column vectors in source and target dictionaries) are used as speech atoms.

The conversion stage includes three tasks: extract source MFCC and F0 using STRAIGHT; estimate activation matrix from Eq. 2; utilize the activation matrix and the target dictionary to generate the converted MFCC 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

The proposed emotional speech synthesis combined from HMM-based TTS and exemplar-

based VC is represented in Fig. 3. Phoneme durations are generated in the form of output label files.

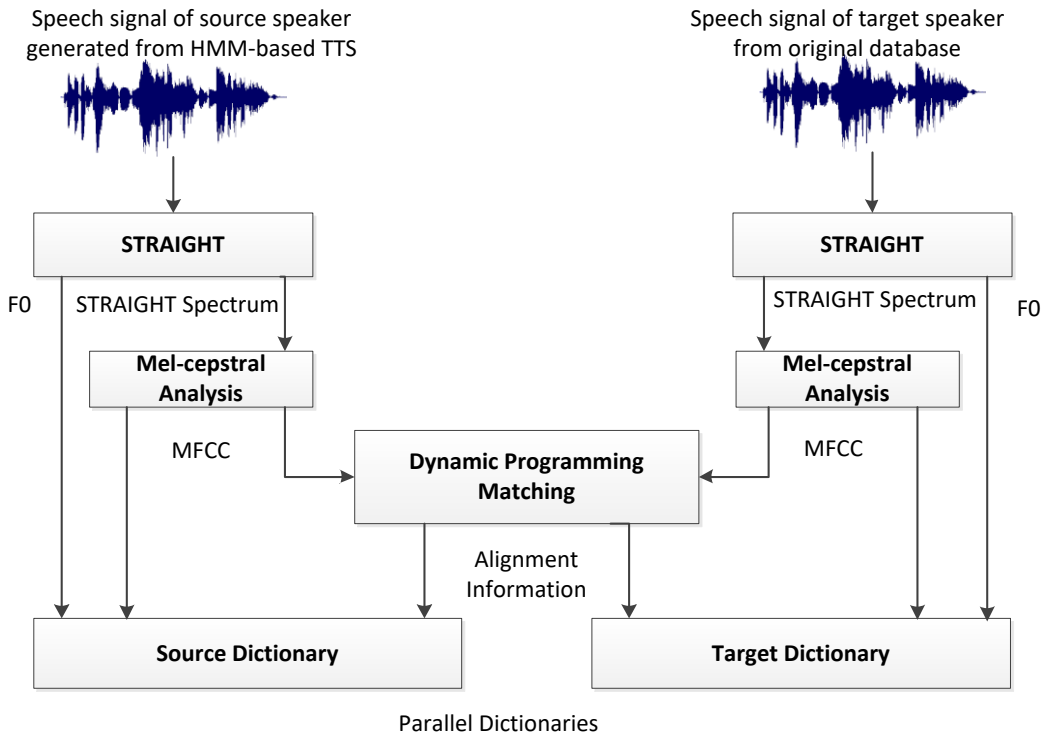

Fig. 1: Construction of source and target dictionaries for each utterance in training stage

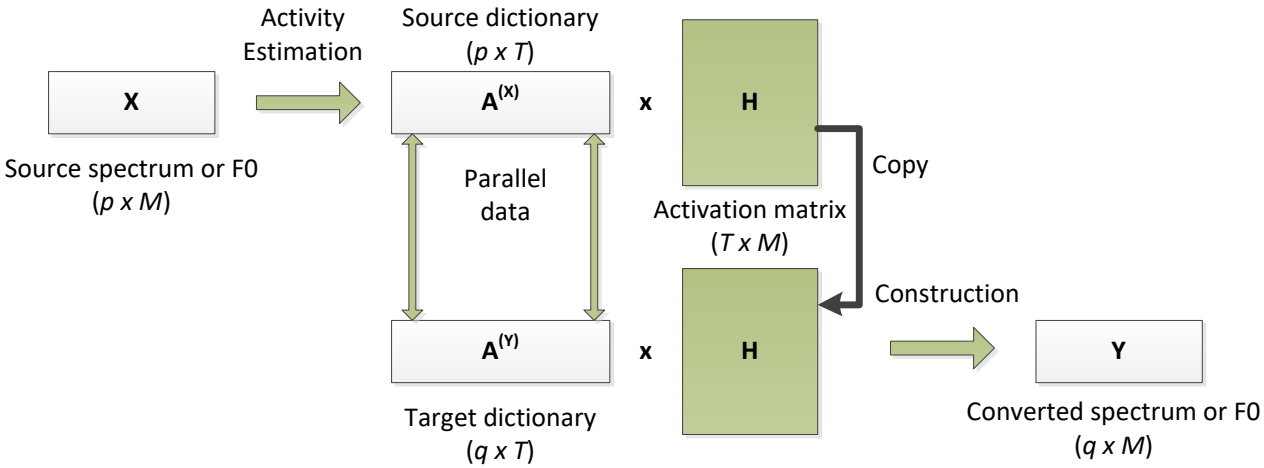

Fig. 2: Conversion stage

The pairs of HMM-based TTS outputs and the corresponding original emotional 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 synthesized speech with the target emotion.

4. Experimental evaluations

4.1. Data corpus

The Vietnamese speech corpus used for HMM-based TTS is DEMEN (Phung et al., 2012) including 567 sentences spoken by a single female speaker. The sampling frequency used in DEMEN database was 16000 Hz with 16 bit resolution. We extended the DEMEN data corpus to an emotional speech database with 19 utterances using six different emotions that were: happiness, cold anger, sadness, hot anger, and neutral.

4.2. Experimental conditions

500 utterances of the single female speaker extracted from Vietnamese DEMEN corpus was used for Vietnamese HMM-based TTS.

The Vietnamese HMM-based TTS was developed from a HMM-based TTS called as HTS in Zen et al. (2007) with modifications as in Phan et al. (2013).

For Vietnamese emotional voice adaption and conversion, we used 15 utterances of emotion “hot anger” for training, and corresponding 4 utterances for testing.

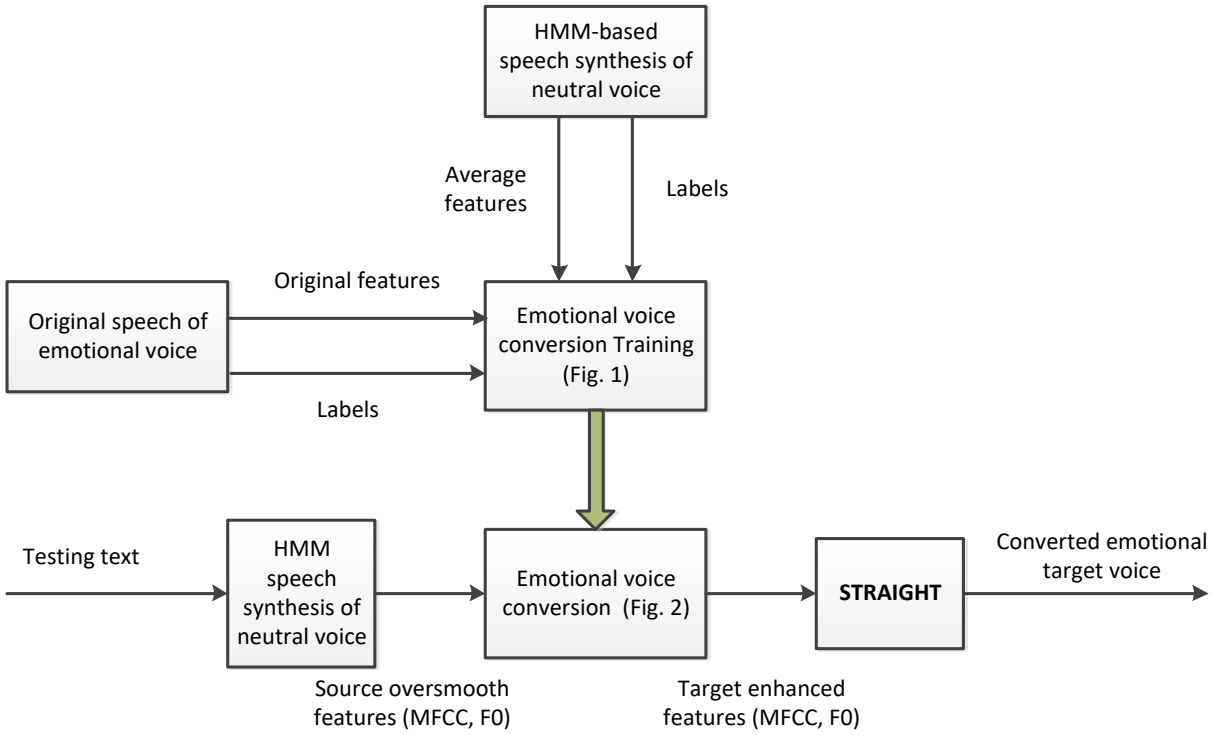

Fig. 3: Combination between HMM-based TTS and exemplar-based VC

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.

4.3. Objective measures

Mel-cepstral distortion was used as an objective spectral measure. The mel-cepstral distortion (MCD) is calculated as Eq. 4.

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

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

MCD is calculated between an original target emotional frame and the corresponding frame adapted from neutral voices by HMM, converted from neutral voices by GMM (Aihara et al., 2012) 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 spectral 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.

Root mean square error (RMSE) of F0 was used as an objective F0 measure (Eq. 5).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f0_i - f0_i')^2} \quad (5)$$

where $f0, f0'$ are the *i-th* value of the source and target F0, respectively.

RMSE of F0 is calculated between an original target emotional speech and the corresponding speech adapted from the neutral voice to the "hot anger" voice by HMM, linearly converted from the neutral voice to the "hot anger" voice by GMM and converted by the proposed combined system. A lower of RMSE indicates the better adaptation or conversion methods.

Table 2 shows RSME results. As shown in the table, the speech converted by using the exemplar-based VC is closest with the original target "hot anger" speech.

4.4. Subjective measures

In the subjective test of synthesized emotional speech, an ABX test was conducted. A means the original neutral source speech, B means the original "hot anger" target speech, and X means the converted or adapted speech.

Ten Vietnamese 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 emotions when comparing. The score of 1 means that the adapted / converted speech is very similar to the neutral speech (source emotion); and the score of 5 means that the adapted / converted speech is very similar to the "hot anger" speech (target emotion).

Results of the ABX test are shown in Table 3. This result shows that the speech emotion of converted speech of our proposed method is the most similar

to the target emotion among the methods. The results also show that the efficiency of all adaptation and conversion methods present in this paper is not really high. The possible reason may be that the size of Vietnamese emotional speech dataset is too small with only 19 Vietnamese utterances.

Table 1: Mel-cepstral distortion measures

	MCD (dB)
Between the original "hot anger" target speech and the corresponding speech adapted by HMM	8.62
Between the original "hot anger" target speech and the corresponding speech converted by GMM-based VC	7.24
Between the original "hot anger" target speech and the corresponding speech converted by exemplar-based VC	5.85

Table 2: F0 distortion measures

	RMSE(F0)
Between the original "hot anger" target speech and the corresponding speech adapted by HMM	61.7
Between the original "hot anger" target speech and the corresponding speech converted by GMM-based VC	45.3
Between the original "hot anger" target speech and the corresponding speech converted by exemplar-based VC	35.9

Table 3: ABX results for HMM-based voice adaptation (1); GMM-based emotional VC (2); and Exemplar-based emotional VC (3)

ABX scores		
(1)	(2)	(3)
2.2	2.6	3.2

5. Conclusion

Both of the voices 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 emotions may be missing. Therefore, HMM-based and GMM-based methods are difficult to synthesize target speech while keeping the detail information related to speaker emotions of the target voice. In this paper, we proposed to use exemplar-based VC combined with HMM-based TTS to synthesize multiple high-quality emotional voices with a few amount of target data. The subjective and objective evaluation results confirmed the advantages of the proposed method.

Acknowledgment

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

References

Aihara R, Takashima R, Takiguchi T, and Ariki Y (2012). GMM-based emotional voice conversion using spectrum and prosody features. American Journal of Signal Processing, 2(5): 134-138.

Beller G, Obin N, and Rodet X (2008). Articulation degree as a prosodic dimension of expressive speech. In the 4th International Conference on Speech Prosody, Campinas, Brazil. Available online at: citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.527.2960

Chappell DT and Hansen JH (1998). Speaker-specific pitch contour modeling and modification. In the IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Seattle, USA, 2: 885-888. https://doi.org/10.1109/ICASSP.1998.675407

Gillett B and King S (2003). Transforming F0 contours. In the 8th European Conference on Speech Communication and Technology, Geneva, Switzerland. Available online at: http://hdl.handle.net/1842/1078

Helander EE and Nurminen J (2007). A novel method for prosody prediction in voice conversion. In the IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Honolulu, USA: 509-512. https://doi.org/10.1109/ICASSP.2007.366961

Kawahara H (1997). Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited. In the IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, Munich, Germany, 2: 1303-1306. https://doi.org/10.1109/ICASSP.1997.596185

Lavner Y, Rosenhouse J, and Gath I (2001). The prototype model in speaker identification by human listeners. International Journal of Speech Technology, 4(1): 63-74.

Phan TS, Duong TC, Dinh AT, Vu TT, and Luong, CM (2013). Improvement of naturalness for an HMM-based Vietnamese speech synthesis using the prosodic information. In the IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future, IEEE, Hanoi, Vietnam: 276-281. https://doi.org/10.1109/RIVF.2013.6719907

Phung TN, Mai CL, and Akagi M (2012). A concatenative speech synthesis for monosyllabic languages with limited data. In the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, IEEE, Hollywood, USA: 1-10.

Takashi NOSE, Tachibana M, and Kobayashi T (2009). HMM-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation. IEICE Transactions on Information and Systems, 92(3): 489-497.

Tokuda K, Zen H, and Black AW (2002). An HMM-based speech synthesis system applied to English. In the IEEE Workshop on Speech Synthesis: 227-230. https://doi.org/10.1109/WSS.2002.1224415

Tomoki T and Tokuda K (2007). A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. IEICE Transactions on Information and Systems, 90(5): 816-824.

Wu Z, Virtanen T, Kinnunen T, Chng ES, and Li H (2013). Exemplar-based voice conversion using non-negative spectrogram deconvolution. In the 8th ISCA Speech Synthesis Workshop, Barcelona, Spain: 201-206.

Zen H, Nose T, Yamagishi J, Sako S, Masuko T, Black AW, and Tokuda K (2007). The HMM-based speech synthesis system (HTS) version 2.0. In the 6th ISCA Workshop on Speech Synthesis, Bonn, German: 294-299.