Quality Improvement of Vietnamese HMM-based Speech Synthesis System based on Decomposition of Naturalness and Intelligibility using Non-negative Matrix Factorization

Anh-Tuan Dinh[1], Thanh-Son Phan[2], and Masato Akagi[1]

[1] School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, Japan
{tuan.dinh,akagi}@jaist.ac.jp
[2] Faculty of Information Technology, Telecommunications University
101 Mai Xuan Thuong, Nha Trang, Khanh Hoa, Vietnam
ptson@tcu.edu.vn

Abstract Hidden Markov model (HMM)-based synthesized speech is intelligible but not natural especially under limited data condition because of over-smoothing of the speech spectra and F0 envelope. One solution is using voice conversion methods to convert over-smoothed speech parameters to natural ones. Although conventional conversion methods transform speech spectra and F0 envelope to natural ones to improve naturalness, they cause unexpected distortions in acceptable intelligibility of synthesized speech e.g. destroying tonal information. The aim of this study is to develop a method for improving naturalness without violating acceptable intelligibility by employing our novel asymmetric bilinear model (ABM) involving non-negative matrix factorization (NMF) to separate the naturalness and intelligibility of synthesized speech. Subjective evaluations carried out on Vietnamese data confirm that the achieved synthesis quality is higher than other methods under limited data condition. Moreover, proposed method is capable of modifying over-smoothed F0 envelope without destroying tonal information.

Keywords: Hidden Markov model, non-negative matrix factorization, naturalness-intelligibility decomposition

1 Introduction

Although Vietnamese is spoken by about 100 million people, there is no huge public speech-corpus with labelling for Vietnamese. In other word, Vietnamese is an under-resourced language.

Hidden Markove model (HMM)-based speech synthesis (HMMSS) is a state-of-the-art method due to its flexibility and compact footprint [1], [11], [12]. The HMM can model not only the statistical distribution of speech parameters but also their rate of change. As a result, synthesized speech is intelligible but not

natural due to statistical averaging or over-smoothing effect under limited data conditions. There have been several attempts to overcome the over-smoothing effect. Thus, it is a challenge to improve naturalness without violating intelligibility especially for Vietnamese synthesized speech.

One approach is using objective evaluations of this effect such as global variance (GV) [2], and modulation-spectrum [3], integrating them into the parameter generation phase to obtain better speech parameter values. Context-dependent models are usually trained for objective evaluations. However, under limited data conditions, there is not enough data to train the context-dependent models for all possible contexts. Another possible approach to reduce the gap between the spectra of natural and synthetic speech is to learn the acoustic differences directly from the data. If we have a parallel set of natural and synthesized speech, voice conversion techniques [4], [5] can be used as mapping from natural speech to synthetic speech. Since quality improvement is independent from synthesizers, we can improve the naturalness of current speech synthesizers. Thus, a voice-conversion approach is used to improve naturalness.

With the majority of previous voice-conversion approaches, all spectra and F0 are modified to improve naturalness. However, applying these approaches often negatively affect intelligibility e.g., destroying tonal information. This drastically affects intelligibility of tonal languages such as Vietnamese and Chinese. To improve naturalness without violating intelligibility, an asymmetric bilinear model (ABM) [6] was introduced to decompose naturalness and intelligibility. Popa et al. [7] used an ABM to decompose a speech parameter vector into speaker information and phonetic information using singular value decomposition (SVD). From this idea, a speech parameter vector y can be represented as a combination of the naturalness factor $\mathbf{A}$ and intelligibility factor $\mathbf{b}^c$ of an intelligibility class c:

$$\mathbf{y} = \mathbf{A}\mathbf{b}^c \tag{1}$$

In the representation of ABM, naturalness can be modified, whilst intelligibility can be preserved. One problem with applying an ABM is finding an efficient constraint to decompose naturalness and intelligibility from speech spectra and F0 envelope. Although an ABM using SVD is an excellent approach, SVD allows negative combinations of intelligibility and naturalness. Since combinations indicate unrealistic subtractions of intelligibility (or naturalness), negative combinations are unnecessary. To avoid subtractive combinations, we propose a method that uses our novel ABM involving negative matrix factorization (NMF).

2 Overview of applying ABM involving NMF for improving speech spectra

In the section, we describe the process of applying an ABM for improving synthesized speech-spectra. This speech spectra is represented by Mel-cepstral coefficient (MCC) ($\gamma = 0$, $\alpha = 0.42$ for 16 kHz speech) [9]. We used the modulation-spectrum of MCC sequences $\mathbf{c}_k = [c_{1k}, c_{2k}, ..., c_{Dk}]^\top$, $k = 1, 2, ..., T$, in which

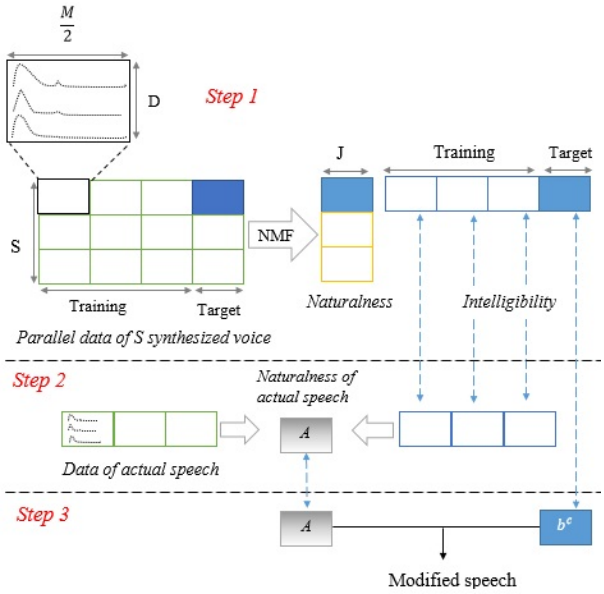


Figure 1. *Scheme of applying ABM for improving synthesized speech-spectra; N training sentences and 1 target sentence*

D is the order of cepstral coefficients and T is the number of frames, to determine the over-smoothing effect in both the time and frequency domains of speech spectra [10]. Short-term spectral analysis of a speech utterance yields a matrix $R = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_T]$ of size $D \times T$. The time trajectory of cepstral coefficient d is defined as $\mathbf{r}_d = [c_{d1}, c_{d2}, ..., c_{dT}]$, $d = 1, 2, ..., D$. The modulation-spectrum of trajectory r_d is defined as:

$$M(d, f) = |FT[\mathbf{r}_d]|, \tag{2}$$

where f is the modulation frequency bin, defined by the number of points in the Fourier transform (FT). The number of points in the FT must be greater than the maximum number of frames T of an utterance. The modulation-spectrum of each utterance is calculated for each coefficient. Using an ABM, the modulation-spectrum of synthetic speech-spectra is modified to be closer to the modulation characteristics of natural one.

The process of improving speech spectra consists of three major steps as shown in Figure 1:

1. Decomposition of naturalness and intelligibility of synthesized voices.
2. Obtaining naturalness of actual speech.
3. Reconstructing modified speech with intelligibility of synthesized voice and naturalness of actual speech.

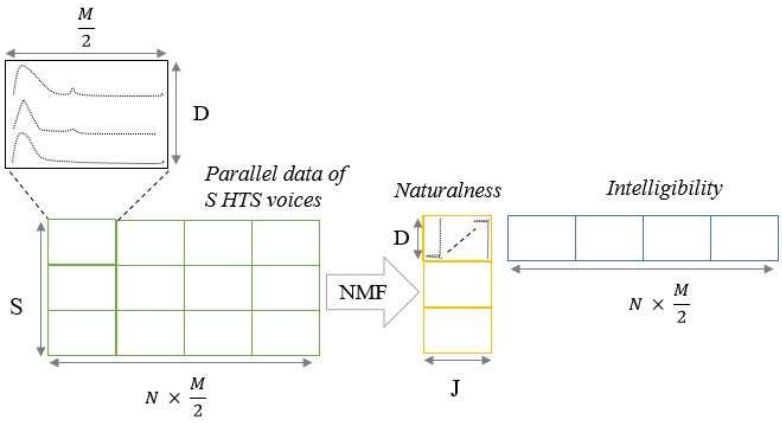


Figure 2. *Our ABM using NMF; S: same sentence and different HTSs*

2.1 Decomposition of naturalness and intelligibility of synthesized voices

The goal with step 1 is to obtain acceptable intelligibility from parallel data of synthesized voices to preserve the intelligibility. The naturalness and intelligibility factors were factorized from the data using NMF.

Stacking parallel data of S synthesized voices: We first prepared parallel data of a number of S HTS voices (PDSHV), as shown in Figure 2. In the PDSHV, the variation in different HTS voices' quality is presented in columns and that in phonetic information of different sentences is presented in rows. Phonetic information is assumed to be intelligible. To build this PDSHV, the modulation-spectrum of the MCC sequence from N sentences were stacked horizontally when the PDSHV is decomposed into two components: naturalness and intelligibility, as shown in Figure 2, where M denotes the number of FT points for modulation-spectrum, D is the MCC order, N is the number of sentences, S is the number of Vietnamese HTSs [11],[12] (S $\geq$ 2), and J is the number of model dimensions determined as $J = S \times D$ [7].

2.2 Natural improvement of HMMSS

In step 2, the naturalness of actual speech $\mathbf{A}$ was obtained using a small amount of actual speech y and corresponding intelligibility set C obtained from step 1. We derived the desired naturalness $\mathbf{A}$ by minimizing the total squared error over actual speech data,

$$E = \sum_{c \in C} ||\mathbf{y} - \mathbf{A}\mathbf{b}^c||^2 \tag{3}$$

Table 1. *Structure of Vietnamese syllable*

Tone			
Initial	Final		
	[Onset]	Nucleus	[Coda]

In Equation 3, intelligibility vectors $\mathbf{b^c}$ were learned from step 1. The desired naturalness $\mathbf{A}$ can be found by solving the linear system

$$\frac{\partial E}{\partial \mathbf{A}} = 0 \tag{4}$$

In step 3, the naturalness of actual speech $\mathbf{A}$ and intelligibility of synthesized speech were combined to obtained an improved version of synthesized speech.

3 Overview of applying ABM involving NMF for improving F0 contour

With Vietnamese, intonation managed by F0 contour is important for perceiving naturalness [12]. Moreover, tonal information determined by F0 contour is also important for intelligibility. In order to improve quality of Vietnamese HMM-based speech synthesis, modifying over-smoothed F0 contour is necessary.

3.1 Tonal information in Vietnamese

The Vietnamese is a monosyllabic and tonal language with six tones. According to [13], a syllable structure can be described as in Table 1. A syllable consists of initial and final part. The final part can be further divided into onset, nucleus, and coda. The onset and coda parts are optional. Six Vietnamese tones are: level, falling, broken, curve, rising, and drop.

3.2 Quality improvement of F0 contour

The process of applying an ABM for modifying synthesized F0 envelope is the same as that of applying an ABM for modifying synthesized speech-spectra. In Step 1, continuous F0 contour, interpolated in unvoiced regions, from N sentences were stacked horizontally to build PDSHV. Then, the PDSHV is decomposed into two components: naturalness and intelligibility. Intelligibility component, which consists of phonetic and tonal information, is preserved when we modify naturalness component.

In Step 2, naturalness component of natural F0 envelope is derived from a small data of actual speech. By combining naturalness component of natural F0 contour and intelligibility of synthesized F0 contour in Step 3, we obtain modified F0 contour.

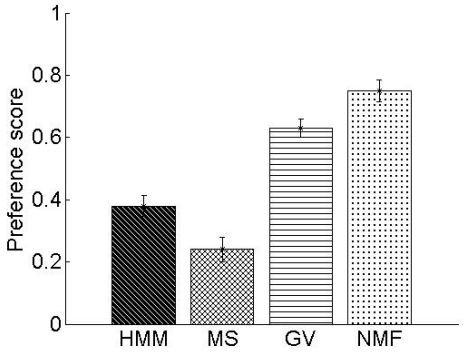
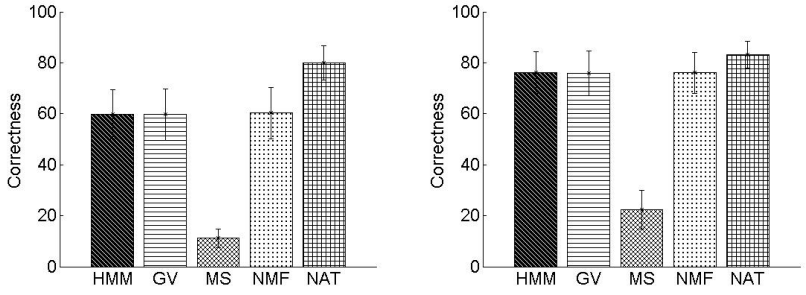


Figure 3. *Preference scores with 95% confidence interval*

4 Evaluation and Discussion

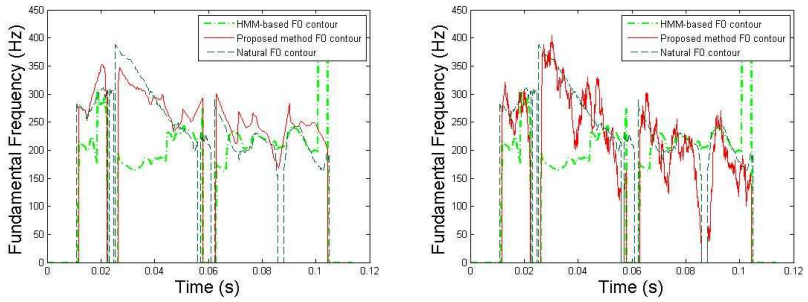


(a) *Tone intelligibility test* (b) *MRT*

Figure 4. *Experimental results with 95% confidence interval*

We evaluated the naturalness and intelligibility of the proposed method using a preference test, a Vietnamese-tone intelligibility test, and a modified rhythm test (MRT) under limited data conditions.

In the preference test, the proposed method was compared with other improvement methods such as those involving GV[2], and modulation-spectrum[4]. Two HMMSS ($S = 2$) were trained using 2 Vietnamese datasets (DEMEN567 and FEMALE1). The DEMEN567 was called TTSCorpus in [14]. This corpus has 567 utterances with sampling rate of 11025 Hz. The FEMALE1 includes 567 utterances with sampling rate of 16000 Hz. This dataset was used in [12] to show an improved naturalness of Vietnamese HMMSS by adding prosodic information to label files. These are limited data conditions comparing with other ordinary training conditions for HMM-based synthesizers. Both of the two datasets have



(a) *Use F0 values as input of proposed method*

(b) *Use modulation-spectrum of F0 contour as input of proposed method*

Figure 5. *The F0 contours of proposed method involving NMF, HMMSS, and natural speech*

a phoneme coverage of one hundred percentage. Fifteen utterances were synthesized for each voice. We applied proposed method and those involving GV, and modulation-spectrum to improve the quality of the samples under limited data conditions. The baseline was HMM trained with 500 utterances from FEMALE1 dataset. Only $N = 5$ natural sentences were used to train proposed method involving NMF. A number of $N = 5$ natural utterances from FEMALE1 dataset was also used in proposed method to derive naturalness factor of human speech in proposed method. A number of $N = 500$ natural sentences were used for training both methods involving GV and modulation-spectrum. All HMM-based synthesized utterances were aligned with their original human-speech using the guide of label files from FEMALE1 dataset. The STRAIGHT vocoder[16] was used to analyze the speech. The frame-shift was 5 ms and the frame-length was 10 ms. It decomposes speech into a spectral envelope, F0, and aperiodicity. Linear interpolation was used to generate F0 values in unvoiced regions. The STRAIGHT-based spectral parameters are further encoded into MCC. The cepstral order was 49 and the MCC sequences were transformed into the modulation-spectrum using FT; $M = 4096$. Eleven individuals (six northern-, and five southern-Vietnamese people) listened to 180 pairs of utterances. The participants are graduate students with normal hearing ability. They listened to each pair only once, then compared the naturalness of utterances on a two-point scale, i.e., 1 (A is more natural), and -1 (B is more natural). Natural speech was defined as actually human-speech.

Figure 3 shows that preference score of proposed method (denoted as NMF) is the best under limited data condition. Since there was not enough data to train context-dependent models considering tonal-information, the method involving GV (denoted as GV) did not perform well. Since the method involving modulation-spectrum (denoted as MS) does not consider tonal-information, it destroyed tonal-information and generated unnatural utterances. Therefore, its preference score is the worst. At the end of the experiments, participants were

asked what factors contribute to their decisions. All participants agreed that speech with buzzing sound and meaningless speech is not natural.

In the tonal-intelligibility test, we evaluate the tonal-intelligibility of synthesized speech after applying proposed method. We prepared 50 syllables synthesized by Vietnamese HMMSS [12]. The syllables were randomly selected from 67 remaining sentences in FEMALE1 dataset. Different methods involving NMF, GV, and modulation-spectrum were used to improve naturalness of the syllables. All configurations were the same as previous experiment. Twelve individuals (seven northern-, and five southern-Vietnamese people) listened to 250 syllables from HMMSS, proposed method involving NMF, other methods involving GV and modulation-spectrum, and human-speech. The 50 syllables spoken by human were used as reference. The participants are graduate students with normal hearing ability. Participants listened to each syllable only once, and selected the most likely syllable they heard among a group of syllables bearing different tones (e.g., la, là, lá, lả, lã, lạ). For each syllable, we put "not like all above syllables" in answer list. The correctness for each method was obtained by pooling responses for the method, tallying the total number of correctly selected syllables, dividing by the total number of listened-syllables, and multiplying by 100.

Figure 4(a) shows that the correctness of proposed method (denoted as NMF) is equal to that of HMMSS (denoted as HMM). It indicates that the tone-intelligibility of synthesized speech was preserved by proposed method. In contrast, the tone-intelligibility of synthesized speech was destroyed by the method involving modulation spectrum (denoted as MS).

In the modified rhyme test (MRT), we evaluate the intelligibility of synthesized speech after applying proposed method. We prepared 60 syllables synthesized by Vietnamese HMMSS [12]. The syllables were randomly selected from 67 remaining sentences in FEMALE1 dataset. Different methods involving NMF, GV, and modulation-spectrum were used to improve naturalness of the syllables. All configurations were the same as previous experiment. Eleven individuals (six northern-, and five southern-Vietnamese people) listened to 300 syllables from HMMSS, proposed method involving NMF, other methods involving GV and modulation-spectrum, and human-speech. The participants are graduate students with normal hearing ability. Participants listened to each syllable only once, and selected the most likely syllable they heard among a group of syllables bearing different final part (e.g., la, lan, lanh, lang). For each syllable, we put "not like all above syllables" in answer list.

Figure 4(b) shows that the correctness of proposed method (denoted as NMF) is equal to that of HMMSS (denoted as HMM). It indicates that the intelligibility of synthesized speech was preserved by proposed method. In contrast, the intelligibility of synthesized speech was destroyed by the method involving modulation-spectrum (denoted as MS). Natural speech is denoted as NAT.

When we apply proposed method involving NMF to modify F0 contour of synthesized speech, both HMM-based F0 values and modulation-spectrum of continuous HMM-based F0 contour were considered as input of our proposed method. In [4], a post-filter was applied to modify modulation-spectrum of con-

tinuous HMM-based F0 contour. However, applying our proposed method on modulation-spectrum of HMM-based F0 contour did not yield a good result. Unexpected fluctuations were added to obtained F0 contour by proposed method as in Figure 5(b). But, applying our proposed method on HMM-based F0 contours yielded better result as in Figure 5(a). The reason may be F0 envelope is smoother than speech spectra. Improving modulation-spectrum of F0 envelope means improving F0 contour's fine-structure which generate unexpected fluctuation for obtained F0 contour from prposed method.

5 Conclusion

We proposed a novel ABM using NMF to decompose the naturalness and intelligibility of Vietnamese HMMSS. The proposed method proved to be efficient in improving naturalness without violating the intelligibility of synthesized speech, especially under limited data condition. The proposed method can model and modify F0 envelope of Vietnamese. Experimental results demonstrated its superiority to other methods under a limited data condition. Our method provides a new way to control naturalness of synthesized speech under limited data conditions.

Acknowledgement

This study was supported by the Grant-in-Aid for Scientific Research (A) (No. 25240026) and the JSPS A3 Foresight program.

References

1. H. Zen, K. Tokuda and W. Black, "Statistical parametric speech synthesis," *Speech Comm.*, vol. 51, no. 11, pp. 1039–1064, 2009.
2. T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans*, vol. E90-D, no. 05, pp. 816–824, 2007.
3. S. Takamichi, T. Toda, A. Black, and S. Nakamura, "Parameter generation algorithm considering modulation spectrum for HMM-based speech synthesis," *Proc. of ICASSP*, pp. 4210–4214, 2015.
4. S. Takamichi, T. Toda, G. Neubig, and S. Nakamura, "A post-filter to modify the modulation spectrum in HMM-based speech synthesis," *Proc. of ICASSP*, pp. 290–294, 2014.
5. L. H. Chen, T. Raitio, C. Valentini-Botinhao, J. Yamagishi, Z. H. Ling, "DNN-based stochastic postfilter for HMM-based speech synthesis," *Proc. of Interspeech*, pp. 1954–1958, 2014.
6. J. Tenenbaum, W. Freeman, "Separating style and content with bilinear models," *Neural Computation*, pp. 1247–1283, 2000.
7. V. Popa, J. Nurminen, M. Gabbouj, "A novel technique for voice conversion based on style and content decomposition with bilinear models," *Proc. of Interspeech*, pp. 2655–2658, 2009.

8. Y. Stylianou, O. Cappe, E. Moulines, "Continuous probabilistic transform for voice conversion" *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 6, pp. 131–142, 1998.
9. K. Tokuda, T. Masuko, S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," *Proc. of ICSLP*, pp. 1043–1046, 1994.
10. T. Dinh-Anh, D. Morikawa, M. Akagi, " A study on quality improvement of HMM-based synthesized voices using asymmetric bilinear model," *Journal of Signal Processing*, vol. 20, no. 4, 2016 (Accepted).
11. T.T. Vu, M.C. Luong, S. Nakamura," An HMM-based Vietnamese Speech Synthesis System," *Proc. of Oriental COCOSDA*, pp. 116—121, 2009.
12. T.S. Phan, T. C. Duong, A.T. Dinh, T.T. Vu, M.C. Luong, " Improvement of naturalness for an HMM-based Vietnamese speech synthesis using the prosodic information," *Proc. of RIVF*, pp. 276–281, 2013.
13. T.T. Doan, " Ngu am tieng Viet (Vietnamese Phonetics)," Hanoi National University Publishing House, pp. 99–148, 1999.
14. L.C. Mai and D.N. Duc, "Design of Vietnamese speech corpus and current status," *Proc. ISCSLP-06*, pp. 748-–758, 2006.
15. H. Scheffe, "An analysis of variance for paired comparisons," *Journal of the American Statistical Association*,vol. 37, pp. 381-400, 1952.
16. H. Kawahara, I. Masuda-Katsue, M. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and a instantaneous frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Comm.*, vol. 27, pp. 187–207, 1999.