Quality improvement of HMM-based synthesized speech based on decomposition of naturalness and intelligibility using non-negative matrix factorization

Anh-Tuan Dinh
Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Nomi, Japan 923–1211
Email: tuan.dinh@jaist.ac.jp

Masato Akagi
Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Nomi, Japan 923–1211
Email: akagi@jaist.ac.jp

***Abstract*—Hidden Markov model based synthesized speech is intelligible but not natural because of over-smoothing of the speech spectra. The purpose of this study is improving naturalness without violating acceptable intelligibility by decomposing the naturalness and intelligibility of synthesized speech using a novel asymmetric bilinear model involving non-negative matrix factorization. Subjective evaluations carried out on English data confirm that the proposed method outperforms original asymmetric bilinear model involving singular value decomposition in factorizing naturalness and intelligibility. Moreover, the performance of the proposed method is comparable with other methods.**

***Index Terms*—Hidden Markov model (HMM), Non-negative matrix factorization (NMF), Singular value decomposition (SVD).**

I. Introduction

The HMM-based speech synthesis is a state-of-the-art method due to its flexibility and compact footprint [1]. 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. There have been several attempts to overcome the over-smoothing effect. One approach is using objective evaluations of this effect such as global-variance [2], and modulation-spectrum [3], integrating them into the parameter generation phase to obtain better speech parameter values. Since the quality improvement is dependent on synthesizers, we cannot improve the naturalness of existing speech synthesizers without re-training the synthesizers. 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 existing speech synthesizers. Thus, a voice-conversion approach is used to improve naturalness.

With the majority of previous voice-conversion approaches, all spectra are modified to improve naturalness. However, applying these approaches often negatively affect intelligibility. To improve naturalness without violating intelligibility, an asymmetric bilinear model [6] was introduced to decompose naturalness and intelligibility. Popa et al. [7] used an asymmetric bilinear model 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 $\mathbf{W}$ and intelligibility factor $\mathbf{h}^c$ of an intelligibility class c:

$$\mathbf{y} = \mathbf{W}\mathbf{h}^c \tag{1}$$

In the above representation, naturalness can be modified, whilst intelligibility can be preserved.

There are two problems with applying an asymmetric bilinear model. The first is finding an efficient acoustic feature vector strongly related to naturalness. The naturalness-associating acoustic feature will be modified, while intelligibility-corresponding acoustic features will be preserved.

The second problem is finding an efficient constraint to decompose naturalness and intelligibility. Although an asymmetric bilinear model 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 a novel asymmetric bilinear model involving non-negative matrix factorization (NMF).

Section II shows that modifying spectral parameters is very important for improving perceived-naturalness. In the work, we modify spectra of HMM-based synthesized speech and retain other acoustic features as in Fig. 1.

II. Finding efficient acoustic feature vector

We carried out an experiment to find an efficient acoustic feature vector strongly related to naturalness. With a certain kind of acoustic feature, the feature values are exchanged between a pair of human speech and synthesized speech. If such an exchanging drastically improves the naturalness of

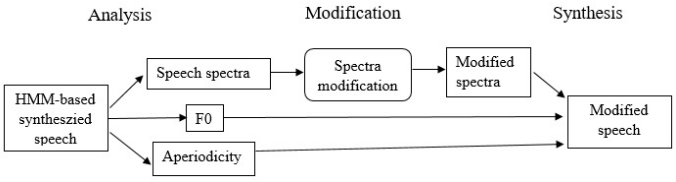


Fig. 1: Schematic of improving naturalness of HMM-based synthesized speech by modifying synthesized spectra

TABLE I: Stimuli obtained from exchanging different acoustic feature vectors between natural speech and synthesized speech

Group	Stimuli	Natural speech with *acoustic feature of synthesized speech*	Stimuli	Synthesized speech *with acoustic feature of human speech*
A	A1	Cepstrum	A2	Cepstrum
B	B1	F0	B2	F0
C	C1	LPC	C2	LPC
D	D1	LPC with power	D2	LPC with power
E	E1	LSF	E2	LSF
F	F1	LSF with power	F2	LSF with power
G	G1	MCC	G2	MCC
H	H1	MFCC	H2	MFCC
J	J1	PLP	J2	PLP

synthesized speech, that acoustic feature strongly relates to naturalness. Therefore, our experiment was composed of three steps:

1) Exchanging acoustic feature values.
2) Comparing naturalness by using a listening test.
3) Finding an efficient acoustic feature

In the first step, several types of acoustic features were prepared, i.e., fundamental frequency (F0), formant-related parameters, e.g., the linear prediction coefficient (LPC) w/wo residual power, linear spectral frequency (LSF) w/wo residual power, and perceptual linear prediction (PLP), and fine–structure– related coefficients, e.g., Mel-frequency cepstral coefficient (MFCC) [8], Mel-cepstral coefficient (MCC) ($\gamma = 0$, $\alpha = 0.42$ for 16 kHz speech) [9] and cepstrum. Since over-smoothing broadens the formants' bandwidth of speech spectra, sharpening the formant peaks can mitigate over-smoothing of the spectral envelope. Therefore, the LPC, LSF, and PLP are taken into consideration to control formants. Since improving the fine structure enhances the dynamics of the generated spectral envelope, the MFCC, MCC, and cepstrum are considered to manipulate the fine-structure of speech spectra. In the experiment, one utterance for one natural speech sentence was synthesized by HMM-based speech synthesis system (HTS) [1]. The synthesized speech was aligned to its original speech with guide–of–label files. The STRAIGHT vocoder[11] was used to analyze the speech. It decomposes speech into a spectral envelope, F0, and aperiodicity. The STRAIGHT-based spectral parameters are further encoded into LPC, LSF, MFCC, MCC, PLP, and cepstrum. After this step, 20 stimuli including the 18 stimuli listed in Table I, natural speech (denoted as I1), and HMM-based synthesized speech (denoted as I2) were obtained.

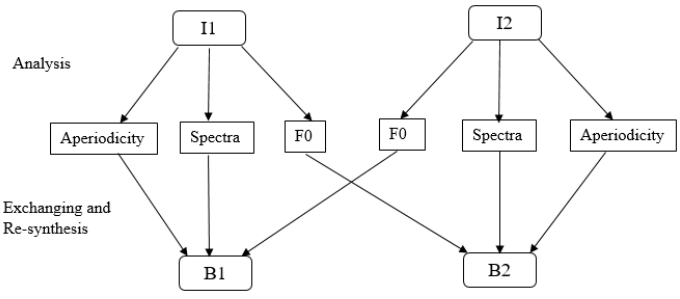


Fig. 2: Schematic of exchanging process; I1 denotes HMM-based synthesized speech; I2 denotes corresponding human speech

Figure 2 shows an example of exchanging F0-related features between HMM-based synthesized utterance and its corresponding natural utterance. The F0 contour is extracted from synthesized speech I2, appropriately time-aligned with the guide-of-label files, and then imposed to natural speech I1. In essence, the HMM-based F0 contour replaces the Natural F0 contour, while all other Natural attributes are retained, thus creating an hybrid waveform B1.

In the second step, the naturalness of the 20 stimuli was compared using Scheffe's method of paired comparison [10]. Six individuals (non native English speakers with fluent English level) participated in the experiment. Each participant listened to 380 pairs of stimuli. With each pair, they compared the naturalness of the stimuli on a five–point scale, e.g., -2 (the former is more natural), 0 (comparable), +2 (the latter is more natural).

In the third step, the efficient acoustic feature was determined by looking for one that improved the naturalness of synthesized speech the most. The experimental results in Fig. 3 indicate that exchanging MCC values improves the naturalness of synthesized speech the most (I2 to G2).

In the frequency domain, fine structure is more important than formant in perceiving naturalness. For the first problem of using asymmetric bilinear model, the MCC is the most efficient acoustic feature for improving naturalness.

Although the MCC can represent the fine structure in the frequency domain, it cannot represent the dynamics of the spectra in the time domain. The modulation spectrum has recently become a popular concept in capturing the fine structure of speech spectra in the time domain. We used the modulation-spectrum of MCC sequences $\mathbf{c}_k = [c_{1k}, c_{2k}, ..., c_{Pk}]^\top$, $k = 1, 2, ..., T$, in which P 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. Short-term spectral analysis of a speech utterance yields a matrix $R = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_T]$ of size $P \times T$. The time trajectory of cepstral coefficient p is defined as $\mathbf{r}_p = [c_{p1}, c_{p2}, ..., c_{pT}]$, $p = 1, 2, ..., P$. The modulation-

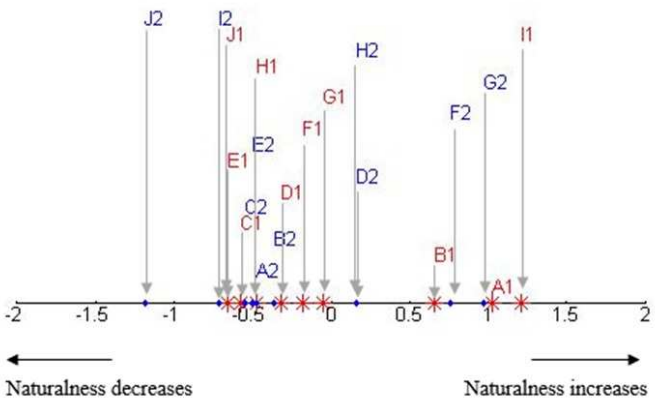


Fig. 3: Results of pair-comparison test

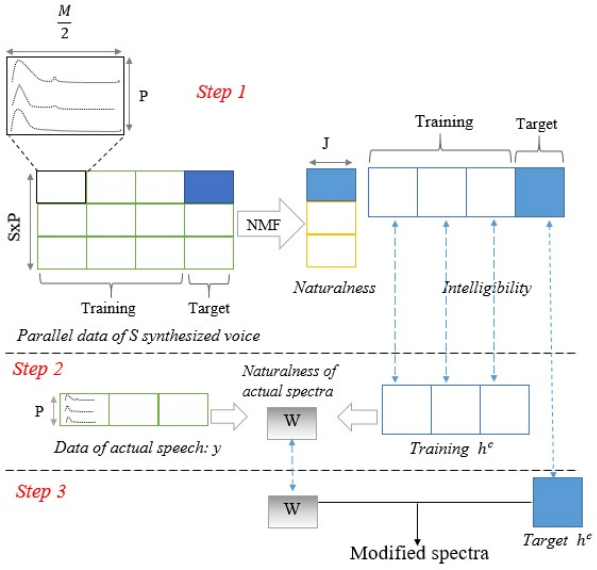


Fig. 4: Schematic of applying asymmetric bilinear model for modifying synthesized spectra; N training sentences with 1 target sentence

spectrum of trajectory r_p is defined as:

$$M(p, f) = |FT[\mathbf{r}_p]|, \quad (2)$$

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 asymmetric bilinear model, the modulation-spectrum of synthetic trajectories is modified to be closer to the modulation characteristics of natural speech.

III. Overview of modifying synthesized spectra using asymmetric bilinear model involving NMF

In the section, we describe the process of applying an asymmetric bilinear model for modifying HMM-based speech

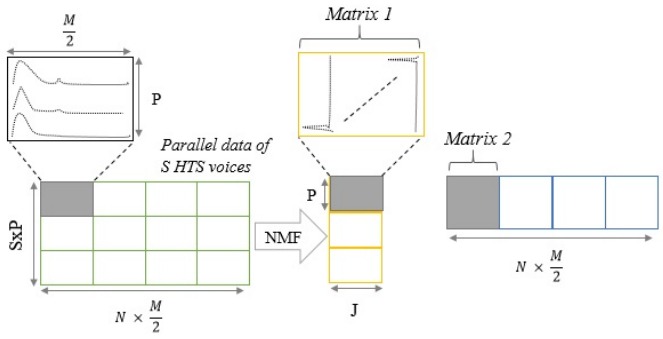


Fig. 5: Our asymmetric bilinear model using NMF; S: same sentence and different HTSs

spectra. The process consists of three major steps as shown in Figure 4:

1) Decomposing HMM-based spectral-parameters into naturalness and intelligibility components.
2) Obtaining naturalness component of actual spectral-parameters.
3) Reconstructing modified spectral-parameters with target-intelligibility of HMM-based spectra and naturalness of actual spectral-parameters.

A. Decomposing HMM-based spectral-parameters into naturalness and intelligibility components

The goal with step 1 was to obtain acceptable intelligibility from parallel data of synthesized spectra to preserve the intelligibility-component. The naturalness and intelligibility components were factorized from the parallel data using NMF.

1) Stacking parallel data of S synthesized voices: We first prepared parallel data of a number of S HTS voices, as shown in Fig. 5. In the parallel data, the variation in the quality of different HTS voices is presented in columns and that in phonetic information of a number of N different sentences is presented in rows. Phonetic information is assumed to be intelligible. To build the parallel data, modulation-spectrum of the MCC sequence from N sentences were stacked horizontally when the parallel data is decomposed into two components: Matrix 1 and Matrix 2, as shown in Fig. 5, where M denotes the number of FT points for modulation-spectrum, P is the MCC order, S is the number of HTSs [1] (S $\geq$ 2), and J is the number of model dimensions determined as $J = S \times P$ [7]. The next step was to prove that the proposed method involving NMF can decompose naturalness and intelligibility components from the parallel data of synthesized spectral-parameters. Evidence about strong relation between Matrix 1 and perceived-naturalness, and between Matrix 2 and perceived-intelligibility is provided in the next subsections.

2) Examining relation between Matrix 1 and perceived-naturalness: We conducted an experiment to prove the relation between Matrix 1 and perceived-naturalness. To do so, several pairs of HMM synthesized speech and natural speech were decomposed into two components. The first component called Matrix 1 is exchanged between the HMM-based synthesized

speech and natural speech. If exchanging Matrix 1 improves the perceived-naturalness of synthesized speech and decreases that of human speech, Matrix 1 strongly relates to naturalness. The experiment consisted of the following steps:

(i) Decomposing HMM-based synthesized speech and natural speech using our asymmetric bilinear model with NMF
(ii) Exchanging Matrix 1 between HMM-based synthesized speech and natural speech
(iii) Comparing naturalness of obtained stimuli using preference test

Figure 6 shows the process of decomposing human speech and synthetic speech in step (i). First, parallel data of human voices and that of HMM-based voices were prepared. A number of $S = 3$ HMM-based synthesized voices were trained using 3 CMU datasets (SLT, CLB, and RMS). A number of $N = 18$ sentences were synthesized using the $S = 3$ HMM-based synthesizers with the guide-of-label files (in total 54 utterances). The 54 HMM-based synthesized utterances were used to build a parallel data of HMM-based voices. We used 54 original utterances with the same sentences as synthesized ones from the 3 CMU datasets (SLT, CLB and RMS) to build a parallel data of human voices. To stack the parallel data, all utterances were analysed into F0, spectral envelope, and aperiodicity using STRAIGHT. The frame-shift was 5 ms and the frame-length was 10 ms. The spectral envelope was represented by the MCC. The cepstral order was 49 and the MCC sequences were transformed into the modulation-spectrum using FT; $M = 4096$. In other word, each utterance was described with $M = 4096$ parameter vectors. The modulation-spectrum of $N = 18$ utterances from one dataset was horizontally stacked as row of the parallel data. The modulation-spectrum of 3 utterances which has same content and come from different speakers was stacked vertically as column of the parallel data. Then the parallel data were decomposed by NMF. In Fig. 6, the D denotes the speech parameters of a synthesized utterance generated by SLT-synthesizer. The A denotes the speech parameters of a human utterance of the same sentence and same SLT dataset; A is decomposed into two components denoted as (1) and (2) and D is factorized into two factors denoted as (3) and (4).

In step (ii), the first components (or Matrix 1) in the factorized matrices were exchanged between HMM-based synthesized speech and natural speech of the same sentence, and speaker. For example, in Fig. 6, components (1) and (3) were exchanged. Therefore, with each sentence of a certain speaker, there are four types of stimuli as follows:

- Stimuli A consists of spectral parameters A comprising of (1) and (2).
- Stimuli B consists of spectral parameters B comprising of (1) and (4).
- Stimuli C consists of spectral parameters C comprising of (3) and (2).
- Stimuli D consists of spectral parameters D comprising of (3) and (4).

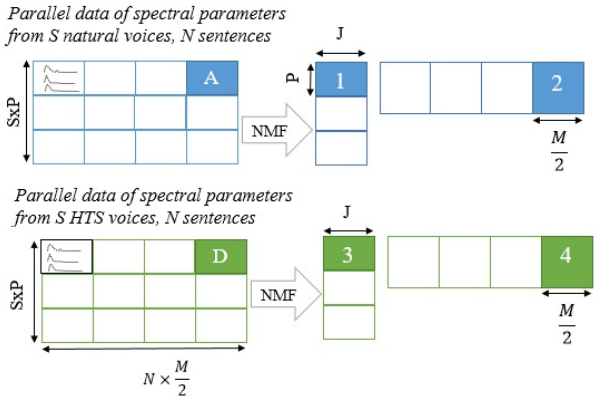


Fig. 6: Decomposing of synthesized speech and human speech; S: same sentence and different HTSs for HTS voice; S: same sentence and different natural voices for human voice; M: order of modulation spectrum; N: no. different sentences

In step (iii), the naturalness of the stimuli was compared using a preference test. Ten individuals (8 non-native English speakers and 2 native English speakers) participated. Each participant listened to 216 pairs of stimuli. With each pair, they compared the naturalness of stimuli on a two-point scale, i.e., 1 (the former is more natural), and -1 (the later is more natural). In Fig. 8(a), the preference score of Stimuli A (or natural speech) reduced to that of Stimuli C, while, that of Stimuli D (or HMM-based synthesized speech) increased to that of Stimuli B after exchanging component (1) and component (3). In other words, the naturalness of Stimuli A decreased and the naturalness of Stimuli D increased after exchanging Matrix 1 (e.g., component (1) and component (3) in Fig. 6). This indicates a strong relation between Matrix 1 and naturalness.

Figure 7 shows the differences between Matrix 1 of one synthesized speech and that of the original speech. The columns of the matrix contain information about the magnitude of the cepstral coefficients. The difference becomes clear in high-order cepstral coefficients, which represent the fine structure of speech spectra. The magnitude of cepstral coefficients from the human speech is larger than that of the HMM-based synthesized speech, especially in the high-order region. Moreover, the fine structure is strongly related to naturalness. Therefore, by emphasizing the magnitude of the cepstral coefficients, especially in the high-order region, naturalness can be improved.

3) Examining relation between Matrix 2 and perceived-intelligibility: Another experiment was conducted to prove the relation between Matrix 2 (e.g., component (2) and component (4) in Fig. 6) and perceived-intelligibility with the same procedure and configuration as the previous experiment. Unlike the previous test, testing words in modified rhythm test were synthesized by using 3 HMM-based synthesized voices trained from SLT, CLB and RMS datasets. Matrix 2 was exchanged between synthesized word segments and natural word segments. Four types of stimuli were obtained, similar to the previous experiment. The intelligibility of the stimuli

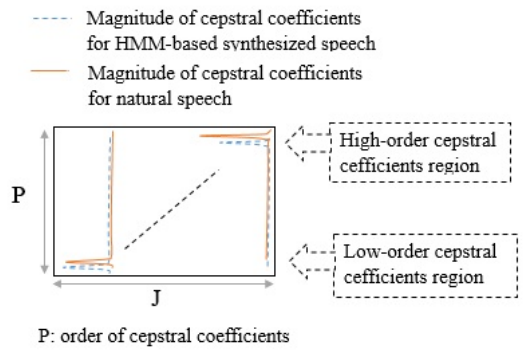
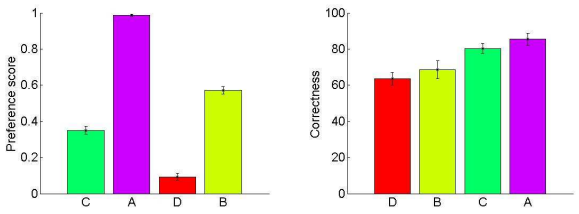


Fig. 7: Difference between Matrix 1 of synthesized speech and that of human speech



(a) Preference test (b) Modified rhythm test

Fig. 8: Experimental results with 95% confidence interval

was compared using the modified rhythm test. Ten individuals (eight native English speakers and two non-native English speakers) with normal hearing ability participated. As shown in Fig. 8(b), the word correctness of Stimuli A decreased to that of Stimuli B, and that of Stimuli D increased to that of Stimuli C. Expected results are Stimuli B is comparable to Stimuli D while Stimuli C is comparable to Stimuli A. The experimental results indicate a strong relation between Matrix 2 and intelligibility. Our asymmetric bilinear model involving NMF can decompose the naturalness and intelligibility of HMM-based synthesized speech.

B. Modification of HMM-based spectral-parameters

In step 2 for modifying HMM-based speech spectra in Fig. 4, the naturalness-component of actual spectral-parameters $\mathbf{W}$ was obtained using a small set of of actual speech y and corresponding intelligibility set C obtained from step 1 as shown in Fig. 4. We derived the desired-naturalness $\mathbf{W}$ by minimizing the total squared error over actual speech data,

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

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

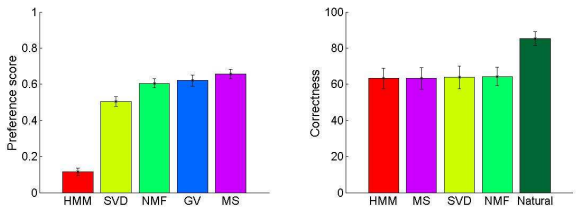
$$\frac{\partial E}{\partial \mathbf{W}} = 0 \tag{4}$$

In step 3 in Fig. 4, the naturalness of actual spectral-parameters $\mathbf{W}$ and intelligibility of synthesized spectral-parameters: target $\mathbf{h}^c$ are combined to obtained an improved spectral-parameters as shown in Fig. 4.

IV. Evaluation and Discussion

We evaluated the naturalness and intelligibility of the proposed method using a preference test and modified rhythm test. In the preference test, the proposed method was compared with other improvement methods such as those involving global-variance [2], and modulation-spectrum post-filter [4]. Two HMM-based synthesizers ($S = 2$) were trained using 2 CMU datasets (SLT and RMS). With each dataset, 500 utterances were used for training an HMM-based synthesizer. Ten sentences were chosen as testing sentences. Ten sentences were synthesized using synthesizer trained from SLT dataset as baseline samples. We applied proposed method and those involving SVD, global-variance, and modulation-spectrum to improve the quality of the samples. A number of 500 training utterances which were used to train HMM-based synthesizer were used for training both methods involving global-variance and modulation-spectrum. To improve spectral parameters of the baseline samples using our proposed method involving NMF, we synthesized 5 training sentences using $S = 2$ synthesizers trained from SLT and RMS datasets (in total 10 training utterances) in step 1 in Fig. 4. We also synthesized fifteen testing sentences using $S = 2$ synthesizers trained from SLT and RMS datasets (in total 20 testing utterances). With each testing sentence, a parallel data of spectral parameters was formed using spectral parameters of 10 training utterances and 2 testing utterances from $S = 2$ different HMM-based synthesized voices and $N = 5$ different sentences. The objective is modifying the testing utterance generated by SLT-synthesizer. In step 2 in Fig. 4, original utterances of 5 training sentences from SLT dataset was also used to derive naturalness of actual speech. We combined the naturalness of actual speech with intelligibility of testing sentence obtained from step 1 to obtain modified spectral parameter for the testing sentence. The process of modifying F0 contour is the same as that of modifying spectral parameter. All HMM-based synthesized utterances were aligned with their original human-speech using the guide of label files from SLT dataset. The configurations were similar to the previous experiments. Eleven participants (ten non-native and one native English speakers) listened to 400 pairs of stimuli. 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 (the former is more natural), and -1 (the latter is more natural). Natural speech was defined as actually human-speech.

Figure 9(a) shows that preference score of proposed method involving NMF is higher than method involving SVD, and comparable to the method involving global-variance (denoted as GV) and modulation-spectrum post-filter (denoted as MS). It's important to notice that a number of 500 training sentences was used for methods involving global-variance and modulation-spectrum post-filter, while only a number of 5 training sentences was used for methods involving NMF and



(a) Preference test (b) Modified rhythm test

Fig. 9: Experimental results with 95% confidence interval

SVD. It indicates that proposed method involving NMF outperform original asymmetric bilinear model involving SVD. It also indicates that NMF is more efficient than SVD in decomposing naturalness and intelligibility. At the end of the experiments, participants were asked what factors contribute to their decisions. All participants agreed that speech with buzzing sound and speech that were flat is not natural.

In modified rhythm test, we evaluated intelligibility of synthesized speech after applying proposed method. Ten individuals (eight non-native and two native English speakers) with normal hearing ability participated. Figure 9(b) show that the correctness are comparable in all methods. The results indicate that intelligibility of synthesized speech is preserved with the proposed method (denoted as NMF).

For the second problem of applying asymmetric bilinear model in factorizing naturalness and intelligibility of synthesized speech, the experimental results indicate that non-negativity constrain is an efficient constraints to decompose naturalness and intelligibility.

V. Conclusion

We proposed a novel asymmetric bilinear model using NMF to decompose the naturalness and intelligibility of HMM-based synthesized speech. The MCC was determined as an efficient acoustic feature strongly related to perceived-naturalness of speech. The proposed method outperforms original asymmetric bilinear model involving SVD. The NMF is more efficient than the SVD in decomposing naturalness and intelligibility. In other words, non-negativity constraint is efficient in decomposing naturalness and intelligibility of synthesized speech. The proposed method is comparable to other methods although our method use only 5 training sentences. Our method provides a new way to control naturalness of speech through modifying the magnitude of high-order cepstral coefficients.

Acknowledgment

This study was supported by the Grant-in-Aid for Scientic Research (A) (No. 25240026), SECOM Science and Technology Foundation 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," *ICASSP Proceeding*, 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," *ICASSP Proceeding*, 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," *Interspeech Proceeding*, 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," *Interspeech Proceeding*, 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] H. Scheffe, "An analysis of variance for paired comparisons," *Journal of the American Statistical Association*,vol. 37, pp. 381-400, 1952.

[11] 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," *Journal of Speech Communication*, vol. 27, pp. 187–207, 1999.