Journal of Signal Processing, Vol.20, No.4, pp.205-208, July 2016

SELECTED PAPER AT NCSP'16

Study on Quality Improvement of HMM-Based Synthesized Voices Using Asymmetric Bilinear Model

Tuan Dinh-Anh, Daisuke Morikawa and Masato Akagi

School of Information Science, Japan Advanced
Institute of Science and Technology
1–1 Asahidai, Nomi, Ishikawa 923–1292, Japan
E-mail: {tuan.dinh, morikawa, akagi}@jaist.ac.jp

Abstract

Hidden Markov model (HMM)-based synthesized voices are intelligible but not natural especially under limited-data conditions due to over-smoothed speech spectra. Improving naturalness is a critical problem of HMM-based speech synthesis. One solution is to use voice conversion techniques to convert over-smoothed spectra to natural spectra. Although conventional conversion methods transform speech spectra to natural ones to improve naturalness, they cause unexpected distortions in the intelligibility of synthesized speech. The aim of the study is to improve naturalness without reducing the intelligibility of synthesized speech by employing our novel asymmetric bilinear model (ABM) to separate the intelligibility and naturalness of synthesized speech. In the study, our ABM was implemented on the modulation spectrum domain of Mel-cepstral coefficient (MCC) sequences to enhance the fine structure of spectral parameter trajectory generated from HMMs. Subjective evaluations carried out on English data confirmed that the achieved naturalness of the method using the ABM involving singular value decomposition (SVD) was competitive with other methods under large-data conditions and outperformed other methods under limited-data conditions. Moreover, modified rhyme test (MRT) showed that the intelligibility of synthesized speech was well preserved with our method.

1. Introduction

Hidden Markov model (HMM)-based speech synthesis is a state-of-the-art method due to its flexibility and compact footprint [1]. An 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 the over-smoothing effect under limited-data conditions.

There have been several attempts to overcome the over-smoothing effect. One approach is to use objective evaluations of over-smoothing effect such as global variance (GV) [2], and modulation spectrum (MS) [3], integrating them into the parameter generation phase to obtain better speech parameters. The joint optimization of HMMs and objective evaluation typically does not have a closed-form solution. Another possible way 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] - [6] can be used for mapping natural spectra to synthetic spectra. The approach benefits from optimizing HMMs with a closed-form solution. Thus, a voice conversion approach is used to improve the naturalness.

In the majority of previous voice conversion approaches, all spectra were modified to improve naturalness. However, applying these approaches often negatively affected intelligibility. To improve naturalness without reducing intelligibility, an experiment was conducted to find efficient acoustic features strongly related to naturalness. Then, the features related to naturalness were converted to improve the quality of synthesized speech, while other intelligibility-related features were preserved.

This paper is organized as follows. In Sect. 2., we review the ABM [7] and show the general framework of using the ABM involving SVD to improve the naturalness of synthesized speech. The problems of using the ABM are also addressed in the section. In Sect. 2.1 we attempt to solve the problems. In Sect. 3., we demonstrate the benefits of the ABM involving SVD through listening test results. Finally, concluding remarks, including a potential future research direction, are presented in Sect. 4.

2. Naturalness Improvement Using ABM

In an ABM [7], an observation y^{sc} from speaker s and phonetic class c can be represented as

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

with A^s denoting speaker information and b^c denoting phonetic information. In this paper, phonetic information b^c is assumed to correspond to intelligibility, and speaker information A^s is assumed to correspond to naturalness. In [7], observation y^{sc} was the line spectral frequency (LSF) vector. The LSF is a way to model speech spectra with emphasis of formants, which are important for speaker characteristics. However, it is not clear whether formants are sufficient to perceive naturalness. Determining features that can be used to efficiently increase naturalness is very important.

2.1 Determining efficient acoustic feature vector

We carried out an experiment to find an efficient acoustic feature vector strongly related to naturalness. For a certain acoustic feature, the feature values are exchanged between a pair of human speech and synthesized speech. If such an exchanged markedly improves the naturalness of the synthesized speech, the acoustic feature is strongly related to naturalness. Therefore, our experiment was composed of three steps:

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

In the first step, several different acoustic features were prepared, i.e., fundamental frequency (F0), formant-related parameters, e.g., the linear prediction coefficient (LPC) w/wo residual power, LSF w/wo residual power, and perceptual linear prediction (PLP), and fine-structure-related coefficients, e.g., the Mel-frequency cepstral coefficient (MFCC) [8], the Mel-cepstral coefficient (MCC) ($\gamma = 0, \alpha = 0.42$ for 16 kHz speech) [9] and cepstrum. To examine each acoustic feature, the feature sequences were exchanged between synthesized speech and natural speech of the same sentence. Exchanging the acoustic feature results in improving the naturalness of the synthesized speech and decreasing the quality of the natural speech. If the quality of the natural speech decreases and the naturalness of the synthesized speech increases markedly after the exchange, the acoustic feature is strongly related to naturalness. In the experiment, one utterance for one natural speech sentence was synthesized by an HMM-based speech synthesis system (HTS) [1]. The synthesized speech was aligned to its original speech using guide-of-label files. The STRAIGHT vocoder [10] was used to analyze the speech. It decomposes speech into a spectral envelope, F0, with aperiodicity. The STRAIGHT-based speech parameters were further encoded into the LPC, LSF, MFCC, MCC, PLP, and cepstrum. After the step, 20 stimuli were obtained as listed in Table 1.

In the second step, the naturalness of the obtained stimuli was compared using Scheffe's method of paired comparison [11]. Six listeners (non native English speakers with fluent English level) with normal hearing condition participated. Each participant listened to 380 pairs of stimuli. For each pair, they compared the naturalness of stimuli on five-point scale, from -2 (A is more natural) through 0 (comparable) to $+2$ (B is more natural).

In the third step, an efficient acoustic feature was determined by finding the feature that improved the naturalness of the synthesized speech the most. The experimental results in Fig. 1 indicate that exchanging the MCC values improves the naturalness of synthesized speech the most (I2 to G2). Exchanging the LSF does not significantly improve naturalness (I2 to E2). In frequency domain, fine structure is more important than formant in perceiving naturalness. The MCC is the most efficient acoustic feature in improving naturalness.

Table 1: Stimuli in experiment

Stimulus	Meaning	Stimulus	Meaning
A1	Natural speech after exchanging (Nat) cepstrum	A2	Synthesized speech after exchanging (HMM) cepstrum
B1	Nat F0	B2	HMM F0
C1	Nat LPC	C2	HMM LPC
D1	Nat LPC with power	D2	HMM LPC with power
E1	Nat LSF	E2	HMM LSF
F1	Nat LSF with power	F2	HMM LSF with power
G1	Nat MCC	G2	HMM MCC
H1	Nat MFCC	H2	HMM MFCC
I1	Natural speech	I2	Synthesized speech
J1	Nat PLP	J2	HMM PLP

Although the MCC can represent the fine structure in the frequency domain, it cannot represent the dynamics of 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 MS of MCC sequences $\mathbf{c}_k = [c_{1k}, c_{2k}, ..., c_{Dk}]^\top$, $k = 1, 2, ..., T$, in which D is the order of cepstral analysis 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 $D \times T$. The time trajectory of cepstral coefficient d is defined as $\mathbf{r}_d = [c_{d1}, c_{d2}, ..., c_{dT}]$, $d = 1, 2, ..., D$. The MS 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 FT points must be greater than the maximum number of frames T of an utterance. The MS of each utterance is calculated for each coefficient. Using the ABM, the MS of synthetic trajectories is modified to make it closer to the modulation characteristics of natural speech.

2.2 Scheme to employ ABM

In the section, we describe the process of applying the ABM to improve naturalness. The process consists of three major steps as shown in Fig. 2:

1. Decomposition of naturalness and intelligibility of synthesized voices
2. Obtaining naturalness of actual speech

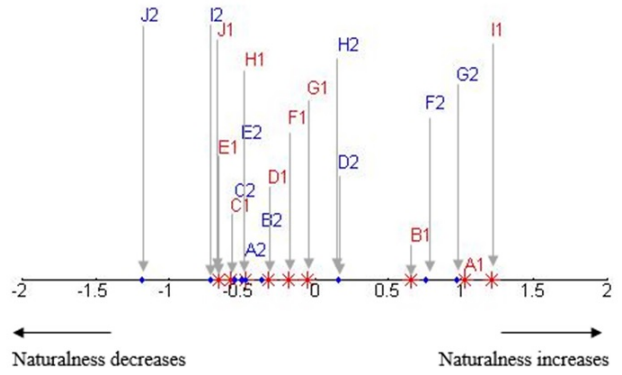


Figure 1: Results of paired comparison test

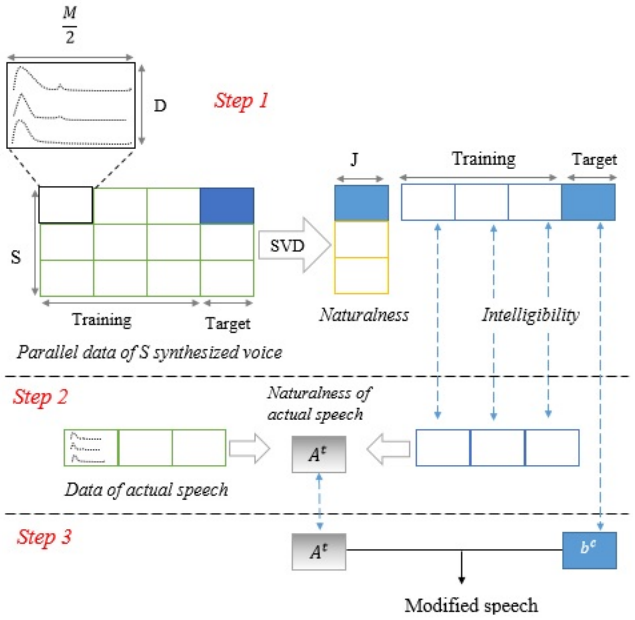


Figure 2: Scheme of applying ABM

3. Reconstructing modified speech with intelligibility of synthesized voice and naturalness of actual speech

The goal of step 1 is to obtain the intelligibility from parallel data of synthesized voices to preserve the intelligibility. The naturalness component and intelligibility component are factored from the data using SVD. Each cell of the parallel data of synthesized speech (PDSS) is the MS of one utterance. In Figure 2, M denotes FT points for the MS, D is the order of the MCC, and S denotes the number of HTSs [1] ($S \geq 2$), and J is the model dimensionality, chosen to be $J = S \times D$. Because SVD results in three matrices $USV^\top$, naturalness matrix is chosen as first J columns of US and intelligibility matrix is chosen as first J rows of $V^\top$. In the PDSS, the variation of naturalness is presented in columns, and the variation of intelligibility is presented in rows. The columns of the naturalness matrix summarize vertical structure of the PDSS associated with naturalness. Likewise the rows of the intelligibility matrix summarize the horizontal structure of the PDSS.

In step 2, the naturalness of an actual speech A^s is obtained using a small amount of data from the actual speech y^{sc} and the corresponding intelligibility set C obtained from step 1. We derive the desired naturalness A^s by minimizing the total squared error over the actual speech data,

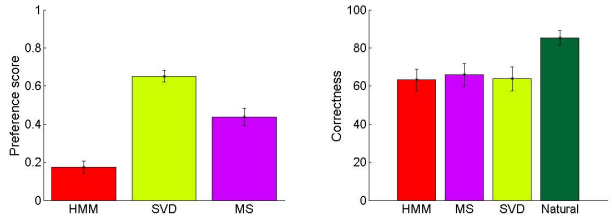
$$E = \sum_{c \in C} ||y^{sc} - A^s b^c||^2 \tag{3}$$

In step 3, the naturalness of the actual speech A^s and the intelligibility of synthesized speech are combined to obtained an improved version of the synthesized speech.

3. Evaluation and Discussion

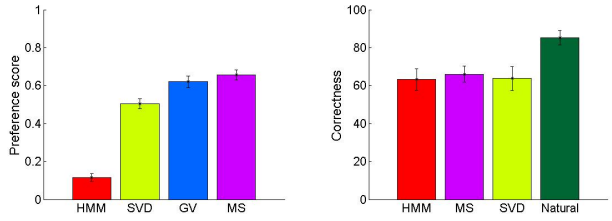
We evaluated naturalness and intelligibility of speech processed by the AMB involving SVD by performing a preference test and a modified rhythm test (MRT). In preference test, the ABM involving SVD was compared with other methods involving GV [2], and MS post-filter [4]. Two HMM-based synthesized voices were trained using 2 CMU datasets (SLT and RMS). Ten utterances were synthesized for each voice. We applied the ABM involving SVD, and methods involving GV, and MS to improve the quality of the samples under large and limited data conditions. The baseline was HMM-based synthesized speech trained by 500 natural sentences. Under both conditions, we used five sentences to train our ABM using SVD. Under the limited-data condition, there were only five training utterances for method involving the MS post-filter. The GV cannot be trained with the small data. Under the large-data condition, 500 utterances were used for training the GV method, and the MS post-filter. Speech was sampled at 16 kHz, the frame shift was 5 ms, $S = 2$, $D = 49$, and $M = 4096$. Eleven listeners (10 non-native speakers and one native English speaker) with normal hearing condition participated. Each participant listened to pairs of samples. For each pair, they compared naturalness of stimuli on a two-point scale: 1 (A is more natural) and -1 (B is more natural) where natural speech was considered to be human-like speech. In Figs. 3. and 4, the HMM denotes HMM-based synthesized speech, and the SVD denotes the ABM involving SVD. Figure 3(a) shows that the preference score of the ABM involving SVD is higher than that for the method involving MS post-filter (denoted as MS). This is because the Gaussian distribution of each speech parameter cannot be well estimated from a small amount of data for the method involving MS post-filter. The results indicated that the ABM outperformed the other methods under limited-data condition. Figure 4(a) shows that the method involving MS post-filter is best under large-data condition. At the end of the experiments, participants were asked which factors contribute to their decisions. All participants agreed that speech with a buzzing sound and flat speech are not natural.

In the MRT, the intelligibility of the synthesized speech after applying the ABM was evaluated. Ten listeners (eight



(a) Preference test (b) MRT

Figure 3: Results of preference test and MRT with 95% confidence interval under limited-data condition



(a) Preference test (b) MRT

Figure 4: Results of preference test and MRT with 95% confidence interval under large-data condition

non-native speakers and two native English speakers) with normal hearing condition participated. As shown in Fig. 3(b), the intelligibility for the ABM involving SVD is equal to that of speech synthesized with the HMM. The results indicated that the intelligibility of the synthesized speech was preserved by the ABM involving SVD. The results under large-data condition in Fig. 4(b) had the same tendency.

4. Conclusions

A novel ABM was utilized on the MS of MCC sequences in improve the quality of HMM-based synthesized speech. Experimental results demonstrated that the performance of the technique is competitive with other techniques under a large-data condition and outperform other methods under a limited-data condition. Moreover, the experimental results also indicated that the ABM involving SVD can preserve the intelligibility of synthesized speech.

Using SVD sometimes results in negative values of naturalness which imply unrealistic subtraction of the intelligibility. In our next work, the use of non-negativity constrain in the ABM will be investigate.

Acknowledgment

This study was supported by the Grant-in-Aid for Scientific Research (A) (No. 25240026) and the A3 Foresight Program made available by the Japan Society for the Promotion of Science (JSPS).

References

[1] H. Zen, K. Tokuda and W. Black: Statistical parametric speech synthesis, Speech Comm., Vol. 51, No. 11, 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. 5, pp. 816–824, 2007.

[3] S. Takamichi, T. Toda, A. W. Black and S. Nakamura: Parameter generation algorithm considering modulation spectrum for HMM-based speech synthesis, Proc. 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. ICASSP, pp. 290–294, 2014.

[5] Y. Jiao, X. Xie, X. Na and M. Tu: Improving voice quality of HMM-based speech synthesis using voice conversion method, Proc. ICASSP, pp. 7964–7968, 2014.

[6] L. Chen, T. Raitio, C. Valentini-Botinhao, J. Yamagishi and Z. Ling: DNN-based stochastic postfilter for HMM-based speech synthesis, Proc. Interspeech, pp. 1954–1958, 2014.

[7] V. Popa, J. Nurminen and M. Gabbouj: A novel technique for voice conversion based on style and content decomposition with bilinear models, Proc. Interspeech, pp. 2655–2658, 2009.

[8] Y. Stylianou, O. Cappe and 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 and S. Imai: Mel-generalized cepstral analysis - A unified approach to speech spectral estimation, Proc. ICSLP, pp. 1043–1046, 1994.

[10] H. Kawahara, I. Masuda-Katsue and M. de Cheveigne: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous frequency-based F0 extraction: Possible role of a repetitive structure in sounds, J. Speech Commun., Vol. 27, pp. 187–207, 1999.

[11] H. Scheffe: An analysis of variance for paired comparisons, J. Am. Stat. Assoc., Vol. 37, pp. 381–400, 1952.