RISP
NCSP'16
International Workshop on
Nonlinear Circuits, Communications and Signal Processing
Honolulu, Hawaii, USA, March 6-9, 2016

A study on quality improvement of HMM-based synthesized voices using asymmetric bilinear model

Tuan Dinh-Anh, Daisuke Morikawa, Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292 Japan
Phone/FAX:+81-761-55-1111
E-mail: {tuan.dinh,morikawa,akagi}@jaist.ac.jp

Abstract

HMM-based synthesized voices are intelligible but not natural especially in limited data condition because of over-smoothing speech spectra in time-frequency domain. Improving naturalness is a critical problem of HMM-based speech synthesis. One solution for the problem is using voice conversion techniques to convert over-smoothed spectra to natural spectra. Although conventional conversion techniques transform speech spectra to natural ones to improve naturalness, they cause unexpected distortions on acceptable intelligibility of synthesized speech. The aim of the paper is to improve naturalness without violating intelligibility of synthesized speech employing an asymmetric bilinear model (ABM) to separate intelligibility and naturalness. In the paper, an ABM was implemented on modulation spectrum domain of Mel-cepstral coefficient (MCC) sequence to enhance fine structure of spectral parameter trajectory generated from HMMs. Subjective evaluations carried out on English data confirm that the achieved naturalness of proposed method is competitive with other methods in large data condition and outperform other methods in limited data condition. Moreover, modified rhyme test (MRT) shows that acceptable intelligibility of synthesized speech is well-preserved with proposed method.

1. Introduction

HMM-based speech synthesis is one of state-of-the-art techniques due to its flexibility and compact footprint [1]. 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 in data limited condition. There have been several attempts to overcome the over-smoothing effect. One approach is using objective evaluations of over-smoothing effect such as GV [2], and modulation spectrum (MS) [3], integrating them into parameter generation phase to obtain better speech parameters. The joint optimization of HMMs and objective evaluation typically does not have close-form solution. Another possible way to reduce the gap between 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 [5], [6] can be utilized as a mapping from natural speech to synthetic speech. The approach benefits from optimizing HMMs with close-form solution. Thus, voice conversion approach is employed to improve the naturalness.

With majority of voice conversion techniques, all spectra are transformed to improve naturalness. However, this often negatively affect intelligibility. To improve naturalness without violating intelligibility, an experiment was conducted to find out efficient acoustic feature strongly relating to naturalness. Then, decomposed features related to naturalness are converted to improve quality of re-synthesized speech, while other intelligibility-related features are preserved.

This paper is organized as follows. Section 2. reviews ABM [8] and shows the general framework of using ABM in proposed method. The problem of using asymmetric bilinear model is also addressed in the section. Section 2.1 tried to solve the problems. In section 3., we demonstrate the benefits of the proposed approach through listening test results. Finally, concluding remarks, including some potential future research direction, are presented in Section 4.

2. Naturalness improvement using asymmetric bilinear model

In an ABM [8], an observation y^{sc} from speaker s and phonetic class c can be represented as following:

$$y^{sc} = A^s b^c$$

with A^s denoting speaker information and b^c denoting phonetic information. In the paper, phonetic information b^c is assumed as intelligibility, and speaker information A^s is assumed as naturalness. In [8], observation y^{sc} is line spectral frequency (LSF) vector. LSF is a way to model speech spectra

with the emphasis of formants which is important for speaker characteristics. However, it is not clear whether formants are not enough for perceiving naturalness. Finding efficient kind of observations which suitable for naturalness improvement is very important.

2.1 Finding efficient acoustic feature

In the section, an experiment was carried out to find out the relationship between acoustic features and over-smoothing effect. There are three steps in the experiment:

1. Exchanging acoustic feature.
2. Comparing naturalness on listening test.
3. Finding efficient acoustic feature

In the first step, several kinds of acoustic features are prepared. They are fundamental frequency F0, peak related parameters such as linear prediction coefficient (LPC) w/wo residual power, LSF w/wo residual power, and perceptual linear prediction (PLP), fine structure related coefficients such as Mel-frequency cepstral coefficient (MFCC), MCC and cepstrum. To examine one kind of acoustic features, the feature sequences are exchanged between synthesized speech and natural speech of the same sentence. Exchanging acoustic feature means improving naturalness of synthesized speech and decreasing quality of natural speech. If quality of natural speech decreases and naturalness of synthesized speech increases a lot after exchanging, the kind of acoustic feature strongly relates to naturalness. In the experiment, one utterance for one natural speech sentence is synthesized by HTS[1]. The synthesized speech is aligned to its original speech with guide of labels. STRAIGHT vocoder was used to analyze speech. It decomposes speech into a spectral envelope, F0, and aperiodicity. The STRAIGHT-based speech parameters are further encoded into LPC, LSF, MFCC, MCC, PLP, and cepstrum. After the step, 20 stimuli are obtained as in Table 1.

In the second step, naturalness of obtained stimuli is compared using Scheffe's method of paired comparison [9] to sort them based on naturalness. There are six participants (non native English speakers with fluent English level). Each participant listened to 380 pairs of stimuli. With each pair, they compare naturalness of stimuli using five grades from -2 (A is more natural), 0 (comparable), +2 (B is more natural) in A-B comparison.

Lastly, the efficient acoustic feature is decided by looking for the one that improves naturalness of synthesized speech the most. Experimental results in Figure 1 indicate that exchanging MCC values improves naturalness of synthesized speech the most (I2 to G2). Exchanging LSF does not significantly improve naturalness (I2 to E2). In frequency domain, fine structure is more important than formant in perceiving naturalness. MCC is the most suitable acoustic feature in improving naturalness.

Table 1: Stimuli in experience

Stimuli	Meaning	Stimuli	Meaning
A1	Natural speech after exchaning (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 MCC can represent the fine structure in frequency domain, it cannot represent the dynamics of spectra in time domain. In recent years, modulation spectrum becomes a popular concept in capturing the fine structure of speech spectra in time domain. In the paper, modulation spectrum (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, is utilized to capture the over-smoothing effect in both time-frequency domain 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:

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

where f is the modulation frequency bin, defined by the number of points in the Fourier analysis. The number of points in the FFT must be greater than the maximum number of frames T of an utterance. The MS of each utterance is calculated for each coefficient. Using ABM, MS of synthetic trajectories is modified to be closer to the modulation characteristics of natural speech.

2.2 Scheme to employ ABM in naturalness improvement

In the section, the process of applying ABM in naturalness improvement is described. The process consists of 3 major steps shown in Figure 2:

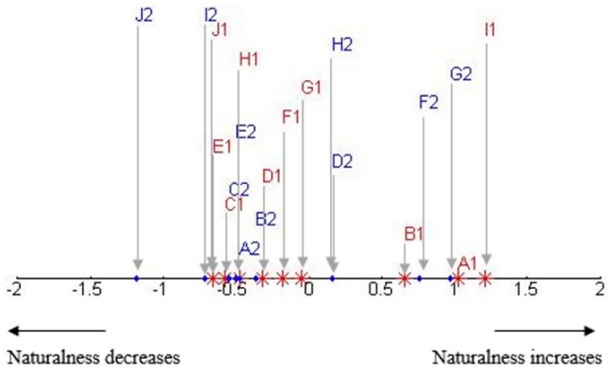

Figure 1: Result of pair comparison test

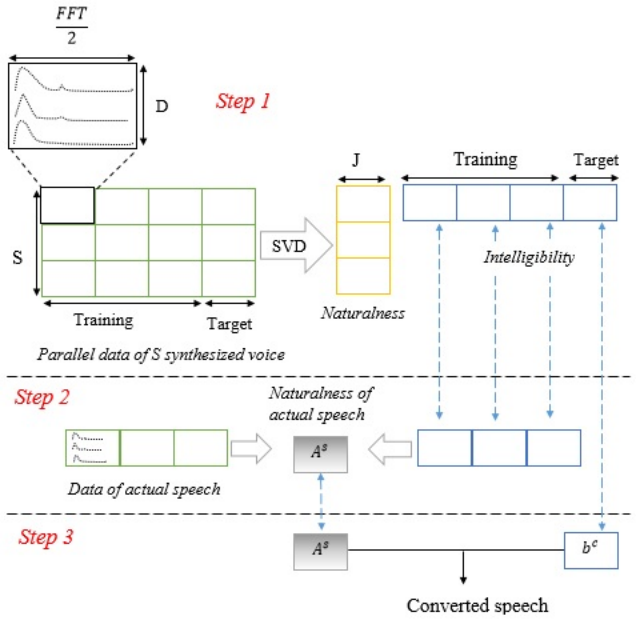

Figure 2: Scheme of applying ABM

1. Separating naturalness and intelligibility.
2. Obtaining naturalness of actual speech.
3. Reconstructing speech.

The goal of step 1 is to obtain acceptable intelligibility from parallel data of synthesized voices to preserve it. Naturalness factor and intelligibility factor are factored from the data using singular value decomposition (SVD). Each cell of the parallel data of synthesized speech (PDSS) is MS of one utterance. In Figure 2, $\frac{FFT}{2}$ denotes half of length of FFT for MS, D is order of MCC, and S denotes number of HTS [1] (S $\geq$ 2). There are 1 target utterance. Number of training sentences can be as small as 5. J is model dimensionality chosen as $J = S \times D$. Because SVD results in 3 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 PDSS, the variation of naturalness is presented in columns, and the variation of intelligibility is presented in rows. The columns of naturalness matrix summarize PDSS's vertical structure associating to naturalness. Likewise the rows of intelligibility matrix do so for the horizontal structure of PDSS.

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

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

Lastly, naturalness of actual speech A^s and intelligibility of synthesized speech are combined to obtained an improved version of synthesized speech. Intelligibility is preserved even in synthesized speech.

3. Evaluation and Discussion

In the section, naturalness and intelligibility of proposed method are evaluated using preference test and MRT. In preference test, proposed method is compared with others improvement methods such as GV method [2], and MS method [4]. Two HMM-based synthesized voices are trained using 2 CMU datasets (SLT and RMS). Ten utterances are synthesized for each voice. We apply proposed method denoted as SVD, GV method, and MS method to improve the quality of the samples in large data condition and limited data condition. In limited data condition, there are only 5 training utterances for each synthesized voice. GV can not be trained with the small data. In large data condition, 500 utterances are used for GV method, and MS method. Speech is sampled at 16 kHz. Frameshift is 5ms. $S = 2$. $D = 49$. $\frac{FFT}{2} = 2098$. There are 11 participants (10 non-native and 1 native English speakers). Each participant listen to pairs of samples. With each pair, participants are asked which sample is more natural. Natural speech is explained as human-like speech. Figure 3 shows that proposed method is competitive with other methods in large data condition. In Figure 4, we can see that the score of proposed method increases over MS. This result indicates that proposed method outperforms other methods in limited data condition. HMM denotes HMM-based synthesized speech. SVD denotes proposed method. At the end of the experiments, participants were asked what factors contribute to their decisions. All participants agreed that speech with buzzy sound and flat speech is not natural.

In MRT, intelligibility of synthesized speech after applying proposed method is evaluated. There are 10 participants (8 non-native and 2 native English speakers). In Figure 5, the correctness of proposed method is equal to synthesized

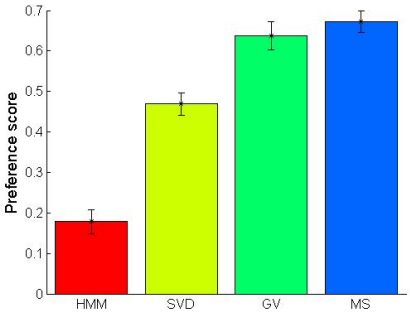

Figure 3: Preference scores in large data condition with 95% confidence interval

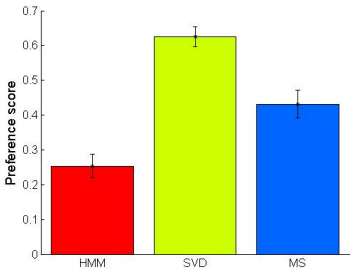

Figure 4: Preference scores in limited data condition with 95% confidence interval

speech HMM. The result indicates that indicate that intelligibility of synthesized speech is preserved with proposed method (SVD). Results in large data condition have the same tendency.

4. Conclusions

A novel ABM was utilized on MS of MCC sequence in the task of quality improvement of HMM-based synthesised speech. Experimental results demonstrated that the performance of the technique is competitive with other techniques in large data condition and outperform other methods in limited data condition. Moreover, experimental results also indicate that the proposed method can preserve acceptable intelligibility of synthesized speech.

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

References

[1] H. Zen, K. Tokuda and W. Black, Statistical parametric speech synthesis · Speech Comm., vol. 51, no. 11, pp. 1039-1064, 2009.

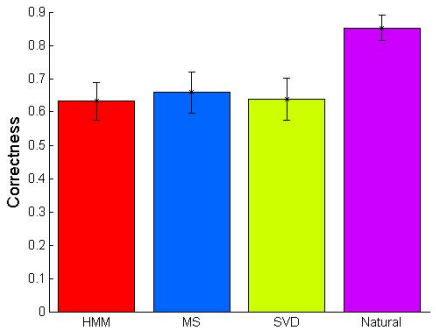

Figure 5: MRT correctness in limited data condition with 95% confidence interval

[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] Shinnosuke Takamichi, Tomoki Toda, Alan W. Black, and Satoshi Nakamura. Parameter generation algorithm considering modulation spectrum for HMM-based speech synthesis. In ICASSP, pp. 4210-4214, 2015.

[4] Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, and Satoshi Nakamura. A post-filter to modify the modulation spectrum in HMM-based speech synthesis, In ICASSP, pp. 290-294, 2014.

[5] Yishan Jiao, Xiang Xie, Xingyu Na, Minh Tu; Improving voice quality of HMM-based speech synthesis using voice conversion method; in ICASSP, pp. 7964-7968, 2014.

[6] Ling-Hui Chen, Tuomo Raitio, Cassia Valentini-Botinhao, Junichi Yamagishi, Zhen-Hua Ling. DNN-based stochastic postfilter for HMM-based speech synthesis; in Interspeech, pp. 1954-1958, 2014.

[7] Joshua Tenenbaum, William Freeman; separating style and content with bilinear models; Neural Computation; pp. 1247-1283, 2000.

[8] Victor Popa, Jani Nurminen, Moncef Gabbouj; A novel technique for voice conversion based on style and content decomposition with bilinear models. In Interspeech, pp 2655-2658, 2009.

[9] H. Scheffe, An analysis of variance for paired comparisons, Journal of the American Statistical Association, vol. 37, pp. 381-400, 1952.