ENHANCEMENT OF SPECTRAL CLARITY FOR HMM-BASED TEXT-TO-SPEECH SYSTEMS

Young-Sun Joo, Chi-Sang Jung, and Hong-Goo Kang

School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea

{disfruta, jtoctos}@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr

ABSTRACT

This paper proposes a method to enhance the spectral clarity of hidden Markov model (HMM)-based text-to-speech (TTS) systems. A simple way of enhancing spectral clarity is increasing the order of spectral parameters in the speech analysis/synthesis stage, but the method has an inherent statistical modeling problem. The proposed algorithm takes a low-to-high-order spectral parameter mapping approach that adopts low-order parameters for HMM training but does high-order parameters for the actual synthesis step. Various ways of mapping criterion to find appropriate high-order parameters are investigated to further enhance the quality of synthesized speech. Performance evaluation results verify the superiority of the proposed method compared to the conventional one.

Index Terms— HMM-based TTS, spectral clarity, statistical modeling, low-to-high-order spectral parameter mapping

1. INTRODUCTION

The HMM-based TTS system has been popularly studied because of its reasonable quality and easy implementation [1, 2]. It is well known that the naturalness of synthesized speech is improved by adopting efficient excitation and spectrum models. As an advanced excitation model, a mixed excitation and waveform interpolation (WI) models are applied to HMM-based TTS systems [3, 4]. For a spectrum model, line spectral frequency (LSF) [5] is popularly used because it is easy to check the stability condition. Note that the clarity of synthesized speech can be enhanced by using a high-order spectral parameter. However, the use of high-order spectral parameters brings a stability problem in the HMM training and parameter generation stage. As the order of LSF is increased, the gaps between consecutive coefficients become small. Therefore, it is not easy to increase modeling accuracy while keeping the constraint of stability condition. In other words, the stability of spectral parameters and the clarity of synthesized speech have a trade-off relationship.

The authors would like to thank Microsoft Research Asia in Beijing, China for funding this project through the Ministry of Knowledge Economy of South Korea.

In order to resolve the stability problem of generated LSFs, a model training criterion of having minimum generation error (MGE) is introduced into the training process [6]. In [7], a parameter generation algorithm is also proposed which utilizes the fact that the stability problem of LSF needs to be solved in the generated LSF domain rather than in the model training procedure. Although these methods reduce the stability problem somehow, the generated LSFs are not clear and detail enough because they are artificially synthesized from trained HMMs.

This paper proposes a low-to-high-order spectral parameter mapping method to represent the details of spectrum. The proposed method substitutes a low-order spectral parameter generated from trained HMMs with a high-order spectral parameter extracted from original speech signals during the synthesis stage. Since the proposed method trains HMMs using a low-order spectral parameter but synthesizes the signal using a high-order spectral parameter, the spectral clarity is enhanced while avoiding the stability problem.

A well-designed mapping criterion from the low-order parameters to the high-order ones is very important to obtain an appropriate spectrum sequence, therefore the proposed mapping criterion considers both static and dynamic characteristics of spectral parameters. The criterion consists of two types of distances; distances between a low-order spectral parameter generated from trained HMMs and low-order spectral parameter candidates in the pre-stored repository, and between high-order spectral parameters obtained in previous frames and high-order candidates in the repository. Since the low-order spectral parameter is coupled with the high-order one extracted at the same frame of the input speech signal, the optimal high-order spectral parameter is simply obtained by taking a table look-up process.

The results of the performance evaluation prove the superiority of the proposed method in terms of both subjective and objective measures.

2. STABILITY PROBLEM IN HIGH-ORDER SPECTRAL PARAMETER TRAINING

Although the synthesized speech by adopting high-order spectral parameters improves intelligibility, there remains a

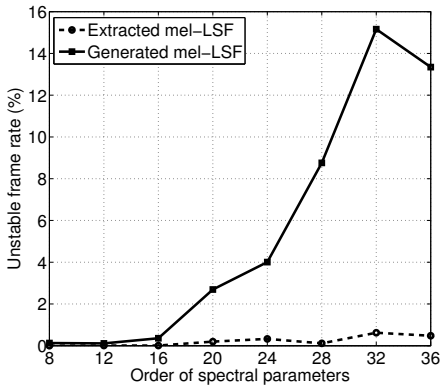

Fig. 1. Unstable frame rate of extracted and generated mel-LSFs on fifty test sentences.

statistical modeling issue if the order of spectral parameters increases or the amount of database for training is insufficient. Fig. 1 shows the unstable frame rate of spectral parameters extracted directly from original speech waveform and generated from the trained HMMs. A frame is classified as an unstable frame if the consecutive coefficients of LSF are reversed or extremely close [8]. The unstable frame rate of generated spectral parameters increases as the order of parameter increases, while there is no significant difference in the spectral parameters directly extracted from original speech waveform. The core idea of the proposed method is to use high-order spectral parameters extracted from original speech for synthesis but to use low-order ones for HMM training. It needs a spectral parameter mapping process that is explained in the following section.

3. PROPOSED HMM-BASED TTS SYSTEM

3.1. Overview of the proposed system

Fig. 2 depicts a block diagram of the proposed method. The main strategy of the proposed method depicted in the box is to substitute the low-order spectral parameters generated from the trained HMMs with the high-order spectral parameters obtained from the original speech waveform. In the mapping process step, it is very important to determine a criterion of choosing the most appropriate parameter. The detailed system flow is explained as follows.

The training procedure consists of two independent parts. One is the context-dependent HMMs training with acoustic parameters such as $f0$ and LSF. This part is identical to the conventional training process [2]. The other is to create a repository that consists of pairs of low and high-order spectral parameters needed for the mapping process in the synthesis stage. The pair of low and high-order LSFs is extracted from the original speech waveform at the same frame data, and the LSFs are linked together to have identical indices.

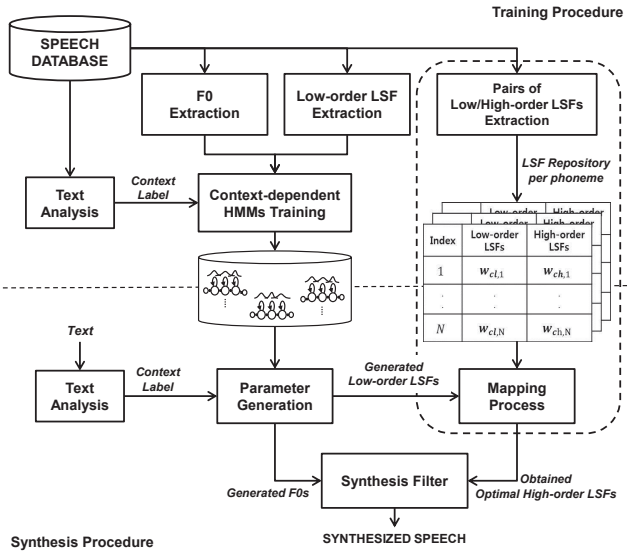

Fig. 2. A block diagram of the proposed HMM-based TTS system.

The synthesis procedure is similar to the conventional approach except for using high-order spectral parameters obtained by the mapping process. At first, the low-order LSF is generated by the trained HMMs using a parameter generation algorithm [9]. Then, the optimal high-order LSF is chosen from the LSF repository by performing a low-to-high-order spectral parameter mapping process. As a final step, a post-processing filter is applied to the sequence of obtained optimal high-order LSFs in order to smooth LSF trajectory.

To minimize the complexity of the mapping process, a concept of partial search is also adopted, which defines multiple number of centroid vectors to restrict the search space. The centroid vectors of the low-order LSF repository are obtained in the training procedure by applying the LBG algorithm [10]. Then, the high-order LSFs corresponding to each search space of low-order LSF repository are deposited to have same indices.

3.2. Proposed low-to-high-order spectral parameter mapping algorithm

Fig. 3 shows the detailed procedure of the proposed low-to-high-order spectral parameter mapping process. The optimal high-order LSF is determined by the proposed mapping criterion using the low-order LSF generated from the trained HMMs and the optimal high-order LSFs determined at previous frames. The mapping criterion is defined as follows :

$$\hat{\boldsymbol{w}}_{oh}(n) = \arg\min_{\boldsymbol{w}_c} \{ \varepsilon\left(\boldsymbol{w}_c;\, \boldsymbol{w}_{gl}(n),\, \hat{\boldsymbol{w}}_{oh}(n-1), \ldots, \hat{\boldsymbol{w}}_{oh}(n-L)\right)\}, \quad (1)$$

where $\hat{\boldsymbol{w}}_{oh}(n) = [w_{oh,1}(n), \ldots, w_{oh,N}(n)]^\top$ is the optimal high-order LSF at the current time n, and $\boldsymbol{w}_{gl}(n) =$

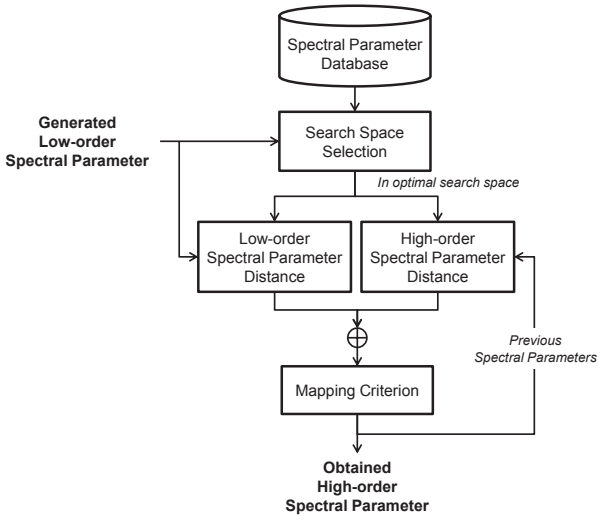

Fig. 3. A block diagram of the proposed mapping process.

$[\omega_{gl,1}(n), \ldots, \omega_{gl,M}(n)]^\top$ is the low-order LSF generated by the trained HMMs. M and N denote the order of the low and high-order LSFs, respectively. $\boldsymbol{w}_c = \left[\boldsymbol{w}_{cl}{}^\top, \boldsymbol{w}_{ch}{}^\top\right]^\top$ is a pair of low and high-order LSF candidates. Given the generated low-order LSF $\boldsymbol{w}_{gl}(n)$ and L optimal high-order LSFs obtained at the previous frames $\hat{\boldsymbol{w}}_{oh}(n-\tau)|_{\tau=1,2,\ldots,L}$, the optimal high-order LSF $\hat{\boldsymbol{w}}_{oh}(n)$ is determined by minimizing the cost function $\varepsilon(\cdot)$ with respect to $\boldsymbol{w}_c$. The cost function $\varepsilon(\cdot)$ is defined as the sum of two distances as follows :

$$\begin{aligned}&\varepsilon(\boldsymbol{w}_c; \boldsymbol{w}_{gl}(n), \hat{\boldsymbol{w}}_{oh}(n-1), \ldots, \hat{\boldsymbol{w}}_{oh}(n-L)) \\ &\quad = D(\boldsymbol{w}_{gl}(n), \boldsymbol{w}_{cl}) + D\left(\sum_{\tau=1}^{L} \alpha_\tau \hat{\boldsymbol{w}}_{oh}(n-\tau), \boldsymbol{w}_{ch}\right).\end{aligned} \tag{2}$$

The second term is included to represent the impact of the dynamic characteristic of spectral parameters. Note that the synthesized speech quality is deteriorated by inappropriate spectral movements if only the independent frame-by-frame-based mapping process is applied. In order to generate the natural trajectory of LSFs, the second term provides the similarity between the high-order LSF candidates $\boldsymbol{w}_{ch}$ and a weighted sum of the optimal high-order LSFs obtained at the previous frames $\hat{\boldsymbol{w}}_{oh}(n-\tau)$. The dynamic characteristic of high-order LSF is also involved because the actual synthesis stage uses high-order LSFs.

The frame weight α_τ in Eq. (2) is determined by considering the statistical characteristic of speech signal. Since the spectral characteristic does not vary rapidly, we assume that the trajectory of LSF coefficients can be modeled by a weighted linear combination. The optimal weight is obtained by solving the Wiener-Hopf equations using large amounts of data [11].

During the mapping process, it is inefficient to use large amounts of repository because its computational complexity is huge. To reduce the complexity, a partial search region is pre-selected in the searching process. The search region is selected to minimize the weighted distance between the generated low-order LSF and low-order partial centroid vectors.

$$\hat{\boldsymbol{c}}_l = \arg\min_{\boldsymbol{c}_l} D(\boldsymbol{\omega}_{gl}(n), \boldsymbol{c}_l), \qquad \forall \boldsymbol{c}_l \in \mathbf{S}_l, \tag{3}$$

where $\boldsymbol{c}_l = [c_{l,1}, \ldots, c_{l,M}]^\top$ is the low-order partial centroid vector that indicates a partial search space. $\mathbf{S}_l$ denotes the entire search space consisting of $\boldsymbol{c}_l$.

The distance between two LSF vectors is measured by following equation :

$$D(\boldsymbol{\omega}_x, \boldsymbol{\omega}_y) = \sqrt{\frac{1}{P}\sum_{i=1}^{P} \frac{v_i}{V}(\omega_{x,i} - \omega_{y,i})^2}, \tag{4}$$

where v_i denotes a weighting factor for the i-th coefficient of the P-order LSF and $V = \sum_{j=1}^{P} v_j$ is a normalization factor. It is an intrinsic characteristic of LSF parameters that neighboring coefficients are related to local spectral peak. Hence, the higher weight is given to the LSF coefficient which is close to its adjacent one. The weight is determined by the inverse harmonic mean (IHM) weighting function [12].

4. PERFORMANCE EVALUATION

To evaluate the performance of the proposed approach, a Korean TTS system is constructed based on the HTS toolkit [2]. For training, around three thousand speech waveforms are recorded by a male speaker with a sampling rate of 16kHz. The excitation is modeled by a pulse-or-noise (PoN) model. The fundamental frequency, $f0$, and 16-order mel-LSFs are used as acoustic parameters. The frame length is set to 25ms and the frame shift is set to 5ms. To reduce the complexity of the low-to-high-order spectral parameter mapping process, each phoneme-dependent LSF repository has 512 partial centroid vectors. The repository consists of around thirty thousand pairs of 16 and 36-order mel-LSFs.

An experiment is carried out to evaluate the feasibility of the proposed mapping criterion. Since the appropriate sequence of speech spectrum should be varied smoothly in time, we measure the spectral distance (SD) between consecutive frames. The SD between the n-th and $(n+1)$-th frames, D_n, is defined (in decibels) as follows

$$D_n^2 = \frac{1}{f_s}\int_0^{f_s} \left[10\log_{10}(P_{n+1}(f)) - 10\log_{10}(P_n(f))\right]^2 df, \tag{5}$$

where f_s is sampling frequency in Hz. $P_n(f)$ is the LPC power spectra of the n-th frame. The SD is also measured using fifty test sentences. The *w/o dynamic characteristics* is the one obtained by the first term of the mapping criterion given in Eq. (2), and the *w/ dynamic characteristics* is the one

Table 1. Average SD of 36-order mel-LSFs sequences

Feature types	Avg. SD (dB)
Natural speech	2.10
w/o dynamic characteristics	4.54
w/ dynamic characteristics	3.95

Table 2. Unstable frame rates of 36-order mel-LSFs

Feature types	Unstable frame rate (%)
Conventional method	13.34
Proposed method	0.02

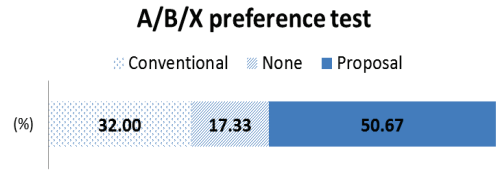

Fig. 4. A/B/X preference test result

obtained by the full mapping criterion. Table 1 shows that although the average SD of the *w/ dynamic characteristics* is not better than that of the *natural speech*, it is much better than that of *w/o dynamic characteristics*.

To evaluate the performance of the trainability, the unstable frame rate is calculated with fifty test sentences. The conventional method with 36-order mel-LSF is used for comparison. As shown in Table 2, the unstable frame rate of the proposed method is very low because the high-order mel-LSF is obtained from the LSF repository of which the data is extracted from original speech waveform.

An A/B/X preference test is also conducted to evaluate subjective quality. Fifteen randomized utterances not included in the training set are used for the test set. Ten experts in the speech signal processing field are participated in the test. As shown in Fig. 4, it is clear that the quality of synthesized speech by the proposed method is superior to that of the conventional one that uses 36-order mel-LSF in both training and synthesis procedures.

5. CONCLUSIONS

The synthesized speech quality of HMM-based TTS systems has been improved by enhancing spectral clarity. The proposed algorithm adopted a low-to-high-order spectral parameter mapping strategy that substitutes low-order spectral parameters generated from trained HMMs with high-order ones extracted from original speech. Experiment results verified the superiority of the proposed method compared to the conventional method.

6. REFERENCES

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH*, 1999, pp. 2374–2350.

[2] K. Tokuda, H. Zen, J. Yamagishi, A. W. Black, T. Masuko, S. Sako, T. Toda, T. Nose, and K. Oura, HMM-based speech synthesis system (HTS). [Online]. Available: http://hts.sp.nitech.ac.jp

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proc. EUROSPEECH*, 2001, pp. 2263–2266.

[4] C.-S. Jung, Y.-S. Joo, and H.-G. Kang, "Waveform Interpolation-Based Speech Analysis/Synthesis for HMM-Based TTS Systems," *IEEE Signal Process. Lett.*, vol. 19, no. 12, pp. 809–812, 2012.

[5] F. Soong. and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. ICASSP*, 1984, pp. 37–40.

[6] M. Lei, Z.-H. Ling, and L.-R. Dai, "Preserve ordering property of generated LSPs for minimum generation error training in HMM-based speech synthesis," in *Proc. ICASSP*, 2011, pp. 4712–4715.

[7] S. Qian, H. Wang, W. Pei, and K. Wang, "Parameter Generation Considering LSP Ordering Property for HMM-Based Speech Synthesis," *IEEE Signal Process. Lett.*, vol. 19, no. 8, pp. 467–470, 2012.

[8] N. Sugamura and F. Itakura, "Speech analysis and synthesis methods developed at ECL in NTT–From LPC to LSP–," *Speech Commun.*, vol. 5, no. 2, pp. 199–215, 1986.

[9] T. Tomoki and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inform. and Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[10] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, pp. 84–95, 1980.

[11] S. O. Haykin, *Adaptive Filter Theory*, 4th ed. Prentice Hall, 2001.

[12] R. Laroia, N. Phamdo, and N. Farvardin, "Robust and efficient quantization of speech LSP parameters using structured vector quantizers," in *Proc. ICASSP*, 1991, pp. 641–644.