Optimum MVF Estimation-Based Two-Band Excitation for HMM-Based Speech Synthesis

Seungho Han, Sangbae Jeong, and Minsoo Hahn

ABSTRACT—The optimum maximum voiced frequency (MVF) estimation-based two-band excitation for hidden Markov model-based speech synthesis is presented. An analysis-by-synthesis scheme is adopted for the MVF estimation which leads to the minimum spectral distortion of synthesized speech. Experimental results show that the proposed method significantly improves synthetic speech quality.

Keywords—HTS, HMM-based speech synthesis, speech synthesis, TTS.

I. Introduction

A hidden Markov model (HMM)-based speech synthesis system (HTS) [1] is suitable for the speech interface of hand-held devices such as PDAs, cellular phones, and car navigation systems because it requires less memory and lower computational power than other systems. In the training procedure, the spectral parameter, the excitation parameter, and the duration of a speech unit are modeled by context-dependent HMMs. In the synthesis procedure, speech is simply produced by mel log spectrum approximation (MLSA) filtering [2] with the parameters generated from the trained HMMs.

For excitation signal generation, the traditional excitation (TE) model is adopted in conventional HTSs [1]. Only the fundamental frequency is used as the excitation parameter. Since the excitation signal in the TE model is generated by either a periodic pulse train for voiced speech or a random noise sequence for unvoiced speech, the synthesized speech has typical artifacts known from vocoders. Yoshimura and others proposed mixed excitation (ME) for the HTS to overcome this problem [3]. Because the ME still has problems, two-band excitation (TBE) for the HTS was proposed [4] to improve synthetic speech quality. In TBE, excitation bands are divided into lower and higher frequency bands by the maximum voiced frequency (MVF). The excitation signal of the lower band is generated by a periodic pulse train, while that of the higher band is generated by a white noise sequence. Therefore, it is very important to estimate the MVF more accurately because synthesized speech quality largely depends on the MVF. This letter proposes an optimum MVF estimation-based TBE for the HTS.

II. Proposed Optimum MVF Estimation

In the TBE model [4], the MVF, and the fundamental frequency (F0) are used as excitation parameters. To determine the MVF, we chose an analysis-by-synthesis (ABS) scheme. An ABS scheme is broadly used as a speech coding technique. In the ABS scheme, a speech signal is represented in the form of the speech production model expressed by some parameters such as linear predictive coefficients and excitation sequences. By changing the parameters, different speech can be produced. A trial and error procedure is usually applied to estimate the optimum parameters. The estimated parameters produce the synthesized speech signal, which matches the real speech signal with minimum error. Fortunately, the ABS scheme can be easily utilized to estimate the MVF in the TBE model.

Figure 1 illustrates the procedure of the proposed MVF estimation. The MVF is estimated in 2 steps. At the first step,

Manuscript received Mar. 16, 2009; revised Apr. 24, 2009; accepted May 6, 2009.

This work was supported by the IT R&D program of MIC/IITA, Rep. of Korea [2006-S-036-01, Development of Large Vocabulary/Interactive Distributed/Embedded VUI for New Growth Engine Industries].

Seungho Han (phone: +82 42 350 6206, email: space0128@kaist.ac.kr) and Minsoo Hahn (email: mshahn@ee.kaist.ac.kr) are with the Department of Information and Communications Engineering, KAIST, Daejeon, Rep. of Korea.

Sangbae Jeong (email: jeongsb@gnu.ac.kr) is with the Department of Electronics Engineering (ERI), Gyeongsang National University, Jinju, Rep. of Korea.

doi:10.4218/etrij.09.0209.0112

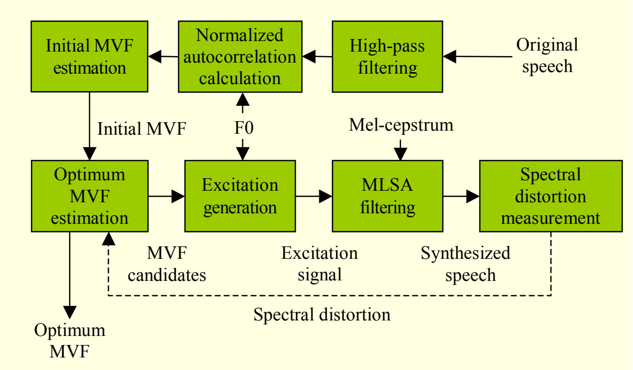

Fig. 1. Proposed MVF estimation procedure.

the initial MVF is roughly estimated by the normalized auto-correlation of the high-pass filtered speech signal $s_{\mathrm{HB}}^{f_c}(n)$ with a cutoff frequency f_c [4]. The auto-correlation $R^{f_c}(\tau)$ is calculated as

$$R^{f_c}(\tau)=\frac{\sum_{n=0}^{N-1} s_{\mathrm{HB}}^{f_c}(n)\, s_{\mathrm{HB}}^{f_c}(n+\tau)}{\sqrt{\sum_{n=0}^{N-1}\left\{s_{\mathrm{HB}}^{f_c}(n)\right\}^2 \sum_{n=0}^{N-1}\left\{s_{\mathrm{HB}}^{f_c}(n+\tau)\right\}^2}}, \tag{1}$$

where τ is the pitch period in a given analysis speech frame, and N is the pitch analysis window size. If the cutoff frequency is smaller than the MVF, $s_{\mathrm{HB}}^{f_c}(n)$ is periodic, and $R^{f_c}(\tau)$ would be close to 1. In the opposite case, $s_{\mathrm{HB}}^{f_c}(n)$ is aperiodic, and $R^{f_c}(\tau)$ would be close to 0. Thus, the cutoff frequency at which $R^{f_c}(\tau)$ becomes less than 0.5 is adopted as the initial MVF. High-pass filters are designed with various cutoff frequencies from 0.5 kHz to 7.5 kHz with an increment of 0.5 kHz.

At the second step, the optimum MVF is searched by the ABS scheme to minimize spectral distortion. First, the excitation signal is generated with the initial MVF. Next, the speech is synthesized by the MLSA filter with this excitation signal and the extracted mel-cepstrum. Then, the quality of the synthesized speech is measured by spectral distortion. In our experiment, as a distortion measure like the concatenation cost in the unit selection method of the corpus-based concatenative speech synthesis, the symmetric Kullback-Leibler distance (SKLD) between the normalized power spectra of adjacent sub-frames near the frame boundary is calculated as

$$D_{\mathrm{SKL}}=\sum_{k=0}^{N-1}\left(S_i(k)-S_{i+1}(k)\log\frac{S_i(k)}{S_{i+1}(k)}\right), \tag{2}$$

where $S_i(k)$ is the normalized power spectrum in a sub-frame, and i and k are the frame and the frequency bin index, respectively. These procedures can be repeated for candidate MVFs around the initial MVF. Then, the MVF which produces the synthesized speech with the minimum spectral distortion is finally determined as the optimum MVF. The calculation time does not matter because the estimation procedure is an off-line process which is used once in the training procedure.

III. Experiments and Results

As the baseline system, a Korean HTS [4] was used. Eighteen mel-cepstrum coefficients and the 0th-order coefficient were extracted, and a 57th-order vector was constructed by concatenating their first and second derivatives and using them as the spectral parameter. The excitation signal was modeled by the proposed optimum MVF estimation-based TBE (OM-TBE). In our experiments, for the efficiency of searching, an initial MVF and 4 MVFs around the initial one were used as the candidate MVFs. The analysis speech frame size was 80, and the sub-frame size was 20.

To verify the effectiveness of the ABS scheme, the excitation signal was also generated by the initial MVF-based TBE (IM-TBE). In addition, the TE and the ME models were compared with the proposed model. To evaluate the synthesized speech quality objectively, the log spectral distance (LSD) and the normalized power spectrum SKLD were calculated between the original and the synthesized speech in both the training and synthesis procedures. The distances were measured for all trained sentences. After the original speech and synthesized speech were time aligned by a dynamic time warping technique, the distances were calculated and averaged for each sentence. In the synthesis procedure, we evaluated the objective spectral distances and carried out subjective mean opinion score (MOS) listening and preference tests [5]. Four female listeners and four male listeners participated in the MOS and preference tests. All listeners were native Koreans in their 20s. The listeners tested eight sentences, among which four sentences were from the training sentences, and the others were randomly selected from newspaper scripts. In the MOS tests, the listeners evaluated the quality of speech synthesized by the TE-based, ME-based, IM-TBE-based, and OM-TBE-based HTSs. In the preference tests, the listeners evaluated pairs of speech generated from the IM-TBE-based and OM-TBE-based HTSs. The order of the speech was mixed for each evaluation. The same procedure was applied to the preference test of the synthesized speech from the ME-based and OM-TBE-based HTSs.

As seen in Table 1, in terms of speech quality, the proposed OM-TBE-based HTS clearly outperforms the others in both the training and synthesis procedures. The MOS test results in Table 2 also confirm that the proposed method is superior to the others. The average MOS of the proposed OM-TBE-based

Table 1. Distortion evaluation results.

	Training procedure		Synthesis procedure	
	LSD	SKLD	LSD	SKLD
TE	76.63	232.21	97.94	346.37
ME	74.10	214.30	97.57	341.21
IM-TBE	73.73	228.43	97.49	345.34
OM-TBE	73.20	225.09	94.39	340.31

Table 2. MOS test results.

	Overall sentences	Training sentences	Non-training sentences
TE	2.50	2.66	2.34
ME	2.70	2.75	2.66
IM-TBE	2.90	2.97	2.83
OM-TBE	3.09	3.09	3.09

HTS is 3.09. The gains are 0.59, 0.39, and 0.19 compared with the TE-based, ME-based, and IM-TBE-based HTSs, respectively. In particular, the synthetic speech quality is significantly improved for the sentences not used for the training, that is, non-training sentences. The MOS difference between the training and non-training sentences is not noticeable for the proposed method, unlike the other methods. From the preference test results, we can also conclude that the OM-TBE-based HTS is preferred over the others. About 83% and 89% of listeners report better or equal quality for speech synthesized by the proposed OM-TBE-based HTS compared with those synthesized by the ME- and IM-TBE-based HTSs, respectively.

The overall required memory for the trained HMM and decision tree data in the conventional TE-based HTS is 1.4 MB. The needed memories in the ME- and the TBE-based HTSs are 2.12 MB and 1.54 MB, respectively. Thus, the memory increase is almost negligible for the proposed method.

IV. Conclusion

We proposed a TBE for the HTS in which the ABS scheme is adopted to estimate the optimum MVF, which leads to the minimum spectral distortion of synthesized speech. As our test results demonstrated, considerable synthetic speech quality improvement was achieved by the proposed excitation method. An MOS gain of 0.59 was obtained while the required memory was increased by only 10% compared with the conventional TE-based HTS. As a future work, we plan to evaluate other distance measuring methods for better MVF estimation.

References

[1] T. Yoshimura et al., "Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis," *Proc. EUROSPEECH*, vol. 5, 1999, pp. 2347-2350.

[2] T. Fukada et al., "An Adaptive Algorithm for Mel-Cepstral Analysis of Speech," *Proc. ICASSP*, vol. 1, 1992, pp. 137-140.

[3] T. Yoshimura et al., "Mixed Excitation for HMM-Based Speech Synthesis," *Proc. EUROSPEECH*, vol. 3, 2001, pp. 2263-2266.

[4] S. Kim, J. Kim, and M. Hahn, "HMM-Based Korean Speech Synthesis System for Hand-Held Devices," *IEEE Trans. Consum. Electron.*, vol. 52, no. 4, Nov. 2006, pp. 1384-1390.

[5] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall, New Jersey, 2001.