Enhanced Maximum Voiced Frequency Estimation Scheme for HTS Using Two-Band Excitation Model

Jihoon Park and Minsoo Hahn

In a hidden Markov model–based speech synthesis system using a two-band excitation model, a maximum voiced frequency (MVF) is the most important feature as an excitation parameter because the synthetic speech quality depends on the MVF. This paper proposes an enhanced MVF estimation scheme based on a peak picking method. In the proposed scheme, both local peaks and peak lobes are picked from the spectrum of a linear predictive residual signal. The average of the normalized distances of local peaks and peak lobes is calculated and utilized as a feature to estimate an MVF. Experimental results of both objective and subjective tests show that the proposed scheme improves the synthetic speech quality compared with that of a conventional one in a mobile device as well as a PC environment.

Keywords: Speech synthesis, HTS, two-band excitation, maximum voiced frequency, harmonic peak.

I. Introduction

In human–device interaction via speech, text-to-speech synthesis (TTS) synthesizes speech from a message of a device. In certain situations, a user may not be able to see or control a device (for example, a blind person approaching a vehicle), in which case, TTS can prove to be a useful, helpful technology.

A corpus-based unit concatenating speech synthesis is the mainstream of TTS [1]. A corpus-based TTS selects appropriate units from a large corpus database and concatenates them. However, for a corpus-based TTS, large amounts of unit data is required to obtain natural, high-quality speech.

In recent decades, mobile devices, such as smartphones, e-book readers, and car navigation systems, have been developed and diffused rapidly. Accordingly, an embedded TTS has long been a requirement of mobile devices. An embedded TTS can be used in several applications, such as in a short message, in an e-mail, in an e-book, or in a car navigation system. However, because mobile devices are limited in terms of memory and computation power, a corpus-based TTS is unsuitable for an embedded TTS. On the contrary, a hidden Markov model (HMM)–based speech synthesis (HTS) developed by Tokuda and others is suitable for an embedded TTS [2]–[6]. A HTS uses context-dependent HMMs to model parameters extracted from a speech database, and speech is then generated from the now *trained* context-dependent HMMs.

The synthesis engine of an HTS requires small memory size and low computation power. An HTS system consists of training and synthesis parts. In the training part, the spectral

Manuscript received Feb. 9, 2015; revised June 22, 2015; accepted July 8, 2015.

Jihoon Park (batho2n@kaist.ac.kr) is with the Department of Research, Center for Integrated Smart Sensors, Daejeon, Rep. of Korea.

Minsoo Hahn (corresponding author, mshahn@ee.kaist.ac.kr) is with the Department of Electrical Engineering, KAIST, Daejeon, Rep. of Korea.

http://dx.doi.org/10.4218/etrij.15.0115.0124

parameter, the excitation parameter, and the state duration of a speech unit are represented by a context-dependent HMM model. In the synthesis part, speech is synthesized by Mel-log spectrum approximation (MLSA) filtering with speech parameters generated from trained context-dependent HMMs [5].

A drawback of Tokuda and others' HTS when compared with a corpus-based TTS is the quality of the synthetic speech. The drawback occurs as a result of their conventional excitation (CE) model, which is generated from the random noise and periodic pulse train of unvoiced and voiced speech, respectively. The CE model highlights the fact that synthesized speech, such as vocoded speech, contains buzzy sounds. Thus, to remove such buzzy sounds, a mixed excitation (ME) model using a multi-band mixing model was developed by Yoshimura and others applied a mixed-excitation linear predictive (MELP) algorithm of a vocoder to HTS [7]. In this model, excitation is divided into five fixed frequency bands and generated by a periodic impulse train or random noise in accordance with the periodicity. Although this ME model reduces buzzy sounds, the resolutions of the five fixed bands are not optimal for wide-band speech. The reason for this is that the MELP vocoder was developed for narrow-band speech. In addition, if the number of bands is increased in the ME model, then it needs much memory for trained data. To overcome this problem, Kim and others incorporated two-band excitation (TBE) based on a harmonic plus noise speech model into the HTS of Tokuda and others [8]–[10]. This TBE model decomposes speech into periodic and non-periodic parts via a maximum voiced frequency (MVF). Periodic and aperiodic excitations are generated by a periodic pulse train and a random noise, respectively. Therefore, estimating an MVF is of extreme importance to the TBE model because the speech quality is highly dependent upon the MVF.

To estimate an MVF, Kim and others proposed an MVF estimation scheme based on a filtering-based TBE (FTBE) scheme. This scheme utilizes the normalized auto-correlation of the high-pass filtered speech around pitch lag [8]–[10]. In addition, to improve the accuracy of an MVF, Han and others proposed an analysis by synthesis (ABS)–based MVF optimization scheme using the initial MVF, in [11]. However, the FTBE scheme occasionally misestimates the initial MVF, because the scheme estimates the MVF from an input signal including a spectral envelope. In particular, it is difficult to find the boundary of the periodic and aperiodic parts due to the spectral envelope that lies within the interval in which the periodic and aperiodic parts are mixed together in the frequency domain.

Thus, this paper proposes an enhanced MVF estimation scheme based on peak picking to improve the accuracy of an initial MVF. The proposed scheme uses a linear predictive (LP) residual signal as an excitation signal instead of an input signal. To estimate an MVF, both a harmonic peak and a peak lobe are *picked* from the spectrum of the LP residual signal and normalized distances of local peaks and peak lobes are utilized.

This paper is organized as follows. Section II describes the TBE model and conventional MVF estimation scheme. Section III explains the overall procedure of the proposed scheme in detail, and Section IV shows the experimental results. Finally, conclusions and future works are given in Section V.

II. TBE Model

The TBE model is essentially similar to the ME model. In the ME model, the excitation signal is divided into five fixed frequency bands [7]. The periodicity of each band is decided by the band-pass voicing strength. The excitation signal in each band is generated by the periodic impulse train or the random noise according to the periodicity. The ME model successfully reduces the buzzy sounds and improves the quality of synthetic speech. However, these bands are not optimal for wide-band speech, because an MELP vocoder is designed for narrow-band speech only. If the number of bands in the ME model is increased to obtain a better resolution, then more memory is required. Therefore, a TBE model was proposed to overcome such a problem [8]. In the TBE model, the MVF, a time-varying parameter, marks the boundary between the periodic and aperiodic bands in the frequency domain, as shown in Fig. 1. Furthermore, this model is very useful in terms of human auditory characteristics, as in the synthetic speech created using the TBE model is of better quality than that created using the ME model [8]. The TBE model generates an excitation signal from a fundamental frequency and an MVF. Therefore, an accurate MVF leads to a high quality of synthetic speech. For accurate MVF estimation, it is important to precisely estimate the fundamental frequency. However, for the purposes of our paper, we make the assumption that the fundamental frequency is given to have already been estimated accurately.

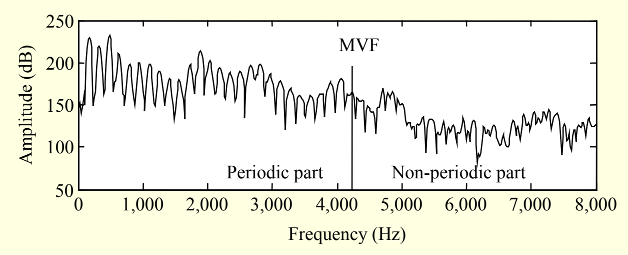

Fig. 1. Spectrum of voiced speech and MVF.

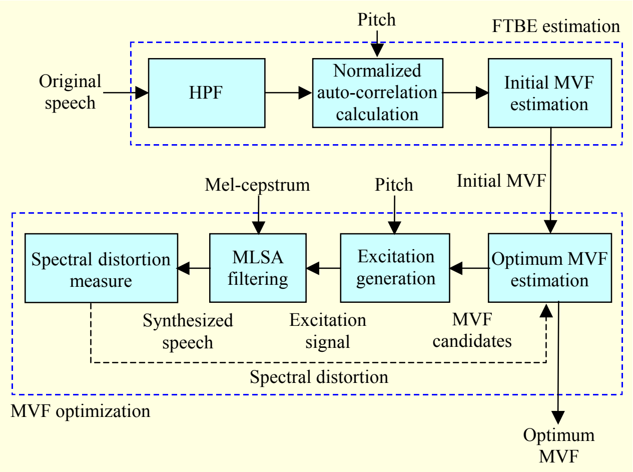

Fig. 2. Procedure of optimum MVF estimation scheme.

Recently, an ABS-based optimum MVF estimation scheme was proposed [11]. Figure 2 describes the procedure of this scheme. This scheme estimates an MVF in two steps. At the first step, an initial MVF is estimated by the FTBE scheme [8]. The FTBE utilizes the normalized auto-correlation of the high-pass filtered speech to estimate an MVF. The filtered speech, denoted by s_{HPF}^{f}, is defined by the convolution between an input speech and a high-pass filter (HPF), where f is the cut-off frequency of the HPF. The normalized auto-correlation, R_f, can be represented as

$$R_f = \frac{\sum_{n=0}^{N-1} s_{\mathrm{HPF}}^{f}(n) s_{\mathrm{HPF}}^{f}(n+\tau)}{\sqrt{\sum_{n=0}^{N-1}\left[s_{\mathrm{HPF}}^{f}(n)\right]^2 \sum_{n=0}^{N-1}\left[s_{\mathrm{HPF}}^{f}(n+\tau)\right]^2}}, \quad (1)$$

where τ, n, and N are an estimated pitch, a time index, and the size of an analysis window, respectively. The normalized auto-correlation of the filtered speech is bounded between –1 and 1; however, practically, it is usually bounded between 0 and 1. If the cut-off frequency is smaller than the MVF, then s_{HPF}^{f} is periodic, and R_f would be close to 1. On the other hand, if the cut-off frequency is larger than the MVF, then s_{HPF}^{f} is aperiodic, and R_f would be close to 0. Thus, if R_f is smaller than 0.5, then its cut-off frequency is adopted as the MVF. We designed 16 HPFs using the Butterworth method, and their cut-off frequencies increase in 500 Hz increments from 500 Hz to 7,500 Hz. In the second step, an optimum MVF is found by an MVF optimization scheme. The scheme utilizes an ABS scheme to minimize spectral distortion. Firstly, an excitation signal is generated with initial MVF and fundamental frequency. Secondly, speech is synthesized by an MLSA filter with an extracted Mel-cepstrum and generated excitation. Then, the synthetic speech quality is measured by the spectral distortion and the symmetric Kullback–Leibler distance (SKLD) is used as the spectral distortion measurement. The SKLD is calculated as

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

where S_i is the normalized power spectrum of the ith frame, and k and K are the index of frequency and size of a discrete Fourier transform (DFT), respectively. This procedure is performed repeatedly among the MVF candidates, which are determined from the initial MVF. Then the MVF candidate having minimum spectral distortion is finally determined as the optimum MVF. This scheme leads to high synthetic speech quality. However, if the accuracy of the initial MVF is decreased, then the synthesized speech would be of a poorer quality.

III. Proposed MVF Estimation Scheme

The main problem with the FTBE scheme is that it misestimates MVFs, because the scheme estimates an MVF from an input signal including a spectral envelope. In particular, this scheme misestimates an MVF at an interval in which periodic and aperiodic components in the frequency domain are mixed together. In such an interval, it is difficult to find the boundary that divides the periodic and aperiodic parts, due to the presence of a spectral envelope. This is the root cause of poor-quality synthesized speech. If an MVF is misestimated and is too high, then any subsequent synthesized speech will contain buzzy sounds. In contrast, if an MVF is misestimated and is too low, then any resulting synthesized speech sounds will be too harsh. To obtain an accurate MVF estimation, an LP residual is utilized as the excitation of speech. The LP residual signal removes the spectral envelope and emphasizes the periodic component. However, the FTBE scheme is difficult to use with residual signals, because the normalized auto-correlation of a high-pass filtered residual signal, such as that in (1), is too small at a cut-off frequency of 500 Hz. Therefore, a residual signal is more unsuitable than an original signal for use as an MVF decision threshold in the FTBE scheme. However, the periodic part of the spectrum of a residual signal is better represented than that of the original spectrum. Harmonic peaks within the spectrum of a residual signal can be detected easily and be a robust feature in the search for the boundary between periodic and aperiodic parts. Therefore, to more accurately estimate an initial MVF, we propose an MVF estimation scheme based on a harmonic peak picking TBE (PTBE) scheme using a residual signal. The proposed scheme, described in Fig. 3, also includes the same MVF optimization

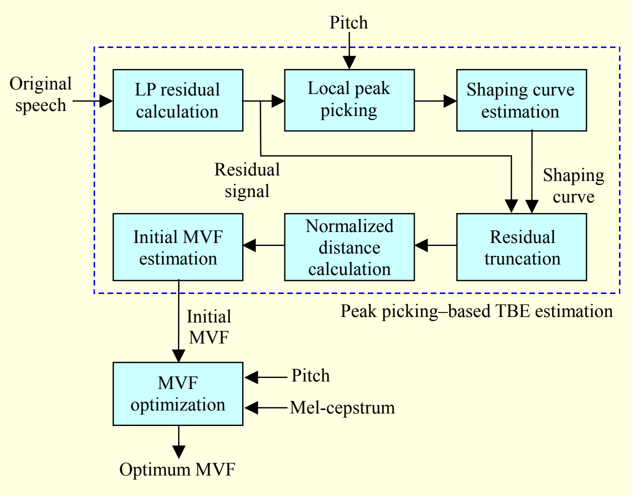

Fig. 3. Procedure of proposed MVF estimation scheme.

technique explained in Section II. This scheme performs an MVF estimation in a voiced frame decided by an unvoiced/voiced decision. For such a decision, an estimated fundamental frequency is used at each frame. If such a fundamental frequency is 0, then the input frame is an unvoiced frame and the MVF is just 0. Otherwise, the input frame is deemed to be a voiced frame.

1. LP Residual Calculation

The first step of the proposed scheme is to extract an LP residual signal using LP filtering from which to select local peaks. To this end, an LP coefficient (LPC), a, is calculated from the Levinson–Durbin algorithm. Upon calculating the LPC, an LP residual signal, r, is obtained by LP filtering and can be written as

$$r(n) = x(n) - \sum_{i=1}^{p} a(i)x(n-i), \qquad (3)$$

where i and p are the index of LPC and LPC order, respectively. Because LP filtering emphasizes only the periodicity, which is highly related with harmonic peaks, the harmonic peaks themselves can easily be found. An example of a residual signal is shown in Fig. 4. From this, we can see that the spectrum of a residual signal is flatter than that of an input signal.

2. Local Peak Picking

In this process, a harmonic *local peak* and its location are calculated using an LP residual signal. Because harmonics are closely related with periodicity, harmonic peaks and their locations are significant in attempts to estimate an MVF. First, an LP residual signal is transformed into a frequency-domain signal by a DFT, and the power spectrum is commonly calculated as

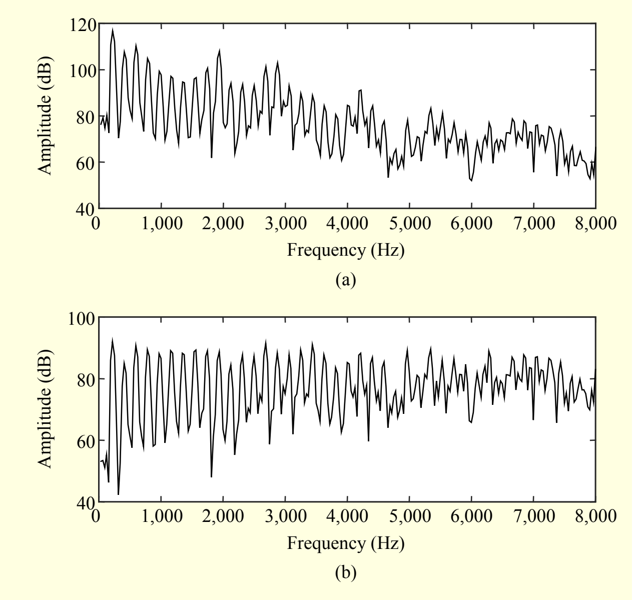

Fig. 4. Spectrum: (a) original signal and (b) LP residual signal.

$$P_R(k) = 20\log_{10}\left(\left|R(k)\right|\right), \qquad (4)$$

where R is a residual signal in the frequency domain and $|\bullet|$ is an absolute notation. To find local peaks, the values of all peaks in P_R are found using

$$\text{peak}(k) = \begin{cases} P_R(k) & \left(\Delta P_R(k) > 0\right) \text{and} \left(\Delta P_R(k+1) < 0\right), \\ 0 & \text{otherwise}, \end{cases} \qquad (5)$$

where

$$\Delta P_R(k) = P_R(k) - P_R(k-1). \qquad (6)$$

Second, the locations of local peaks are determined through

$$l_{\text{P}}(\kappa) = \arg\max_{i} \left[\text{peak}\left(\kappa \frac{F_0}{2} + i\right) \right], \qquad (7)$$

where l_{P} and F_0 are the location of a local peak and the estimated fundamental frequency, respectively. In addition, the condition on (7) is that $-F_0/2 < i < F_0/2$ and the search criteria is $\kappa = 1, 2, \ldots, \lfloor F_{\text{S}} / F_0 \rfloor$, where F_{S} a sampling rate and $\lfloor \bullet \rfloor$ means to round down. Finally, values of local peaks, peak$_{\text{l}}$, are obtained from "peak" and "l_{P}" as follows:

$$\text{peak}_{\text{l}}(k) = \begin{cases} \text{peak}\left(\kappa \dfrac{F_0}{2} + l_{\text{P}}(k)\right) & \kappa = 1, 2, \ldots, \left\lfloor \dfrac{F_{\text{S}}}{F_0} \right\rfloor, \\ 0 & \text{otherwise}. \end{cases} \qquad (8)$$

The distances between adjacent local peaks denote the periodicity of a spectrum. Thus, the location at where such distances begin to change rapidly marks the beginning of the

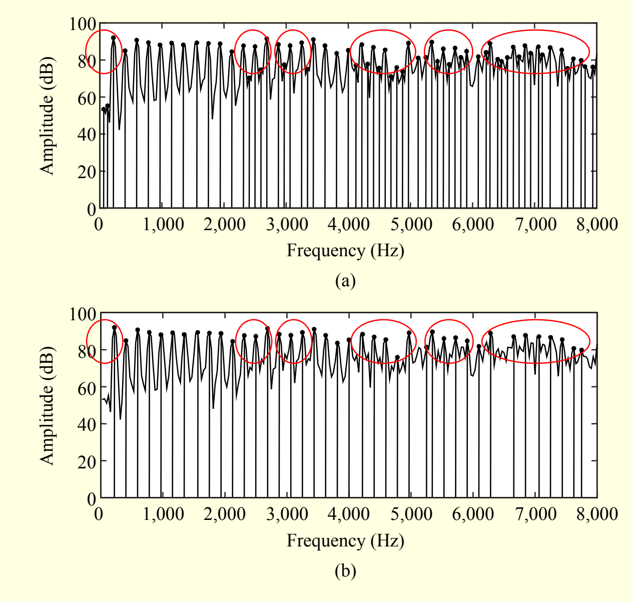

Fig. 5. Peak location: (a) all peaks and (b) local peaks (circle: picking local peak positon from all peaks).

aperiodic part of the spectrum. Figures 5(a) and 5(b) show the peaks and local peaks of a spectrum of a residual signal, respectively. Adjacent local peaks within the periodic part of a spectrum are separated by a uniform distance; this is not true of the aperiodic part of the spectrum, whereby it can be seen that non-uniform distances separate adjacent local peaks (see Fig. 5(b)).

3. Shaping Curve Estimation and Truncation

Even though local peaks can be correctly found from the spectrum of a residual signal, it is not easy to find the location of an MVF using only these local peaks. Not only the local peaks themselves but also the distances between the nearby lobes of local peaks are an important feature for calculating the periodicity of a spectrum. To separate the lobes from the spectrum of the residual signal, we truncate the residual spectrum using a shaping curve of the local peaks. The shaping curve, Sh, is obtained through a linear interpolation of the values of local peaks. After obtaining the shaping curve, a truncation curve, Tr, is calculated by

$$\mathrm{Tr}(k) = \mathrm{Sh}(k) - 3\ \mathrm{dB}, \tag{9}$$

where 3 dB is used as a threshold for truncation because it has been generally defined as the bandwidth of the cut-off frequency. Next, the power spectrum of the truncated residual signal, P_{T}, is obtained by

$$P_{\mathrm{T}}(k) = P_R(k) - \mathrm{Tr}(k). \tag{10}$$

An example of this process is shown in Fig. 6. Figures 6(a) and 6(b) show an estimation of the shaping curve and the point at which to truncate the residual signal by (10), respectively.

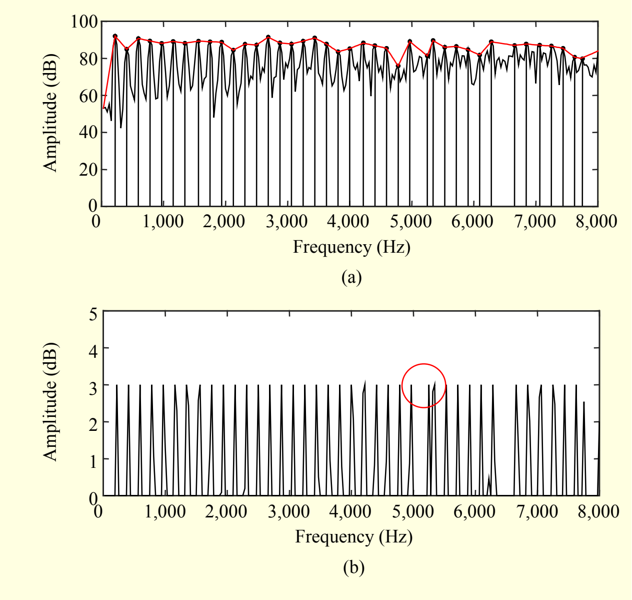

Fig. 6. (a) Estimated shaping curve and (b) truncated residual spectrum (red line: shaping curve, circle: evidence of breaking periodicity).

4. Normalized Distance Calculation and MVF Estimation

The final process determines an initial MVF from local peaks and truncated residual spectrum. The periodicity of the truncated residual spectrum is related to the distances between local peaks. Therefore, these distances are calculated as features to decide an initial MVF. The first distance value is used for normalization and is kept as a reference distance. After the distances between adjacent local peaks have been calculated, they are normalized using the stored reference distance. In addition, the distances between adjacent peak-lobes are also normalized using a reference distance for the lobes in the same way because the distances between the peak lobes is an important feature of the periodicity of the spectrum. The normalized distances help to estimate an accurate MVF. All frequency values that fall below that of the MVF mark the

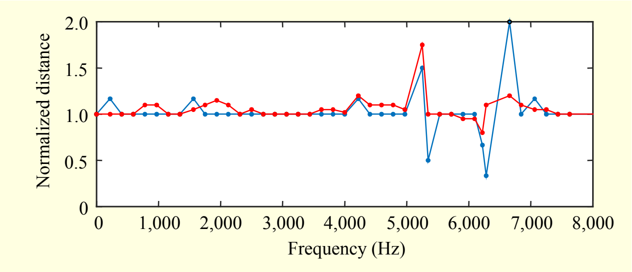

Fig. 7. Normalized distances (blue: local peaks, red: peak lobes).

periodic part of the spectrum. Here, the normalized distances of the local peaks tend to revolve around a value of 1 over the changes in frequency. On the contrary, all frequency values that are higher than the MVF mark the aperiodic part of the spectrum. Here, the normalized distances rapidly increase or decrease over the changes in frequency. The obtained normalized distances are shown in Fig. 7.

In our system, the average of the normalized distances of local peaks and peak lobes is utilized to determine an MVF. If this average becomes less than 0.5 or more than 1.5 at some frequency, then the frequency is used as an initial MVF. In other words, the lowest frequency that satisfies the above condition is estimated as the initial MVF.

IV. Experiment and Results

1. Experimental Setup

For our experiments, we utilized training data consisting of 4,000 sentences uttered by a group of five females. Three thousand of these sentences were used for a training procedure, and the remaining 1,000 sentences were used for a synthesis procedure. The average duration of each sentence is about 3 s, and all sentences were sampled at 16 kHz with a 16-bit quantization level. The version of the HTS system used for the experiments was 2.1 and the same contexture information provided by the HTS homepage was used. Thirteen Mel-cepstrum coefficients including a 0th-order coefficient were extracted from a 25 ms hamming-windowed speech, with a 5 ms frame-shift as a spectral parameter. The excitation signal is modeled by the optimized MVF estimation–based PTBE (O-PTBE) proposed in Section III. In our experiments, the initial MVF is quantized by a 500 Hz step size due to the same condition of [10] and MVF optimization uses the initial MVF and four other MVFs nearby the initial MVF [11]. The fundamental frequency is extracted with the speech transformation and representation based on an adaptive interpolation of a weighted spectrogram (STRAIGHT) [12]. The MVF is determined as described in Section III. For the proposed scheme, we use an LPC of order 16 and a 512-point fast Fourier transform executed as part of a time-to-frequency mapping technique.

To evaluate the performance of the O-PTBE, excitation signals were generated by CE, ME, and optimized MVF estimation–based FTBE (O-FTBE). The ME scheme is modeled by STRAIGHT. All HTS systems produce synthesized speech via MLSA filtering of the input excitation signals. Firstly, we used SKLD and log-spectral distance (LSD) as the objective distortion measurements between the original and the synthesized speech in both the training and synthesis procedures. The SKLD was calculated by

$$D_{\mathrm{SKLD}} = \sum_k \left(\left(P(k) - Q(k) \right) \log \frac{P(k)}{Q(k)} \right), \tag{11}$$

where P and Q denote the power spectra of the original and synthesized speech, respectively. In addition, we obtain the LSD by using

$$D_{\mathrm{LSD}} = \sqrt{\sum_k \left[10 \log \frac{P(k)}{Q(k)} \right]^2}. \tag{12}$$

Distortions were measured for all of the trained sentences. In the synthesis procedure, distortions were measured after aligning the durations of the synthesized speeches and original speeches through use of a dynamic time warping technique. As a subjective listening test, the mean opinion score (MOS) and preference tests were performed [13]. Ten experienced listeners evaluated the synthesized speech quality of five speeches taken from within the set of training sentences and five speeches taken from within the set of non-training sentences. For the listening tests, listeners heard synthesized speech through the loudspeaker of a smartphone (SAMSUNG Galaxy Note 2) and through the ear pieces of a Stax Lambda Pro headphone set attached to a desktop PC. In the MOS test, listeners scored the quality of synthesized speech by the CE, ME, O-FTBE, and O-PTBE schemes. In the preference test, listeners evaluated a pair of speeches synthesized by the O-FTBE and O-PTBE schemes.

2. Experimental Results

Table 1 shows the results of the objective distortions. In the training procedure, the ME scheme looks to have performed slightly better than the others. The ME scheme achieved an SKLD gain of 201.37 and an LSD gain of 68.10 as a minimum distortion, while O-PTBE achieved SKLD and LSD gains of 201.42 and 68.12, respectively. However, the results were different for the synthesis procedure. The results of the O-FTBE and O-PTBE schemes show smaller distortions than those obtained for the ME scheme. The minimum distortion was obtained through the O-PTBE scheme with an SKLD gain

Table 1. Objective tests results.

	Training procedure		Synthesis procedure	
	SKLD	LSD	SKLD	LSD
CE	207.86	69.83	328.37	83.24
ME	201.37	68.10	322.29	82.87
O-FTBE	203.35	68.32	320.32	82.53
O-PTBE	201.42	68.12	317.07	82.18

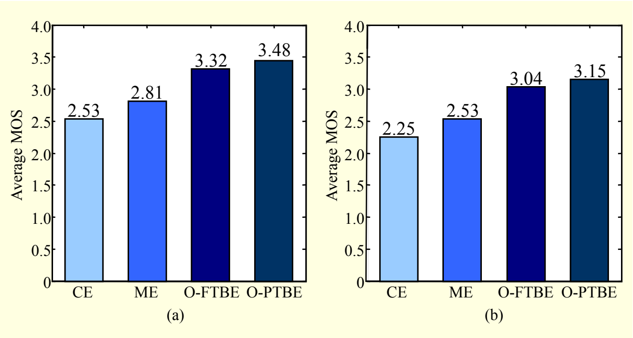

Fig. 8. MOS test results: (a) PC headphones and (b) smartphone speakers.

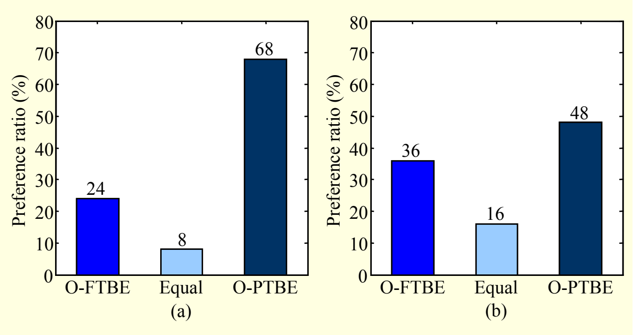

Fig. 9. Preference test results: (a) PC headphones and (b) smartphone speakers.

of 317.07 and an LSD gain of 82.18.

Figures 8 and 9 show the results of the MOS and preference tests, respectively. In the MOS test (Fig. 8), the O-PTBE scheme scores the best score among all the schemes at both the PC and the smartphone. The average MOS of the O-PTBE scheme was 3.315 and its gains were 0.645 and 0.135 compared with the ME and O-FTBE schemes, respectively. In the preference test (Fig. 9), the listeners preferred the speech synthesized by the O-PTBE scheme in the smartphone as well as the PC. In all subjective tests, the O-PTBE scheme showed the best performance among the other schemes. As shown in Table 1, Fig. 8, and Fig. 9, the results of the objective and subjective tests can be summarized as follows: the O-PTBE scheme showed better performance than the CE and O-FTBE schemes for all evaluations. In addition, the O-PTBE scheme outperforms the ME scheme for all evaluations except the objective tests in the training procedure. In summary, the O-PTBE scheme guarantees a more accurate MVF and improves the quality of synthesized speech.

Figure 10 shows an example of a spectrogram and MVF contour comparisons between the O-FTBE and the O-PTBE schemes in the training procedure. The background is a spectrogram and the solid line is the MVF contour. The MVF contours are estimated by the O-FTBE and O-PTBE schemes from identical speech samples. Figure 11 shows an example of the spectrogram and the MVF contour comparisons between the O-FTBE and the O-PTBE schemes in the synthesis procedure. The background and the solid line are the spectrogram and the MVF contour, respectively. The two different schemes synthesize the same text but use different methods, as well as estimating the MVF contour.

Table 2 lists the binary file sizes of the synthesis system. Even though the total memory for the trained HMM data, decision tree data, and a synthesizer engine is only about 2.1 MB, the synthesized speech shows a fairly good quality.

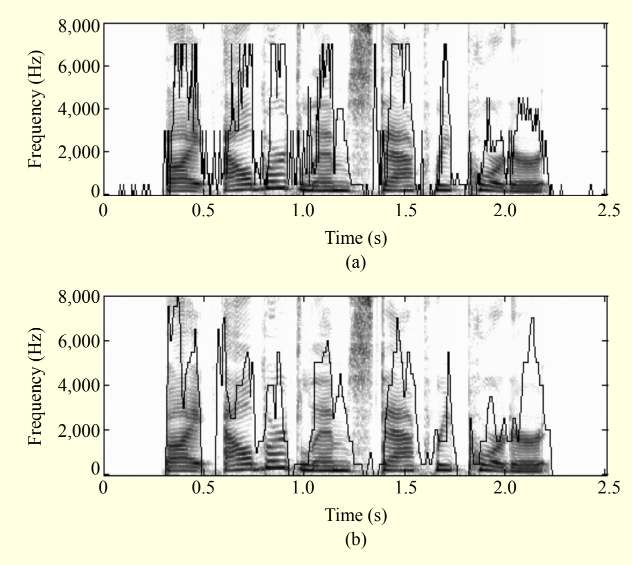

Fig. 10. Spectrograms and MVF contours in training procedure: (a) O-FTBE and (b) O-PTBE.

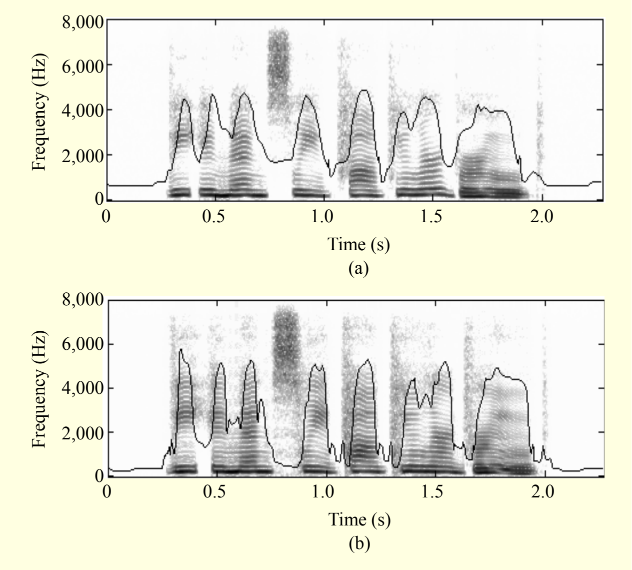

Fig. 11. Spectrograms and MVF contours in synthesis procedure: (a) O-FTBE and (b) O-PTBE.

Table 2. Binary file size of synthesis system.

Module		Size (MB)
Trained HMM data	Spectrum	1.292
	Excitation	0.160
	Duration	0.014
Decision tree data	Spectrum	0.254
	Excitation	0.304
	Duration	0.037
Synthesizer engine		0.073
Total		2.134

The memory sizes of the O-FTBE and O-PTBE schemes are the same, whereas the quality of the speech synthesized by the O-PTBE scheme is better than that of the O-FTBE scheme.

V. Conclusion

The TBE model was a useful excitation model and the MVF is an important feature for the TBE model. However, the FTBE scheme misestimates the MVF because of the spectral envelope of speech. Thus, this paper proposed a harmonic peak picking–based MVF estimation scheme using the spectrum of an LP residual signal and proved the performance of the proposed scheme. From our results, it can be seen that the proposed scheme obtains the best results in both objective and subjective tests in comparison with other schemes. The main reason for this is that the proposed scheme estimates a reliable MVF in the training procedure. Work still remains to be done in improving the excitation generation module in the synthesis procedure.

References

[1] A. Hunt and A.W. Black, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Atlanta, GA, USA, May 7–10, 1996, pp. 373–376.

[2] K. Tokuda, T. Kobayashi, and S. Imai, "Speech Parameter Generation form HMM Using Dynamic Features," *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, Michigan, USA, May 8–12, 1995, pp. 660–663.

[3] K. Tokuda et al., "An Algorithm for Speech Parameter Generation from Continuous Mixture HMMs with Dynamic Features," *Proc. Eurospeech*, Madrid, Spain, Sept. 18–21, 1995, pp. 757–760.

[4] K. Tokuda, H. Zen, and A.W. Black, "An HMM-Based Speech Synthesis System Applied to English," *Proc. IEEE Workshop Speech Synthesis*, Santa Monica, CA, USA, Sept. 11–13, 2002, pp. 227–230.

[5] T. Fukada et al., "An Adaptive Algorithm for Mel-cepstral Analysis of Speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, San Francisco, CA, USA, Mar. 23–26, 1992, pp. 137–140.

[6] K. Tokuda et al., "Speech Parameter Generation Algorithm for HMM-Based Speech Synthesis," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Istanbul, Turkey, June 5–9, 2000, pp. 1315–1318.

[7] T. Yoshimura et al., "Mixed Excitation for HMM-Based Speech Synthesis," *Proc. Eurospeech*, Aalborg, Denmark, Sept. 3–7, 2001, pp. 2263–2266.

[8] S.-J. Kim, J.-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.

[9] S.-J. Kim, J.-J. Kim, and M. Hahn, "Implementation and Evaluation of an HMM-Based Korean Speech Synthesis System," *IEICE Trans. Inf. Syst.*, vol. E89-D, no. 3, Mar. 2006, pp. 1116–1119.

[10] S.-J. Kim and M. Hahn, "Two-Band Excitation for HMM-Based Speech Synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no 1, Jan. 2007, pp. 378–381.

[11] S. Han, S. Jeong, and M. Hahn, "Optimum MVF Estimation-Based Two-Band Excitation for HMM-Based Speech Synthesis," *ETRI J.*, vol. 31, no. 4, Aug. 2009, pp. 457–459.

[12] H. Zen et al., "Details of Nitech HMM-Based Speech Synthesis System for the Blizzard Challenge 2005," *IEICE Trans. Inf. Syst.*, vol. E90-D, no 1. Jan. 2007, pp. 325–333.

[13] X. Huang, A. Acreo, and H.-W. Hon, "*Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*," Prentice Hall: New Jersey, 2001, pp. 840–842.

Jihoon Park received his BS and MS degrees in electronics engineering from the Information and Communications University, Daejeon, Rep. of Korea, in 2005 and 2007, respectively. He received his PhD degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea, in 2013. Since 2103, he has been a research professor at the Center for Integrated Smart Sensors, Daejeon, Rep. of Korea. His research interests include microphone array–based speech enhancement, HMM-based speech synthesis, VoIP, multi-channel audio coding, multi-object audio coding, and ultra-high definition audio coding.

Minsoo Hahn received his BS and MS degrees in electrical engineering from Seoul National University, Rep. of Korea, in 1979 and 1981, respectively. He received his PhD degree in electrical engineering from the University of Florida, Gainesville, USA, in 1989. From 1982 to 1985, he was with the Korea Research Institute of Standards and Science, Daejeon, Rep. of Korea. From 1990 to 1997, he was with ETRI, Rep. of Korea. Since 1998, he has been a professor with the Department of Electrical Engineering and a director with the Digital Media Laboratory, both of the Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea. His research interests include speech; audio and biological signal processing; speech synthesis; noise reduction; and VoIP.

http://dx.doi.org/10.4218/etrij.15.0115.0124