HMM-Based Korean Speech Synthesis System for Hand-held Devices

Sang-Jin Kim, Jong-Jin Kim, and Minsoo Hahn

Abstract — *Speech interface may be the first choice as a user interface for robots or hand-held devices such as personal digital assistants (PDAs) and portable multimedia players (PMPs). However, those devices have the limitation of the memory space and the computation power. The hidden Markov model (HMM)-based speech synthesis is presently considered to be suitable for the embedded systems. In this paper, our HMM-based Korean speech synthesis system is described. Statistical HMM models for Korean speech units are trained with the hand-labeled speech database including the contextual information about phoneme, word phrase, and multilevel break strength. Mel-cepstrum and line spectrum pair (LSP) are compared for the spectrum modeling, and two-band excitation based on the harmonic plus noise speech model is utilized for the mixed excitation source. The developed small-size Korean synthesis system produced considerably high quality speech with a fairly good prosody.*

Index Terms — Speech synthesis, excitation model, context clustering, Korean.

I. INTRODUCTION

Speech is an easy and intuitive communication method. Speech interface with automatic speech recognition (ASR) and text-to-speech (TTS) synthesis, is one of the best methods to manipulate the mobile devices such as PDAs, car navigation systems or robots instead of keyboard inputs and text or graphical outputs. TTS can also enable users to have their e-mail or text messages read on their mobile phones. This is useful for not only normal people but the blinds. Those devices, however, have the limitation of the memory space and the computation power.

Corpus-based unit concatenating speech synthesis [1] is most popular approach mainly because of its high output speech quality, but it needs a huge speech unit database. Thus, it is inappropriate for mobile devices. Tokuda et al. has proposed the HMM-based approach to the speech synthesis [2]. This method uses the statistical HMMs to model the spectra and the prosodic parameters of the speech units. The synthesis engine needs less memory and low computation complexity, so presently considered to be suitable for the embedded systems [3],[4].

Sang-Jin Kim is with Speech & Audio Info. Lab., Information and Communications University, Daejeon, Korea. (e-mail: sangjin@icu.ac.kr).

Jong-Jin Kim is with with Electronics and Telecommunications Research Institute (ETRI), Deajeon, Korea. (kimjj@etri.re.kr)

Minsoo Hahn is with Speech & Audio Info. Lab., Information and Communications University, Daejeon, Korea. (e-mail: mshahn@icu.ac.kr).

Recently, the HMM-based speech synthesis technique has been reported for other languages [4]-[7], although it has been originally developed to support Japanese. In this paper, our HMM-based Korean speech synthesis system is described.

The rest of this paper is organized as follows. Section II summarizes the HMM-based speech synthesis system briefly. In Section III, the contextual information for Korean and its clustering are presented. And then Section IV and V describe the spectral parameters and the excitation model, respectively. Subjective listening tests of the generated speech are discussed in Section VI while the conclusion is given in the final section.

II. HMM-BASED SPEECH SYNTHESIS [3]

The HMM-based speech synthesis system consists of the training and the synthesis part. In the training part, the statistical HMM model represents the spectrum, the excitation, and the state duration of the context-dependent speech units. Each HMM model has left-to-right state transition with no skip. The spectrum and the F0, i.e., the pitch, are related to the vocal tract shape and the excitation source, respectively. The mel-cepstrum has been utilized for the spectrum analysis.

In the synthesis part, the system input is a contextual label sequence of the text. The contextual label format has to be same with that used in the HMM training. Firstly, the corresponding contexture dependent HMMs are concatenated and then state durations for the HMM sequence are

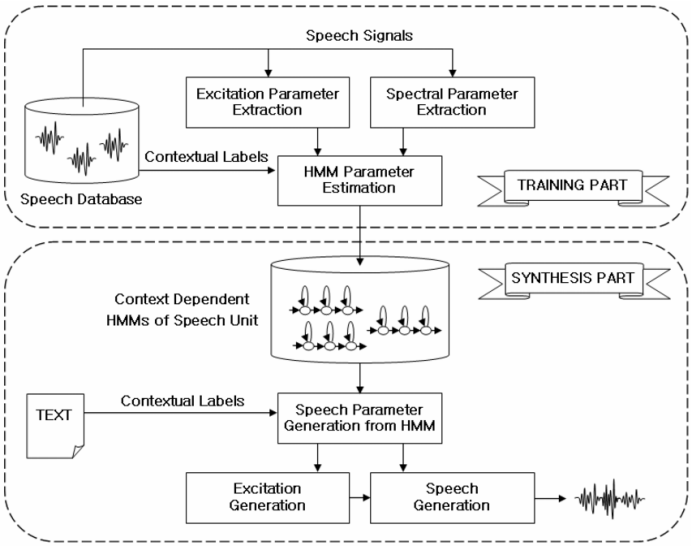


Fig. 1. HMM-based speech synthesis system

Contributed Paper
Manuscript received October 2, 2006

determined. After the speech parameters such as mel-cepstral coefficients and F0 values in log scale are generated from the HMM sequence, the output speech is synthesized by MLSA (Mel Log Spectrum Approximation) filtering [8].

III. Contextual Information and its Clustering

A. Characteristic of the Korean Language

Korean and Japanese are completely different language except for the grammatical structure. Korean is also completely different from Chinese, even though Korean has borrowed many Chinese words and has used Chinese characters.

The Korean alphabet, *Hangul*, consists of forty letters. Twenty-one of them represent vowels, including thirteen diphthongs, and the remaining nineteen represent consonants. Table I shows the Korean vowels while Table II, the Korean consonants.

Phonemes are combined to form a syllable, and several syllables are combined to form a word phrase (*Eojeol* in Korean) which is different from a phrase in English. The syllable structures of Korean are V, CV, VC, and CVC, where C and V stand for a consonant and a vowel, respectively.

B. Contextual Information

Contextual information is language dependent. Besides, a large number of contextual factors doesn't guarantee the better quality output speech. The contextual factors taken into account for Korean in this paper are as follows:

- {preceding, current, succeeding} phoneme
- position of current phoneme in current syllable
- number of syllables in current word phrase (eojeol)
- position of current syllable in current word phrase
- {forward, backward} break strength of current syllable,
- {forward, backward} break strength of current word phrase

More factors were considered to model the Korean prosody better based on our former experiments of the corpus-based Korean TTS system [9]-[11]. However, we found that finer information wouldn't improve the synthesized speech quality with the limited database a lot. Above factors are the final results of our informal listening test. But their combination is still large. Since the training data should be limited, decision-tree based context clustering is utilized to model the parameters with sufficient accuracy.

C. Context & Prosody Clustering

It is obvious that as contextual factors increase, their combinations also increase exponentially. Since there are many contextual factors, if we try to consider all of them, it is impossible to prepare speech database which includes all combinations of contextual factors. It is inevitable that the size of training data should be limited and model parameters wouldn't be sufficiently accurate. Decision-tree based context clustering is utilized to overcome this problem. Not only the contextual information in the previous subsection is clustered but the phonemes are also clustered according to Table III and Table IV. Table III and Table IV shows the classification of Korean vowel and consonant allophones, respectively.

TABLE I
Korean monothong vowels classified according to tongue position and height.

Tongue Height	Tongue Position					
	Front		Center		Back	
	Unrounded	Rounded	Unrounded	Rounded	Unrounded	Rounded
High	ㅣ[i]	-	-	-	ㅡ[ɯ]	ㅜ[u]
High-mid	ㅖ[e]	-	ㅓ[ʌ]	-	-	ㅗ[o]
Mid-low	ㅐ[ɛ]	-	-	-	-	-
Low	-	-	ㅏ[ɑ]	-	-	-

TABLE II
Korean consonants classified according to place and manner of articulation.

Manner of articulation	Place of articulation				
	Bilabial	Dental/Alveolar	Palatal	Velar	Glottal
Plosives	ㅂ[b], ㅃ[p], ㅍ[p^h]	ㄷ[t], ㄸ[t], ㅌ[t^h]	-	ㄱ[g], ㄲ[k], ㅋ[k^h]	-
Affricates	-	-	ㅈ[ʤ], ㅉ[tɕ], ㅊ[$tɕ^h$]	-	-
Fricatives	-	ㅅ[s], ㅆ[s^h]	-	-	ㅎ[h]
Nasals	ㅁ[m]	ㄴ[n]	-	ㅇ[ŋ]	-
Liquids	-	ㄹ[l]	-	-	-

TABLE III
Classification table of Korean vowel allophones.

Classification factors			Korean phonemes
Monothong	Flat open		ㅏ[ɑ], ㅓ[ʌ]
	Back rounded		ㅗ[o], ㅜ[u]
	Front flat closed		ㅣ[i]
	Back closed		ㅡ[ɯ]
	Front flat semi-closed		ㅖ[e], ㅐ[ɛ]
Diphthong	Front part	Front flat closed	ㅑ[jɑ], ㅕ[jʌ], ㅛ[jo], ㅠ[ju], ㅒ[jɛ], ㅖ[je]
		Rounded	ㅘ[wɑ], ㅝ[wʌ], ㅙ[wɛ], ㅞ[we], ㅚ[ɸ], ㅟ[y]
		Back flat closed	ㅢ[ɯɪ]
	Back part	Flat open	ㅑ[jɑ], ㅕ[jʌ], ㅘ[wɑ], ㅝ[wʌ]
		Back rounded	ㅛ[jo], ㅠ[ju]
		Front flat closed	ㅟ[y], ㅢ[mi]
		Front flat semi-closed	ㅒ[jɛ], ㅖ[je], ㅙ[wɛ], ㅞ[we], ㅚ[ɸ]

TABLE IV
Classification table of Korean consonant allophones

Classification factors			Korean phonemes
Syllable initial	Voiced	Nasal	ㄴ[n], ㅁ[m]
		Liquid	ㄹ[r]
	Unvoiced	Tense unaspirated obstruent	ㄲ[k], ㄸ[t], ㅃ[p], ㅆ[s^h], ㅉ[tɕ]
		Tense aspirated obstruent	ㅍ[p^h], ㅌ[t^h], ㅋ[k^h], ㅊ[$tɕ^h$]
		Lax plosive	ㄱ[g], ㄷ[d], ㅂ[b]
		Fricative	ㅅ[s], ㅈ[ʤ]
		Vowel-like fricative	ㅎ[h]
Syllable final	Voiced	Nasal	ㄴ[n], ㅁ[m], ㅇ[ŋ]
		Liquid	ㄹ[l]
	Unvoiced		ㄱ[g], ㄷ[d], ㅂ[b]
Silence			#4, #5, #6

IV. Feature Parameter for Spectrum Modeling

A. Mel-cepstrum and Line Spectrum Pair (LSP)

The HMM-based TTS system is basically a source-filter model, so it is necessary to extract feature parameters that describe the vocal tract shape. Because the mel-cepstrum satisfies the stability and the interpolation performance constraints of the synthesis filter [12], it is utilized for the spectrum analysis since the HMM-based speech synthesis has been introduced [2],[3].

Even though the LSP is a variation of the linear prediction analysis based on the all-pole model, it is also known for good characteristics of the stability and the interpolation performance, and widely used for many standard vocoders. It is of worth to test its feasibility for the HMM-based speech synthesis.

Because the mel-cepstrum and the LSP are well-known parameters, their mathematical details would not be covered in this paper.

B. Speech Analysis and Synthesis

Even though the mel-cepstral analysis has been known for better characteristics mathematically and theoretically than the LSP analysis [12], we executed our own preference listening test to compare them and to confirm the validity. The speech sentences uttered by six announcers (three males and three females) are sampled at 16 kHz with 16 bit resolution. The 25 ms Blackman window is applied and shifted at every 5 ms to extract the mel-cepstrum and the LSP parameters for five speech sentences, and then they are filtered to synthesize speech without any signal processing. In other words, the speech is analyzed and then, just synthesized.

The comparison category rating (CCR) method is tested to compare both the synthesized speech. In the CCR test, which is an ITU recommendation [13], two different speech samples are presented to the listeners in random order. The listeners then compare the quality of the second speech relative to that of the first. Ten listeners tested and the results are shown in Table V.

From Table V, we can see that the listeners prefer the speech of the LSP if the order of analysis is same, although the mel-cepstrum has many good characteristics. Besides, the LSP is still competitive even though the order of the mel-cepstrum is greater. This result makes us decide to test the LSP for the feature of HMM-based speech synthesis.

V. Excitation Modeling

The HMM-based speech synthesis also needs to generate the excitation source signal. If the simple excitation model with the voiced/unvoiced selection is adopted, it is inevitable to get a typical quality of "vocoded sound". Yoshimura et al. incorporated the mixed excitation model used in the mixed excitation linear predictive (MELP) vocoder, which is originally developed for narrowband speech [14], and reported the improved speech quality [15].

TABLE V
Preference test results (MCEP: Mel-cepstrum)

Ref. Feature	Test Feature	Speaker	CCR vote		
			Better	Similar	Worse
MCEP-18	MCEP-24	F1	34	16	0
		F2	32	15	3
		F3	18	28	4
		M1	12	37	1
		M2	18	26	6
		M3	26	20	4
		sum	140	142	18
MCEP-18	LSP-18	F1	50	0	0
		F2	31	13	6
		F3	36	14	0
		M1	30	18	2
		M2	32	12	6
		M3	28	14	8
		sum	207	71	22
MCEP-24	LSP-18	F1	32	18	0
		F2	20	21	9
		F3	20	24	6
		M1	12	28	10
		M2	24	16	10
		M3	14	22	14
		sum	122	129	49

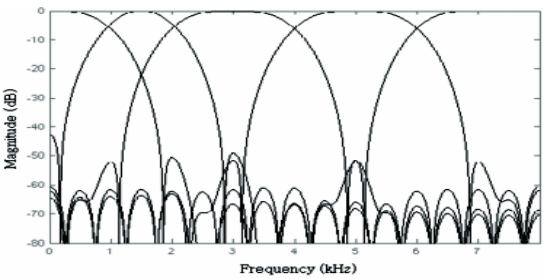


Fig. 2. Filterbanks in the mixed excitation for the wideband speech

A. Shortcoming of Mixed Excitation Model

The MELP vocoder is proposed by McCree et al. for low bit rate narrowband speech coding at 2.4 kbps [14], and has been chosen for the U.S. military standard speech coder in 1996 [16]. This vocoder has the following added capabilities: mixed pulse and noise excitation, the position jitter to reduce the tonal sound quality, adaptive spectral enhancement, and the pulse dispersion filter to match the filtered synthetic and natural speech waveform. More details are described in [14]. The MELP has been applied to the wideband speech coder [17], and Yoshimura et al. incorporated into the excitation model of HMM-based wideband speech synthesis [15]. However, the number of frequency bands for the band-pass voicing analysis is same as that of the narrowband MELP. The 5 bands of 0-0.5, 0.5-1, 1-2, 2-3, 3-4 kHz are simply extended to that of 0-1, 1-2, 2-4, 4-6, 6-8 kHz as shown in Fig. 2. These bands are not optimal even though the buzziness in the synthesized speech is surely reduced. If the number of analysis bands is increased to get finer resolution, more memory space would be required for the trained HMM data

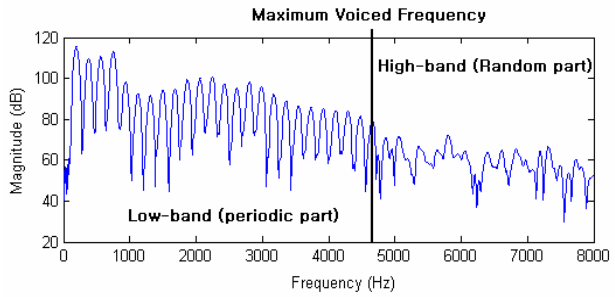


Fig. 3. Maximum voiced frequency in two-band speech model

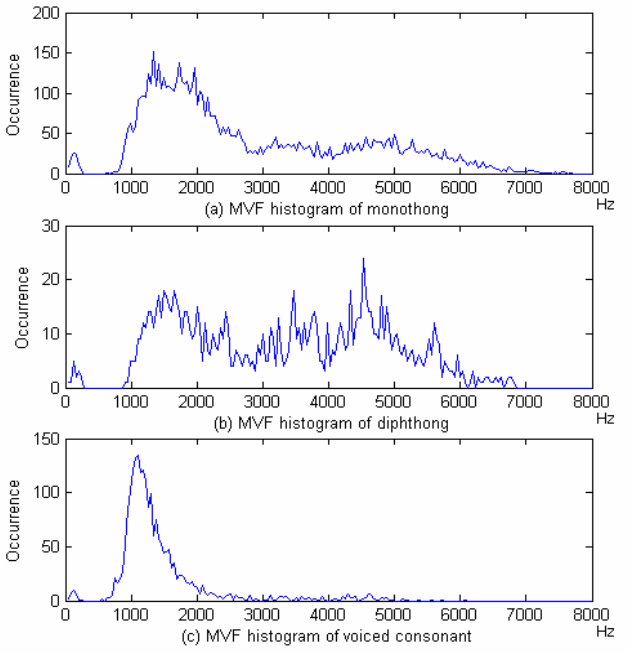


Fig. 4. Histogram of maximum voiced frequency

because of the delta and delta-delta parameters of the increased bands.

B. Two-band Speech Model

In the two-band speech model (or harmonic plus noise speech model), the bands are divided by the time-varying frequency referred to as "maximum voiced frequency (MVF)"[18]. The lower band has voiced characteristics and the upper band has unvoiced characteristics. If the MVF is misestimated too low, a large part of originally voiced region is synthesized as unvoiced signal and this leads to a harsh sound. On the contrary, if the MVF is misestimated too high, it leads to a buzzy sound. From this point of view, 5 bands of analysis in the mixed excitation are not enough for the wideband speech and finer frequency bands are required. An example of the MVF is shown in Fig. 3, and more details are described in [18]. Fig. 4 shows the histogram of the MVF calculated from 20 sentences.

C. Estimation of Maximum Voiced Frequency

Stylianou introduced a “harmonic test” to determine the MVF, but its heuristic threshold was not proper to the various

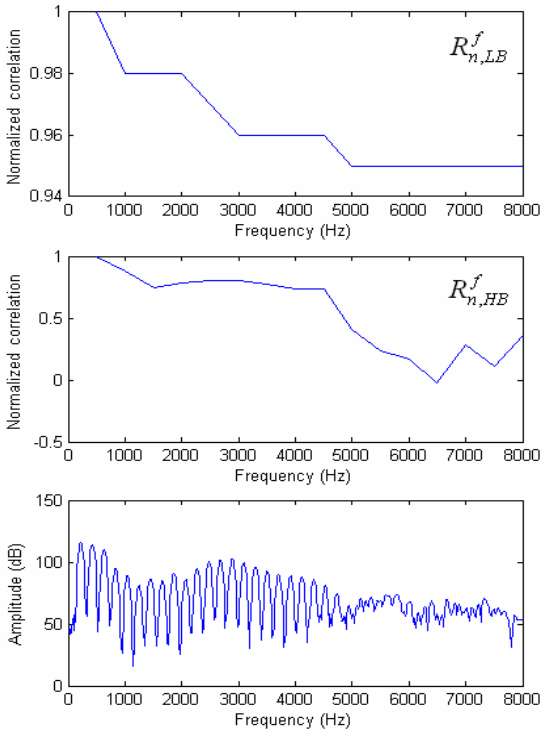


Fig. 5. Normalized correlation of the low- and high-band speech

speech databases. Instead, we utilized the normalized correlation of the high-pass filtered speech to find the MVF. If we define h^f_{LPF} and h^f_{HPF} is a low-pass and high-pass filter with a cutoff frequency, f, respectively, the filtered speech can be defined as follows:

$$s^f_{LB}(n) = h^f_{LPF} * s(n), \tag{1}$$

$$s^f_{HB}(n) = h^f_{HPF} * s(n), \tag{2}$$

$$s(n) = s^f_{LB}(n) + s^f_{HB}(n). \tag{3}$$

And their normalized correlation can be represented as:

$$R^f_{n,LB}(\tau) = \frac{\sum_{n=0}^{N-1} s^f_{LB}(n) s^f_{LB}(n+\tau)}{\sqrt{\sum_{n=0}^{N-1}\left\{s^f_{LB}(n)\right\}^2 \sum_{n=0}^{N-1}\left\{s^f_{LB}(n+\tau)\right\}^2}}, \tag{4}$$

$$R^f_{n,HB}(\tau) = \frac{\sum_{n=0}^{N-1} s^f_{HB}(n) s^f_{HB}(n+\tau)}{\sqrt{\sum_{n=0}^{N-1}\left\{s^f_{HB}(n)\right\}^2 \sum_{n=0}^{N-1}\left\{s^f_{HB}(n+\tau)\right\}^2}}. \tag{5}$$

Where, τ is the estimated pitch in a frame, and N is the pitch analysis window size. If the cutoff frequency, f, is smaller then the MVF, the filtered high-band speech, $s^f_{HB}(n)$, is periodic. So $R^f_{n,HB}$ would be close to 1. On the contrary, if

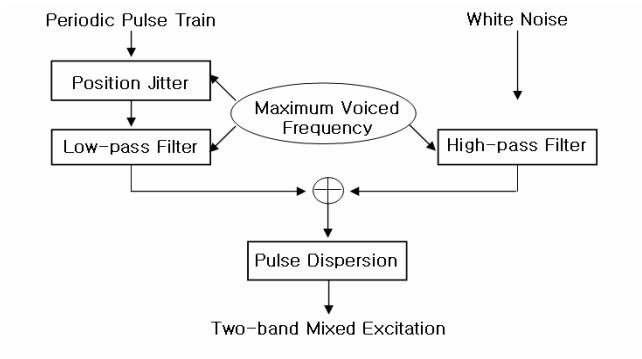


Fig. 6. Two-band excitation model.

the f is greater than the MVF, $s_{HB}^{f}(n)$ is aperiodic. So $R_{n,HB}^{f}$ would be close to 0. $R_{n,LB}^{f}$ and $R_{n,HB}^{f}$ values depending on f are shown in Fig. 5. In our system, the normalized correlation of the high-pass filtered speech, $R_{n,HB}^{f}$, is utilized to determine the MVF.

D. Two-band Excitation Analysis

15 pairs of the 6th-order Butterworth low- and high-pass filters are designed and their cutoff frequencies vary from 0.5 to 7.5 kHz with 0.5 kHz step increment. These filters will be used in the synthesis phase again. By applying the filters, the full band of 8 kHz is divided into the lower and the higher frequency band.

Firstly, voiced/unvoiced decision is performed and the pitch is estimated for each frame. If the input frame is voiced, $R_{n,HB}^{f}$ of the high-passed speech for each filter is calculated sequentially. And if $R_{n,HB}^{f}$ becomes less than 0.5, its cutoff frequency of the filter, i.e., the lowest cutoff frequency which satisfies the above condition, is used as an MVF. Because of the high-pass filtering, the processing would take time. But, it is the analysis phase for the training data preparation, so its calculation time is not a shortcoming. If the finer cutoff frequency steps are utilized, the finer MVF will be determined.

E. Two-band Excitation Synthesis

Fig. 6 shows the two-band excitation model incorporated into our HMM-based Korean speech synthesis system. It is basically similar to the mixed excitation for HMM-based speech synthesis suggested by Yoshimura et al., except the band-pass voicing strength decision and the shaping filter part. The low- and high-pass filters are the 6th-order Butterworth filters used in the analysis phase.

When the MVF is generated from the trained HMM data, the nearest frequency among the 16 frequency steps is selected. And its corresponding low- and high-pass filters are applied to the pulse and the white noise excitation, respectively. Then the filtered excitations are added together. The position jitter and the pulse dispersion filter are same with those of the MELP.

VI. Experiments and Evaluation

A. Speech Database and Training Condition

The training data, which are about the weather forecast, consist of carefully selected 540 sentences (about 110 minutes) uttered by a professional female announcer. They are recorded at 16 kHz sampling rate with 16 bit resolution. The phonemes are firstly automatically segmented, and then manually refined. The LSP parameter and the mel-cepstrum parameter are extracted from 15 ms and 25 ms speech frame, respectively, with the 5 ms frame shift. The 0^{th}-coefficient (in case of the LSP, the log energy) and the MVF are also calculated. These feature vectors are extended with delta and delta-delta features. We used 5 states left-to-right HMMs with no skip.

TABLE VI
Preference Test Results of the Spectral Parameters.

Ref. Feature	Test Feature	Test Sentence	CCR vote		
			Better	Similar	Worse
MCEP-18	LSP-18	database	19	25	6
		newspaper	15	24	11
MCEP-24	LSP-18	database	12	25	13
		newspaper	15	20	15

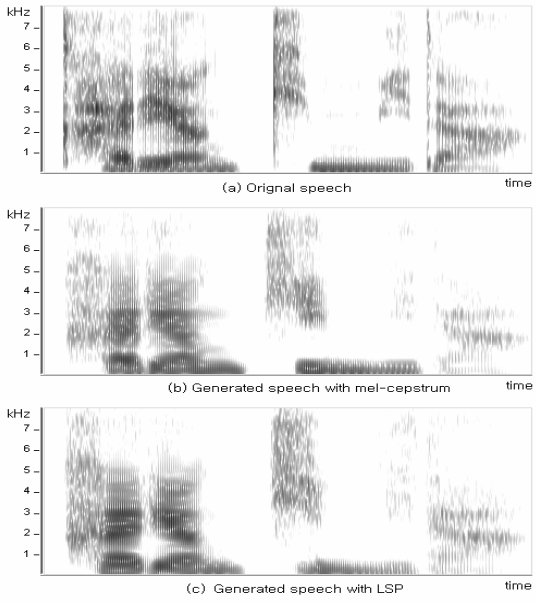


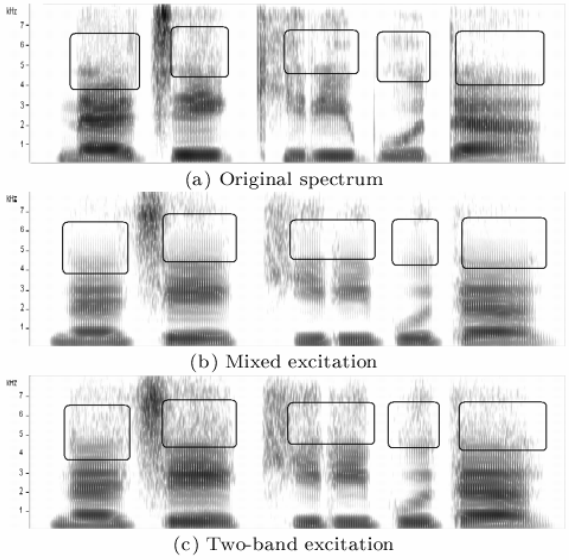
Fig. 7. Spectral parameter comparison.

B. Spectral Parameter Evaluation

Table VI shows the CCR preference results of the synthesized speech of the system between the LSP and the mel-cepstrum. In this test, the traditional excitation model of V/UV selection is utilized. Ten listeners tested ten sentences. Five sentences are from the database, and the others are chosen randomly from a newspaper.

TABLE VII
MOS LISTENING TEST RESULTS
(ME: MIXED EXCITATION, TBE:TWO-BAND EXCITATION).

Ref.	Test	Distance measurement		
		MFCC	Log-spectrum	SKL
Analysis/synthesis				
Original	ME	0.889	30.797	43.579
	TBE	0.883	31.965	43.738
HMM-based synthesis				
Original	ME	1.149	37.146	47.767
	TBE	1.135	37.040	46.958



(a) Original spectrum

(b) Mixed excitation

(c) Two-band excitation

Fig. 8. Spectrogram comparison of the excitation models

The results show that the LSP parameters could be a good alternative for the spectrum modeling of the HMM-based speech synthesis. The listeners, however, say that the synthesized speeches with both parameters have almost same quality and the difference is almost negligible. Fig. 7 shows an example of the spectrogram comparison. They are the generated speech of the system with 24th-order mel-cepstrum and the 18th-order LSP parameter.

C. Excitation Model Evaluation

Table VII shows the spectral distances among the original and the synthesized speech with the two excitation methods. 10 sentence speech pairs are time aligned. Then, cepstral distance with mel-frequency cepstral coefficients (MFCCs), spectral distance with log-spectrum, and symmetric Kullback-Leibler(SKL) distance [20] are calculated and averaged. In the analysis/synthesis the HMM training is not applied. We can see that the distances of the analysis/synthesis are smaller than those of the HMM-based synthesis. HMM training seems to affect the speech parameters and to result in the degraded outputs. In case of the analysis/synthesis, the mixed excitation method looks slightly better than the other one. When the speech parameters, however, are trained with HMMs, the results are changed. The output with the suggested two-band excitation method is slightly better for the HMM-based speech synthesis than the mixed excitation method although the difference of the distance between the two methods is very small. Fig. 8 shows an example of the spectrogram comparisons among the original and the HMM-based synthesized speech with the two different excitation methods.

TABLE VIII
PREFERENCE TEST RESULTS OF THE EXCITATION MODELS.
(TE:TRADITIONAL EXCITATION)

Ref. Excitation	Test Excitation	Test Sentence	CCR vote		
			Better	Same	Worse
TE	ME	database	48	2	0
		newspaper	45	5	0
	TBE	database	48	2	0
		newspaper	45	5	0
ME	TBE	database	10	37	3
		newspaper	12	30	8

TABLE IX
MOS LISTENING TEST RESULTS.

Spectrum &Excitation	Test Sentence	MOS		
		Naturalness	Clearness	Overall quality
MCEP-18 & TE	database	2.78	2.81	2.81
	newspaper	2.53	2.13	2.36
MCEP-18 & TBE	database	2.94	3.43	3.13
	newspaper	2.65	2.98	2.85
LSP-18 & TBE	database	2.98	3.51	3.28
	newspaper	2.78	3.19	3.08

TABLE X
BINARY FILE SIZE OF THE SYNTHESIS SYSTEM.

module		size(kbyte)
trained HMM data	spectrum	1,292
	excitation	160
	duration	14
decision tree data	spectrum	254
	excitation	304
	duration	37
synthesizer engine		73
total		2,134

Table VIII shows the preference listening test results among the excitation models. In this test, the 24th-order Mel-cepstrum is utilized for the spectrum analysis. From the results, we can see that both the mixed excitation methods are preferred to the traditional one. And the two-band excitation method shows slightly better results than the mixed one. Even if the output speech is comparable, the suggested excitation is still meaningful because it can reduce the excitation parameter order for the data.

The MVF of two-band excitation model is one scalar value to divide the frequency into a voiced and an unvoiced region. It can generate the mixed excitation signal with a finer resolution, and reduce the memory of the excitation parameters for the trained HMM data.

Overall Synthesized Speech Evaluation

Table IX shows the conventional mean opinion score (MOS) listening test results on a scale from 1 to 5. We can see that the two-band excitation surely enhanced the synthesized speech quality, especially the clearness. We also found that the LSP analysis could be a good alternative for the HMM-based speech synthesis. Besides, according to our experimental results, the LSP looks slightly better than the mel-cepstrum. Table X shows the binary file size of the synthesis system. Even though the size of the system is only about 2MB, the synthesized speech quality is fairly good.

VII. CONCLUSIONS

This paper presents our HMM-based Korean speech synthesis system. Refined contextual information utilized in the system is described. The LSP parameter is tested and compared with the mel-cepstrum for the synthesis system. The two-band excitation model is also described. Finally, the LSP analysis and the two-band excitation models are incorporated into the system for the spectrum and the excitation generation, respectively, and produced best results. With only about 2MB data, fairly good quality speech is synthesized and shows the possibility of an embedded TTS for hand-held mobile devices. We are planning to test some more different speech databases with the system to generalize our results further in the near future.

REFERENCES

[1] A.J. Hunt and A.W. Black, “Unit selection in a Concatenative speech synthesis system using a large speech database,” *Proc. IEEE ICASSP*, pp.959-962, 1996.

[2] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features", *Proc. IEEE ICASSP*, vol.1, pp.660-663, 1995.

[3] K. Tokuda, H. Zen, and A.W. Black, "Chapter 7. An HMM-based approach to multilingual speech synthesis", *Text to Speech Synthesis: New Paradigms and Advances*, Prentice Hall, pp.135-153, 2004.

[4] K. Tokuda, H. Zen, A.W. Black, "An HMM-Based Speech Synthesis System Applied to English," *Proc. IEEE Workshop on Speech Synthesis*, pp.227-230, 2002.

[5] R.S. Maia, H. Zen, K. Tokuda, T. Kitamura, and F.G.V. Resende, Jr., “Towards the development for a Brazilian Portuguese text-to-speech synthesis system based on HMM,” *Proc. Eurospeech*, pp.2465-2468, 2003.

[6] B. Vesnicer, and F. Miheliv, "Evaluation of the Slovenian HMM-Based speech syntehsis system", *Proc. TSD2004, LNAI3206*, pp.513-520, 2004.

[7] 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, pp1116-1119, 2006.

[8] Fukada, T., Tokuda, K., Kobayahsi, T., and Imai, Satoshi, "An Adaptive Algorithm for Mel-Cepstral Analysis of Speech", *Proc. IEEE ICASSP*, vol.1, pp137--140, 1992.

[9] S.H. Kim, Y.J. Lee, K. Hirose, "A New Korean Corpus-Based Text-to-Speech System," *International Journal of Speech Technology*, vol.5, pp.105-116, 2002.

[10] S.H. Kim, Y.J. Lee, K. Hirose, “Unit Generation Based on Phrase Break Strength and Pruning for Corpus-Based Text-to-Speech,” *ETRI Journal*, vol.23, no.4, 2001.

[11] J.J Kim, M.O. Choi, S.S. Oh, S.H. Kim, J. Park, Y.J. Lee, “Introduction of the ETRI Dialog-style TTS System,” *Proc. KSPS(the Korean Society of Phonetic Sciences and Speech Technology)*, pp.79-82, December, 2004.

[12] Koishida, K., Tokuda, K., Kobayashi, T., and Imai, S., "Spectral Representation of Speech Based on Mel-Generalized Cepstral Coefficients and Its Properties", *Elec.&Comm. in Japan*, vol.83, no.3, pp1999-2006, 2000.

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

[14] A.V. McCree, T.P. Barnwell III, "A mixed excitation linear LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech and Audio Processing*, vol.3, no.4, pp.242--250, 1995.

[15] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, “Mixed excitation for HMM-based speech synthesis,” *Proc. Eurospeech*, vol.3, pp.2263--2266, 2001.

[16] L.M. Supplee, R.P. Cohn, J.S. Collura, A.V. McCree, "MELP: The New Federal Standard at 2400bps," *Proc. ICASSP*, vol.2, pp.1591-1594, 1997.

[17] W. Lin, S. N. Koh and X. Lin, “Mixed excitation linear prediction coding of wideband speech at 8kbps” *Proc. ICASSP*, vol.2, pp.1137-1140, Jun. 2000.

[18] Y. Stylianou, "Modeling speech based on harmonic plus noise models," *Nonlinear speech modeling, LNAI3445*, pp.244-260, 2005.

[19] S-.J. Kim, M. Hahn, “Two-band Excitation for HMM-based Speech Synthesis,” *IEICE trans. Inf.&Syst.*, submitted for publication.

[20] E. Klabbers, R. Veldhuis, "Reducing Audible Spectral Discontinuities," *IEEE Trans. Speech and Audio Processing*, pp.39-51, 2001.

Sang-Jin Kim received the B.S. degree in electronic engineering from Inha University, Incheon, South Korea, in 2000, and the M.S. degree in electronic engineering from Information and Communications University (ICU), Daejeon, South Korea, in 2002. He is currently a Ph.D. candidate in school of engineering at ICU. His research interests include speech synthesis, recognition, coding, enhancement and their applications.

Jong-Jin Kim received the B.S. and the M.S. degrees in computer engineering from Wonkwang University, Chunbuk, South Korea, in 1995 and 1997, respectively. Currently he is with Electronics and Telecommunications Research Institute (ETRI), Deajeon, South Korea.
His research interests include speech synthesis, speech coding and speech recognition.

Minsoo Hahn received the B.S. and the M.S. degrees in electrical engineering from Seoul National University, Seoul, South Korea, in 1979 and 1981, respectively, and the Ph.D. degree in electrical and electronics engineering from University of Florida, Florida, USA, in 1989. From 1982 to 1985, he was with Korea Research Institute of Standards and Science (KRISS), Deajeon, South Korea. From 1990 to 1997, and he was with Electronics and Telecommunications Research Institute (ETRI), Deajeon, South Korea. In 1998, he has been a faculty member of the School of Engineering, Information and Communications University (ICU). Currently, he is a Full Professor in School of Engineering, ICU and a Director in Digital Media Laboratory, ICU. His research interests include speech and audio coding, speech synthesis, noise reduction and VoIP. Recently, he is located general chair for Speech Engineering Community, South Korea.