Implementation and Evaluation of an HMM-Based Speech Synthesis System for the Tagalog Language

Quennie Joy Mesa(한남대), 김경태(한남대), 김종진(ETRI)

<차 례>

<Abstract>

Implementation and Evaluation of an HMM-Based Speech Synthesis System for the Tagalog Language

Quennie Joy Mesa, Kyung-Tae Kim, Jong-Jin Kim

This paper describes the development and assessment of a hidden Markov model (HMM) based Tagalog speech synthesis system, where Tagalog is the most widely spoken indigenous language of the Philippines. Several aspects of the design process are discussed here. In order to build the synthesizer a speech database is recorded and phonetically segmented. The constructed speech corpus contains approximately 89 minutes of Tagalog speech organized in 596 spoken utterances. Furthermore, contextual information is determined. The quality of the synthesized speech is assessed by subjective tests employing 25 native Tagalog speakers as respondents. Experimental results show that the new system is able to obtain a 3.29 MOS which indicates that the developed system is able to produce highly intelligible neutral Tagalog speech with stable quality even when a small amount of speech data is used for HMM training.

* Keywords: Tagalog, HMM, Corpus, Training, Synthesis, Mean opinion score (MOS).

1. Introduction

Over time, the dominant speech synthesis techniques have been evolved from physical and acoustic models to data-driven models. This enables the creation of synthetic voices that sound more natural and possess recognizable identities, but the flexibility of such systems is limited by the amount and type of data collected [1]. With these given limitations of the data-driven techniques, there is now an emerging trend to reincorporate some of the flexibility of the statistical model-based techniques to reduce the amount of necessary data. The hidden Markov model (HMM) [2][3], one of the statistical time series models, presents a tangible way to synthesize speech data.

In the HMM-based approach to speech synthesis, as originally proposed by Tokuda *et. al.* [4][5], speech parameters such as spectrum, fundamental frequency and phoneme duration, used in the synthesis stage are generated directly from HMMs by using a parameter generation algorithm [6].

These techniques have the advantage of providing a means to automatically train the specification-to-parameter module, bypassing the problems associated with hand-written rules, thus produce high quality of synthesized speech [3][4][7]. Furthermore, they have the advantage of being compact and amenable to modification for voice transformation and other purposes [3][4][5][7].

Recently, we can find trainable synthesis systems for Japanese [6], English [8] and a few other languages [9][10]. As for the Tagalog speech synthesis research, so far only experiments using phoneme concatenation synthesis have been reported [11].

Tagalog is the most widely spoken language of the Philippines. It is native to the southern part of the island of Luzon, and is historically only one of several fairly widely spoken languages of the region. The Ethnologue reported that as of 1995, there were 15.9 million Tagalog native speakers in the world of which 14.5 million were in the Philippines [12]. Moreover, according to the 1990 and 2000 United States Census, Tagalog is the second most commonly-spoken Asian language (after Chinese) in the United States and the sixth non-English language spoken in America [13].

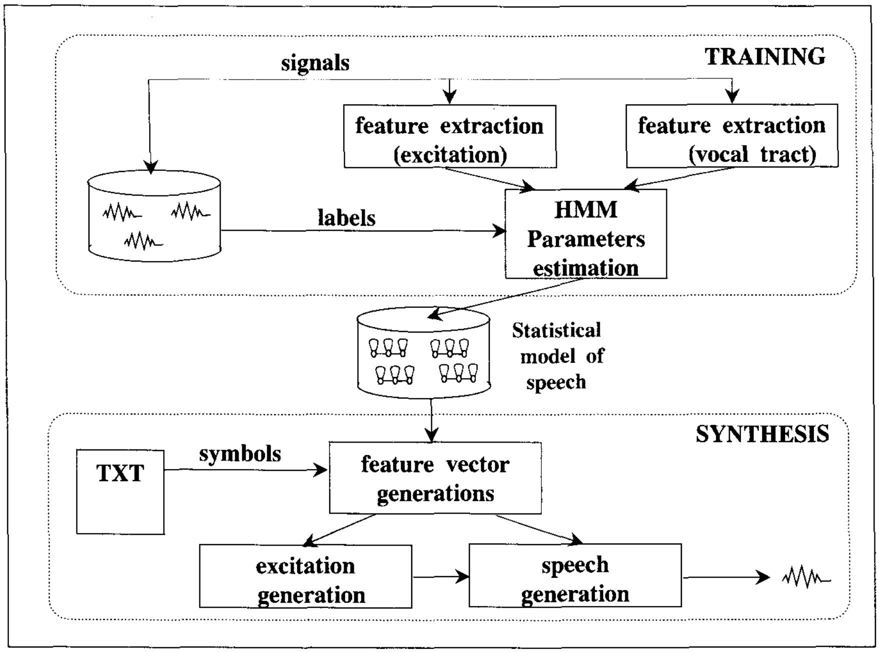
In this paper, we present an HMM-based Tagalog speech synthesis system. The aim of this research is to build a Tagalog synthesizer. The research focuses on the generation of speech for a given word string of known pronunciation. Text-to-phoneme conversion and fundamental frequency contour generation are not attempted.

The rest of this paper is organized as follows. Section 2 summarizes an HMM-based speech synthesis technique. Section 3 concerns with the development of the Tagalog synthesizer. The training database and the experimental results are discussed in Section 4 and the conclusion is given in Section 5.

2. HMM-Based Speech Synthesis Technique

The HMM-based approach to speech synthesis differs from the other techniques because it uses the statistical framework of HMMs not only for segmentation and labeling of the database but also as a model of speech production [4].

The text-to-speech (TTS) synthesis procedure consists of two main phases. As shown in the upper part of <Figure 1>, there is a training step, wherein the statistical model of speech is estimated. In addition, at the lower part of the figure, there is a synthesis step, where the speech signal is generated.


signals
TRAINING
feature extraction (excitation)
feature extraction (vocal tract)
labels
HMM Parameters estimation
Statistical model of speech
SYNTHESIS
TXT
symbols
feature vector generations
excitation generation
speech generation

<Figure 1> Schematic diagram of HMM-based training and synthesis procedure [9]

2.1. Training Part

The training part includes spectral and excitation parameter extraction. The feature vectors of extracted mel-cepstrum and fundamental frequency (F0) parameters, together with their dynamic feature, are concatenated and used for training context-independent and context-dependent acoustic models [3].

The training of phone HMMs using pitch and mel-cepstrum simultaneously is enabled in a unified framework by using multi-space probability distribution HMMs and multi-dimensional Gaussian distributions [7].

2.2. Synthesis Part

In the synthesis part of the HMM-based speech synthesis, an arbitrarily text to be synthesized is converted to a context-based label sequence. Then, according to the label sequence, a sentence HMM is constructed by concatenating context dependent HMMs. State durations of the sentence HMM are determined so as to maximize the output probability of state durations [14] and then a sequence of mel-cepstral coefficients and log(F0) values including voiced/unvoiced decisions is determined in such a way that its output probability for the HMM is maximized using the speech parameter generation algorithm [6]. Finally, speech waveform is synthesized directly from the generated mel-cepstral coefficients and F0 values by using the mel log spectrum approximation (MLSA) filter [15].

3. Tagalog Speech Characteristics

3.1. Tagalog Phonological System

In this section, the most important features of the Tagalog language from the speech synthesis perspective are discussed.

Tagalog has 26 phonemes; five vowels (<Table 1>), six dipthongs (<Table 2>), and 17 consonants (<Table 3>). Tagalog is close to a "phonetic language" [16]. The Tagalog writing system follows the phonemic principle: there exists a different grapheme corresponding to each phoneme that is not used for any other purpose [17]. This fact makes many aspects easier in speech and in language research. For example, the conversion of written text to phoneme strings in speech is straightforward.

<Table 1> Articulatory properties of Tagalog vowels [16]

Tongue Position	Front	Central	Back
High	*i*		*u*
Mid	*e*		*o*
Low		*a*	

<Table 2> Articulatory properties of Tagalog diphthongs [18]

Sound Position	Front	Central	Back
High	*iw*		*uy*
Mid	*ey*		*oy*
Low		*aw, ay*	

<Table 3> Articulatory description of Tagalog consonants [16]

Phonetic Description	Phones
Voiced Bilabial Plosive	*b*
Voiced Dental or Alveolar Plosive	*d*
Voiced Velar Fricative	*g*
Voiceless Velar Fricative	*k*
Voiceless Bilabial Plosive	*p*
Voiceless Dental or Alveolar Plosive	*t*
Voiceless Postalveolar Affricate	*ts*
Voiceless Alveolar Palatal Fricative	*s*
Voiced Bilabial Nasal	*m*
Voiced Dental Nasal	*n*
Voiced Velar Nasal	*ng*
Retroflex Alveolar Lateral	*l*
Voiced Alveolar Trill	*r*
Voiced Labial-Velar Approximant	*w*
Front Alveolar Lateral	*y*

Number of words: The morphology of the Tagalog language is complex [18], which makes the number of different words immense. This is due to several factors:

(a) Tagalog is an agglutinative language [18]. The meaning of a word is typically changed by adding affixes to the word. These kinds of affixes are used with nouns, pronouns, adjectives, numerals and even verbs.

(b) New words are formed by composing other words [17].

Quantity of Phonemes: Phoneme duration is a distinctive feature in Tagalog. The meaning of a word can be changed by varying the temporal durations of the phonemes [17][18].

3.2. Contextual Information

The richness and appropriateness of the prosodic patterns depend on the number and type of the considered contextual factors [9]. In this work, the following contextual informations are considered.

Phoneme

- {preceding, current, succeeding} phoneme
- Position of current phoneme in current syllable

Syllable

- Number of phonemes at {preceding, current, succeeding} syllable
- Accent of {preceding, current, succeeding} syllable
- Stress of {preceding, current, succeeding} syllable
- Position of current syllable in current word
- Number of {preceding, succeeding} stressed syllables in current phrase
- Number of {preceding, succeeding} accented syllables in current phrase
- Number of syllables {from previous, to next} stressed syllable
- Number of syllables {from previous, to next} accented syllable
- Vowel within current syllable

Word

- Guess at part of speech {preceding, current, succeeding} word
- Number of syllables in {preceding, current, succeeding} word
- Position of current word in current phrase
- Number of {preceding, succeeding} content words in current phrase
- Number of words {from previous, to next} content word

Phrase

- Number of syllables in {preceding, current, succeeding} phrase
- Position in major phrase
- TOBI end tone of current phrase

Utterances

- Number of syllables in current utterance

3.3. Context and Prosody Clustering

It has been argued and proven in practice that as contextual factors increases, their combinations also increases exponentially. Hence, model parameters cannot be estimated accurately with limited training data. Moreover, it is impossible to prepare a database which will include all combinations of contextual factors [9].

To reduce the number of parameters the decision-tree based context clustering [19] has been performed.

4. Experiments

4.1. Speech Corpus

The speech corpus is the essential part of all spoken technology systems. The quality and volume of speech data in the corpus defines the performance of the system [1]. Enough speech data is essential in all statistical approaches to speech modeling such as HMMs in order to estimate all the parameters of the model. The training of the HMM models for speech synthesis is based on speech utterances and their transcription.

Ideally, in the construct of a speech corpus, the training data should reflect the task for which the synthesizer is to be used [1], since using training data similar in style and contents to that of the data to be synthesized ensures an appropriate balance of acoustic information across contexts likely to be encountered during the synthesis. Problems can arise when contexts required in synthesis is different from anything seen in training. For this reason some degree of database preparation to ensure a minimum coverage of possible contexts is recommended, though it is not performed in this paper.

4.1.1. Manuscript selection

We have used text gathered from the news articles that were in Tagalog from the Internet. News articles were chosen because, although the grammar and structure of the Tagalog language utilized here were not the same as the spoken language, it had informal content, and included everyday as well as non-grammatical expressions, thus approximating the natural language better. Also, there was a relative abundance of these data in the web, compared to other documents that were in Tagalog. Each news clip contained 5~47 sentences. Each sentence contained 2~13 words and had either one of the following themes: (1) relating a news events, (2) describing a scene, (3) relating an emotional event, and (4) story telling.

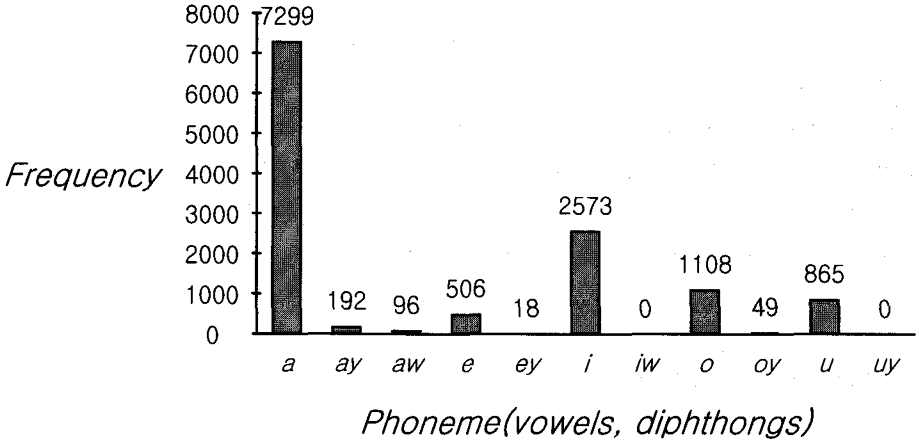
The speaker read 28 questions, 10 exclamations and 558 declarative sentences. The speaker read a total of 6047 words in the recording session. These data were collected for about two weeks.

For readability we put several limits on the sentence length:

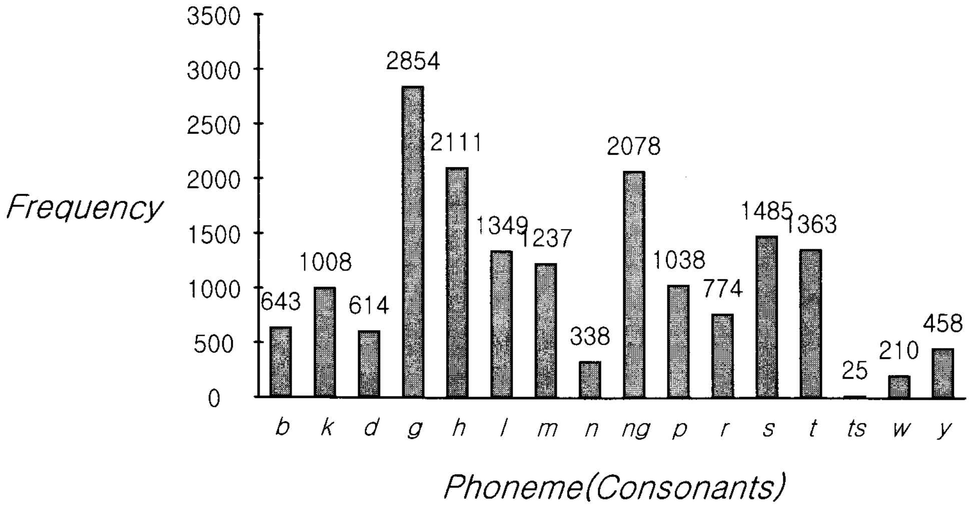
- Maximum word length: 20 characters
- Maximum number of words in a sentence: 13 words

4.1.2. Corpus analysis

In parallel with corpus compilation a constant process of analysis was going on to diagnose its possible weak points. First we had to find out if all Tagalog phonemes, existing as well as theoretically, were presented in the speech corpus and with what frequency. <Figure 2> and <Figure 3> show the frequency of occurrences of vowels and dipthongs, and that of consonants in the speech corpus, respectively. The diphthongs /iw/ and /uy/ were missing in our corpus, however possible theoretically.



<Figure 2> Frequency of occurrences of vowels and diphthongs in the speech corpus



<Figure 3> Frequency of occurrences of consonants in the speech corpus

4.1.3. Speaker selection

The main criterion for the speaker selection was the ability to read out the whole text at normal speaking rate and pronunciation. As a result, a female Tagalog native speaker was chosen.

4.1.4. Recording

The recordings were made in a sound proof and low echo room at a sampling rate of 44.1 kHz with 16 bits resolution by using a professional recording equipment. Each speech utterance was later down sampled to 16 kHz for our experiment. The microphone and software used were as follows.

- Desktop PC: Hard Recording Workstation 2.4 GHz, Pentium 4.
- Microphone: SENNHEISER HMD 280
- Software for recording: Cool Edit Pro 2.0

4.1.5. Manual annotation

The corpus was manually annotated using Praat [20]. The annotation was performed with the following steps:

- Manuscript conversion

- Manual phonemic conversion
- Manual correction of phonemic annotation
- Manual phonemic labeling

The speech utterances were transcribed on the word level, so before training we had to perform initial phone level segmentation. Phone level segmentation was achieved using manual alignment of speech signal and word transcription. The phoneme set employed in this system consisted of 28 phonemes, including one silence and one pause model. Despite the absence of diphthongs in some known Tagalog phoneme sets employed for TTS purposes [11], some diphthongs were considered in this paper because of the hard task involved in separating the vowels from the semi-vowels during the phonetic segmentation. <Table 1>, <Table 2> and <Table 3> show the phoneme set which was employed to label the database, excluding silence and pause models.

4.2. Training Condition

In this paper, feature extraction has been performed on 25 ms long Hamming-windowed speech frames with 5 ms frame shift. Every frame was analyzed by a 512-point FFT, and 24 mel-frequency cepstral coefficients (MFCCs) including the zeroth coefficient and log F0 were calculated. Lower and upper limit for F0 extraction were 60 Hz and 400 Hz, respectively.

We used 5 states left-to-right HMMs without skips, where the first and the last states were non-emitting states. First, we trained 28 monophone models, 26 for the Tagalog phoneme set and 2 for special events like silence and pause.

The triphone models were made out of monophone models and trained. Then, the state-tying procedure based on Tagalog phonetic rules was performed, where the phonetic rules were used for describing the class of allophones according to their articulatory and acoustic characteristics.

State duration densities were estimated on the trellis which was obtained in the embedded training stage. State durations of each phoneme HMM were regarded as a multi-dimensional observation, and the set of state durations of each phoneme HMM was modeled by a multi-dimensional Gaussian distribution. Dimensions of the state duration densities were equal to the number of state of HMMs, and the n-th dimension of state duration densities was corresponding to the n-th state of HMMs [14]. The last step in the training part was parameter generation for unseen triphones according to the Tagalog phonetic decision trees.

4.3. Speech Synthesis

The synthesis part was processed as follows. First, the given text to be synthesized was converted into a contextual label sequence. Then, according to such label sequence, an HMM sequence was constructed by concatenating context-dependent HMM. After this, state durations for the HMM sequence were determined so that the output probability of the state durations was maximized. According to the obtained state, the sequence of the mel-cepstral coefficients and log F0 values including voiced/unvoiced decisions were determined by maximizing the output probability of HMM. Finally, speech waveform was synthesized directly from the generated mel-cepstral coefficients and F0 values by using the MLSA filter.

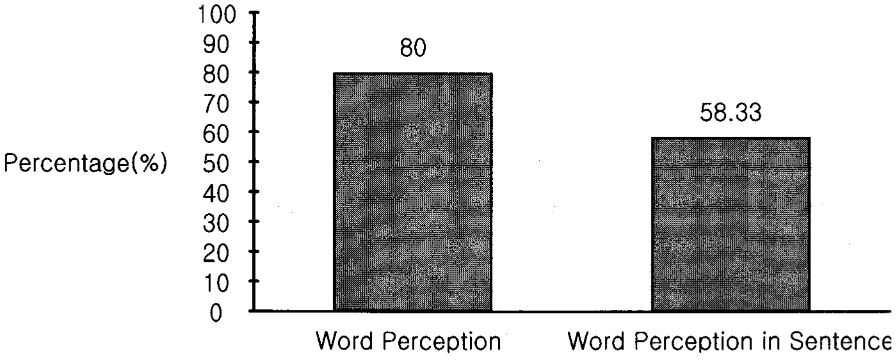
4.4. Evaluation of Synthesized Speech

In evaluating the quality of the synthetic voice produced by the developed system, formal listening test has been carried out. The test required the listeners to rank the voice quality using a mean opinion score (MOS) like scoring.

In our experiment, the synthesized speech material was presented to 25 (including 6 females) Tagalog subjects aged between 20 to 40 years old. Nobody has had former experiences in the field of spoken language technology.

The MOS test was carried out by synthesizing a set of 20 pairs of words with varying levels of confusability for the word perception tests. Also, 112 words were synthesized for the 12 sentences used for the word perception in a sentence test.

The test was divided into 3 parts. The first part consisted of twenty pairs of words with different levels of confusability. The respondents listened to the words at a time, and were asked to mark on the answer sheets which of the words they thought was correct. For the second part, the respondents were asked to listen to 12 distinct sentences and were instructed to write down what they have heard from the audio file. For the third part, the respondents were asked to evaluate the system by giving ratings in terms of intelligibility, speed, pronunciation, naturalness, articulation, ease of listening, comprehension and pleasantness. The listeners were required to make a single rating from five choices: excellent (5), good (4), fair (3), poor (2), and bad (1). A naive listening test was performed. The result (<Figure 4> and <Figure 5>) implies the following:



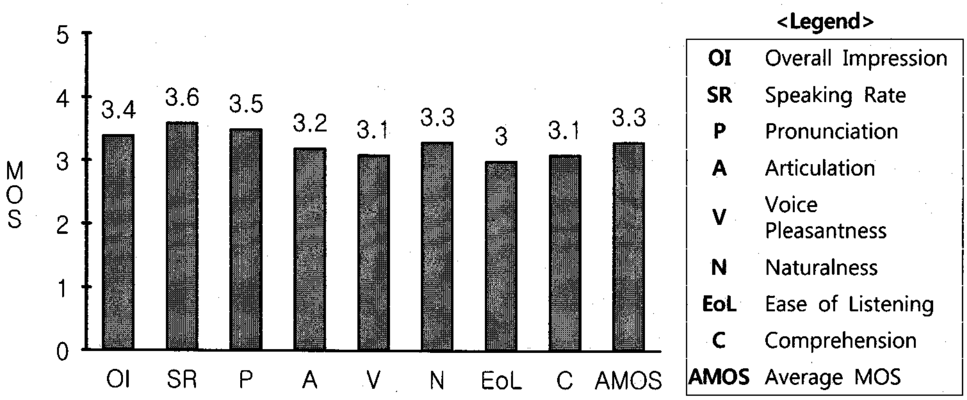
< Figure 4 > Intelligibility Test Result

The intelligibility test results (<Figure 4>) were promising. 80% of the words in the word perception test and 58.33% of the words in the word perception in a sentence test, were correctly identified by the respondents.

An error rate of 20% was obtained in the word perception test. This could be attributed to the fact that in general words without their contexts were difficult to perceive. Context was the main factor that affects the perception rate.

The reason why a 41.67% error rate from the word perception in sentence test was obtained was due to the sentence length. The longer the sentence was the more difficult to perceive all the words it contained.

From the MOS test results (<Figure 5>), the system in general obtained an average of 3.29 MOS score which indicates that the HMM based system was able to produced relatively good speech in the sense of naturalness and intelligibility with a small amount of trained data.



<Figure 5> MOS Test Result

5. Conclusion

In this paper we presented the HMM-based Tagalog speech synthesis system, wherein Tagalog is the most widely spoken indigenous language of the Philippines. The text-to-speech system was trained on approximately 89 minutes of Tagalog speech organized in 596 spoken utterances. In this system, database recording and labelling were carried out to build the synthesizer. Furthermore, contextual informations were determined. The system was able to produce highly intelligible neutral Tagalog speech with stable quality even with a small amount of speech data used for training. The quality of the synthetic voiced produced by the system was measured using MOS tests. An average of 3.29 MOS score was obtained, an error rate of 20.0% (for the word perception tests) and 41.67% (for the word perception in the sentence tests) were also recorded.

This paper introduced the first version of the Tagalog HMM-based speech synthesis system. As a next step we would like to record another speech corpus to ensure minimum coverage of the phonemic embodiment of the Tagalog language (which will include all acceptable phoneme combinations e.g. /iw/, /uy/ which were missing in our speech corpus). Furthermore, future works will also include the development of a text processing module, a grapheme-to-phoneme conversion module, construction of phonetically balanced sentences for the speech corpus, segmentation of more utterances to increase the training database and improvement on the contextual informations.

Acknowledgement

The authors gratefully acknowledge the generous assistance and valuable information provided to us by Dr. J. Park, Division Chief of the Spoken Technology Research Team at ETRI, and Professor Y. J. Lee, Director of SiTec (Speech Information Technology and Industry Promotion Center) at Wonkwang University for allowing us to use their Speech Laboratory for the recordings of the speech corpus. We are also grateful to the Digital Signal Processing Team of the University of Philippine for making their speech corpus publicly available.

This paper has been supported by the 2008 Hannam University Research Fund (2008A095).

References

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

[2] L. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[3] S. Young, G. Everman, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, *The HTK Book Version 3.2.1*, Dec. 2002.

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

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

[6] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis", *Proc. ICASSP*, pp. 1315-1318, 2000.

[7] K. Tokuda, T. Masuko, N. Miyazaki, T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling", *Proc. ICASSP*, pp. 229-232, 1999.

[8] K. Tokuda, H. Zen, A. W. Black, "An HMM-based speech synthesis applied to English", *Proc. IEEE Workshop on Speech Synthesis*, pp. 227-230, 2002.

[9] S. J. Kim, J. J Kim, M. Hahn, "Implementation and evaluation of an HMM-based Korean speech synthesis system", *IEICE Transactions on Information and Systems*, Vol. E89-D, No. 3, pp. 1116-1119, 2006.

[10] B. Vesnicer, F. Mihelic, "Evaluation of the Slovenian HMM-based speech synthesis system", *Lecture Notes in Artificial Intelligence*, Vol. 3206, pp. 513-520, 2004.

[11] M. Corpus, J. Liampo, M. Co, R. Guevara, "Development of a Filipino TTS system using concatenative speech synthesis", *Proc. 2nd National ECE Conference*, 2001, and also available at http://www.upd.edu.ph/~dsp/DSP_research_compilation/ ttsfullpaper_rev.pdf.

[12] http://www.ethnologue.com.

[13] http://www.census.gov/prod/2003pubs/c2kbr-29.pdf.

[14] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Duration modeling in HMM-based speech synthesis system", *Proc. ICSLP*, Vol. 2, pp. 29-32, 1998.

[15] T. Fukada, K. Tokuda, T. Kobayashi, S. Imai, "Adaptive algorithm for mel-cepstral analysis of speech", *Proc. ICASSP*, vol. 1, pp. 137-140, 1992.

[16] R. Guevara, M. Co, E. Espina, I. Garcia, E. Tan, R. Ensomo, R. Sagun, "Development of a Filipino speech corpus", *Proc. 3rd National ECE Conference*, 2002, and also available at www.upd.edu.ph/~dsp/DSP_research_compilation/d006-guevara-development.pdf.

[17] http://wika.pbwiki.com/f/ORTOPDF.pdf.

[18] T. Ramos, *Tagalog Structures*, University of Hawaii Press, 1971.

[19] J. J. Odell, *The Use of Context in Large Vocabulary Speech Recognition*, Ph.D. Dissertation, Cambridge University, 1995.

[20] http://www.fon.hum.uva.nl/praat/.

접수일자: 2008년 10월 28일
게재결정: 2008년 12월 6일

▶ Quennie Joy Mesa
주소: 306-791, Ojung-dong, Daejeon City, Korea
소속: Dept. of Computer Engineering, Hannam University
Isabela State University, Philippines
전화: 042) 629-8024
E-mail: quincy.quennie@gmail.com

▶ 김경태 (Kyung-Tae Kim) : Corresponding Author
주소: 306-791, Ojung-dong, Daejeon City, Korea
소속: Dept. of Information and Communication Engineering, Hannam University
전화: 042) 629-7574
E-mail: ktkim@hnu.kr

▶ 김종진 (Jong-Jin Kim)
주소: 305-350, 161 Gagung-dong, Yuseong-gu, Daejeon, Korea
소속: ETRI
전화: 042) 860-5759
E-mail: kimjj@etri.re.kr