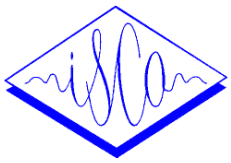


First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

SPEECH SYNTHESIS USING DEMISYLLABLES FOR KOREAN: A PRELIMINARY SYSTEM

Jung-Chul Lee, Yong-Ju Lee, Hee-il Han, Eung-Bae Kim, Chang-Joo Kim and Kyung-Tae Kim

Signal Processing Section
Electronics & Telecommunications Research Institute
Daejon, Korea

ABSTRACT

This paper describes a preliminary version of the text-to-speech system for Korean, which is referred to as "Geul-Sori". Input letters include Korean characters, numerals, and punctation marks. As a synthetic unit, we use the demisyllables and the control parameters of prosody are generated by rule. Geul-Sori was implemented on IBM-PC/AT using the TMS320C25 DSP chip. Geul-Sori is well operated, but yet poor in naturalness of speech. We are now developing many kinds of rules to achieve more naturalness and more intelligibility of the synthesized speech.

I. INTRODUCTION

As speech synthesis comes to play an important role in information services, the demand for the text-to-speech synthesizer is increasing rapidly. Since a number of laboratory English text-to-speech synthesizer including MITalk-79[1,2] has been developed in 1970s, nowadays there are a number of commercially available text-to-speech systems. In Japan, Hakoda *et al*.[3] used the line spectrum pairs(LSP) representations of LPC parameters and residual signals instead of white noise as a source of unvoiced sound to improve the quality of speech. Recently, multi-lingual text-to-speech system has been developed[4].

In Korea, since 1986 some research activities to develop the text-to-speech system for Korean has been in universities , institute and companies. Despite of short time of research, the quality of the synthetic speech for some text-to-speech systems is acceptable. But there are some problems in both lingistic and prosodic process. On the other hand, we have been developed Geul-Sori at ETRI, which converts unrestricted Korean text into a speech. The Korean language, also termed "Hangeul", has its own characteristics from the point of veiw for text-to-speech synthesis[5,6].

In this study, we make rules for detection of boundary information of phrases and clauses. And we also study prosodic processing according to boundary information and phonological processing by both letter-to-sound rule and exception dictionary. As a synthetic unit, the demisyllable speech unit , which makes it possible to realize with current technology while preserving the acceptable quality of synthetic speech[1,2], is concatenated to produce continuous speech waveforms.

This paper is organized as follows. Section II describes the system configuration; it presents the both linguistic process and synthetic process. Section III describes the implementation of hardware for real time synthesizer. Section IV gives the experimental results. And conclusion is given in Section V.

II.TEXT-TO-SPEECH CONVERSION

Fig. 1 shows the overview of the text-to-speech conversion process. In the linguistic process, input text from either keyboard or text file is converted to a phonemic representation using letter-to -sound rules and word dictionary, which is under developing status. The length of the pause are determined by using boundary information included in text.

In the synthetic process of speech, the duration of each syllable in the sentence is computed and then that of each phoneme is extracted from

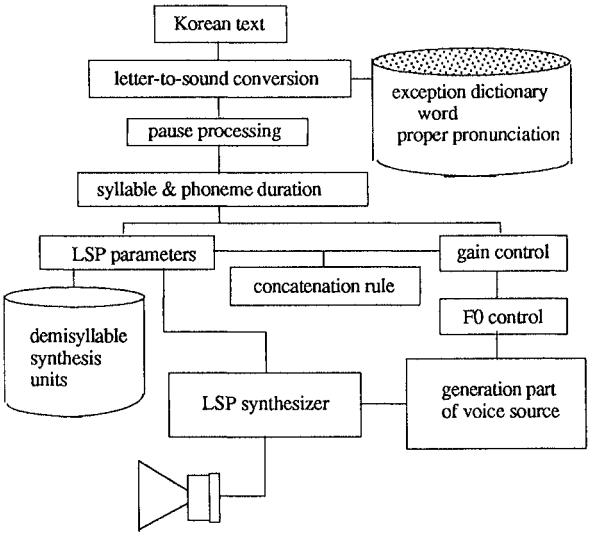


Fig. 1. Overview of text-to-speech conversion.

the syllable duration. The contour of the fundamental frequency is obtained by rules using the boundary information of phrase, clause, sentence and then adjusted according to the coarticulation method of the consonants in the syllables. The energy contour of the excitation source in each word is obtained and then merged into the clause unit according to the rules. The continuous synthetic speech is produced by concatenating the CV/VC unit.

II.1 Linguistic processing

The purpose of the linguistic processing is as follows:

(1) convert input text into a string of phonetic notations.
(2) insert the pause with proper length at the boundary of the sentence.

In this section, we describe the rule of Korean phonology.

II.1.1 Phonological rules of Korean phonetics

Basically Korean syllable consists of initial consonant, vowel and final consonant. However, final consonant doesn't always exist. The number of characters for initial consonant is 19, for vowel 21, and for final consonant 27. The number of the phonemes for initial consonant and vowel is the same as that of characters, but that for final consonant can be reduced to 7. As a result of perception test, phonemes for 3 vowels were redundant. So 44 phonemes are sufficient to represent Korean syllables and the phonemic representation for a syllable is unique. But when morphs are joined together, pronunciation on morphic boundary is often

changed depending on the nature of morphs and phonetic environment[5]. We call these changes as phonological phenomena and these are occured regularly and orderly, but there are some exceptions.

We have formalized these rules and applied them to syllable boundaries of the input text , which is written by Hangeul orthography.

II.2 Speech synthesis by rule

II.2.1 CV/VC synthesis

Korean is written and pronounced with syllable by syllable and each syllable has one of the following formats; 1) V, 2) CV, 3) VC, and 4) CVC, i.e., 아 /a/, 가 /ka/, 악 /ak/, 각 /kak/. Even though the number of all possible syllables is 3,520, i.e., 22 for V, 154 for VC, 418 for CV, 2926 for CVC, it can be practically reduced to 1096 because of the restrictions in the process of concatenation of each phonemes. In our study, we selected 640 demisyllables as synthetic units to obtain the good quality of speech while reducing the number of synthetic unit at the same time.

II.2.2 Concatenation rule

It is necessary to smooth out adjacent parameters, such as LSP parameters and excitation energy, in the process of the concatenation with demisyllable unit to improve the quality of synthetic speech. In order to smooth out the speech parameters more effectively, we classified the demisyllable with 9 types. We finally fixed to 4 kinds of smoothing mode depending on the combination of each type. The types of demisyllable and the smoothing rules are tabulated in Tables I and II, respectively.

Table I. Types of demisyllable.

Type	Format	Type	Format	Type	Format
1	pause	4	CuV+	7	+VCv
2	V+	5	CpV+	8	+VCu
3	CvV+	6	+V	9	+VCp

V+ : pre-demisyllable vowel, +V : post demisyllable vowel,
Cv : voiced consonant, Cu : unvoiced consonant, Cp : stop consonant.

Table II. Smoothing rule.

type / type p	1 2 3 4 5 6 7 8 9	type / type p	1 2 3 4 5 6 7 8 9
1	4 3 3 4 4 - - - -	6	2 1 1 2 2 - - - -
2	- 1 1 2 2 1 1 1 1	7	2 1 1 2 2 - - - -
3	- 1 1 2 2 1 1 1 1	8	4 3 3 4 4 - - - -
4	- 1 1 2 2 1 1 1 1	9	4 3 3 4 4 - - - -
5	- 1 1 2 2 1 1 1 1		

Also, Fig. 2 shows the smoothing model between adjacent parameters.

II.2.3 Rule of duration

We computed and used phonemic duration to produce the synthesizer control parameters even though Geul-Sori adopts demisyllable as a synthetic unit. This is because the syllable duration of Hangeul is highly dependent on the number of syllables in a word, boundaries of phrase, clause and sentense, and phonemes within a syllable[1,7,8]. And the duration of each phoneme is not directly proportional to that of syllable because each phoneme has its own minimum and inherent duration to preserve intelligibility and naturalness. The dependency of vowel duration on consonant is higher than that of consonant duration on vowel. Also phonemic environment has influences on phoneme duration.

So, first we extracted the duration information of each syllable and lengthened from the information of word, phrase, clause and sentence. The model is heuristically obtained based on the characteristics of Korean

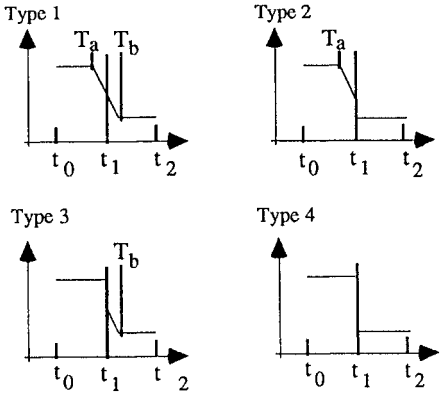


Fig. 2. Smoothing model.

language.

The WDdur, duration of a word, is given by

$$\mathrm{WDdur} = \mathrm{RFdur} \times [\alpha(N-1)+1], \quad (1)$$

where RFdur is a standard duration of a syllable, N is the number of syllables in a word , and α is a constant.

The SYLdur, duration of a syllable in a word, is

$$\mathrm{SYLdur}=\mathrm{WDdur}/N. \quad (2)$$

Next, the duration of each phoneme is obtained using segmental duration model of Klatt[1,2]. The PHONdur, duration of phoneme in a syllable, is given by[1]

$$\mathrm{PHONdur}=\mathrm{MINdur}+(\mathrm{INHdur}-\mathrm{MINdur}) \times \mathrm{PRCNT}/100. \quad (3)$$

II.2.4 Rule of fundamental frequency

In the standard speech of Seoul, there is found no pair of lexical items which are distinguished exclusively by a difference of pitch, even though some pitch pattern in a word is found according to the structure of syllables, stress and personal habit. But it is difficult to formalize pitch pattern in a word of Korean. In Seoul speech, therefore, the role of pitch is intonational, not tonal[7].

Korean has characteristics of declination like other languages and of "fall and rise" phenomena at both phrase and clause boundary. However, First F0 peak of Korean has almost no influences on length of the sentences, which is different from any other languages[9].

In our study, we find F0 contour of the sentence as follows:

(1) the contour baseline of the fundamental frequency for the sentence is determined by both semantics of the sentences and the boundary information of phrase, clause, and sentence.

(2) The coarticulation method of each phonemes makes the local peak and valley.

We adopted Fujisaki model [10,11,12] to produce baseline of F0.

II.2.5 Representation of vocal tract parameters

The linear predictive coding(LPC) is one of the most popular approaches for processing speech. For most of applications, it is necessary to interpolate the speech parameters in order to improve the speech quality. The LSP method, introduced by Itakura and Sugamura[13], is generally known as one of the efficient speech analysis/synthesis techniques because the LSP representations have small spectrum distortion for linear interpolation of parameters and by parameter quantization[13]. Therefore, we adopted the LSP representations as a means of coding speech parameters.

III. HARDWARE IMPLEMENTATION

We simulated the possibility for implementing the synthesizer in real time. Our simulation shows that the time required to extract the information on the text processing and prosody has almost no influences on the real time implementation because it is very short. But both extraction for a set of control parameters and the production of synthetic speech requires a lot of time. Therefore, in order to implement Geul-Sori

in real time, it is necessary to design and manufacture a special purpose synthesizer. We manufactured the real time synthesizer at a reasonable cost by considering the time required to produce the speech. The hardware is controlled by IBM PC/AT. The IBMPC/AT produces a set of control parameters with demisyllable unit and then transmit them to the real time synthesizer. By this scheme, it makes the system possible to operate with real time.

Fig. 3 shows the block diagram of the synthesizer. As shown in Fig. 3, we designed the hardware by using TMS320C25 DSP chip and 12 bit D/A converter etc. It's operation is as follows: First, after performing the text processing and prosody processing with the sentence unit at a time, IBM-PC/AT produces a set of synthesizer control parameters with the demisyllable unit by using CV/VC synthetic unit stored at hard disk, and then transmit them to the synthesizer. Second, the synthesizer produces the synthetic speech and output analog speech waveforms by D/A converter. For the interface between both PC and synthesizer, we adopted the memory mapped I/O method. On the other hand, we used the I/O port in order to handshake the protocol of the data transmission. By mentioned so far, we implemented the hardware for the synthesizer. The hardware developed is well operated.

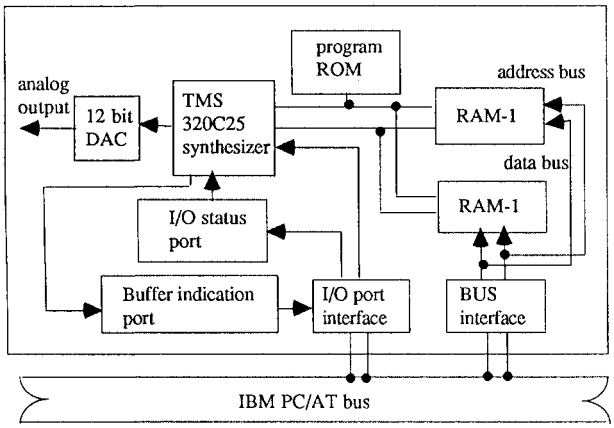
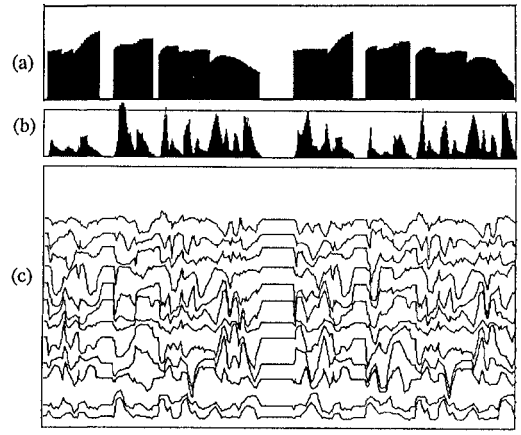


Fig. 3. Block diagram of the synthesizer.

IV. EXPERIMENTAL RESULTS

Fig. 4 shows an example of our experiments. The 12th order of the vocal track filter was selected to allow both good spectral matching and real time implementation. Fig. 4(a) shows the F0 contour generated by rule. The residual energy contour is constructed by both rule and information of database as shown in Fig. 4(b). Figs. 4(c) and 4(d) show the LSP parameters and the waveforms synthesized by rule. Fig. 4(e) shows the source waveforms, which are sampled with frequency 10 KHz.



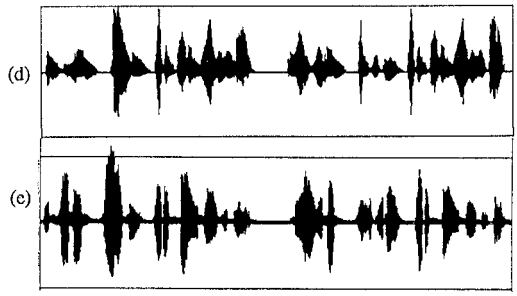


Fig. 4. Example of synthesis for the sentence:
Transcription; /ɯnjunɯn maɯmi arɯmdaun jʌinimjʌ, mijʌŋinɯn ʌlguri arɯmdaun jʌinida./
Meaning; Eun-Ju is a lady with the warm heart, and Mi-Young, with the pretty face.
(a) Pitch contour. (b) Energy contour.
(c) LSP parameters. (d) Synthesized speech.
(e) Source waveforms.

From Figure 4, we can see that the synthesized waveforms are similar to the source waveforms.

V. CONCLUSION

We have described Geul-Sori, a text-to-speech system for Korean language developed by ETRI since 1988. We developed and used the rules for linguistic processing such as letter-to-sound rule and pause processing method. We adopted and modified the duration model of Klatt. We adopted and modified the F0 contour rule of Fujisaki. As a concatenation unit, we used the demisyllable taking into account both the speech quality and the real time implementation. To implement the synthesizer in real time, we designed the special purpose hardware, which is operated on the IBM PC/AT. The experiment results described in Section IV shows that Geul-Sori is well operated, but yet poor in naturalness. So, we are now developing many rules to achieve more natural and more intelligible speech output from the synthesizer.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Prof. Ryunen Teranishi at Kyushu Institute of Design and Dr. Higuchi at KDD for their constructive advices at the earlier stage of this work.

REFERENCES

[1] J. Allen, M. S. Hunnicutt, and D. Klatt, *From text to speech : The MITalk system*, Cambridge University Press, 1987.
[2] D. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Am.* 82(3), pp737-793, Sept. 1987.
[3] K. Hakoda *et al.*, "Japanese Text-to-Speech Synthesizer based on Residual Excited Speech Synthesis," in *Proc. ICASSP86*, pp2431-2434, Apr. 1986.
[4] H. Javkin *et al.*, "A Multi-lingual Text-to-Speech System," in *Proc. ICASSP89*, 1989.
[5] U. Hoe, *Korean phonology*, Seoul : Sam publishing company , 1985, (in Korean).
[6] Y. S. Kim, *Study of Korean sound*, Seoul : Gwa-hag-sa, 1981,(in Korean).
[7] Hyun Bok Lee, "Korean prosody: Speech rhythm and intonation," *Korea Journal, Vol. 27, No. 2*, Feb. 1987.
[8] Minje Zhi *et al.*, "Acoustic Phonetic Studies for Speech Synthesis by Rule of Korean II: An Experiment on Speech Rhythm in Korean," *Proceeding of KICS summer conference, Vol. 9, No. 2*, pp613-616, Aug. 1990.

[9] Do Heung Ko, *Declarative intonation in Korean: An acoustical study of F0 declination*, Seoul : Hanshin publishing company , 1988.
[10] Hiroya Fujisaki, Keikichi Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Jpn.* (E) 5.4, pp233-242, 1984.
[11] Keikichi Hirose, *et al*., "Generation of prosodic symbols for rule-synthesis of connected speech of Japanese," in *Proc. ICASSP86*, pp2415-2418, April 1986.
[12] Hiroya Fujisaki,Hisashi Kawai, "Realization of linguistic information in the voice fundamental frequency contour of the spoken Japanese," in *Proc. ICASSP88*, pp663-666, April 1988.
[13] N. Sugamura and N. Farvardin, "Quantizer Design in LSP Speech Analysis-Synthesis," *IEEE J. Select. Areas Commun., Vol. 6 No.2*, pp432-440, Feb. 1988.