Implementation of Prosodic Information for an Unlimited-Vocabulary Korean Speech Synthesis System

Jun Heo and Jae Hong Lee

The rules of prosodic information are extracted from Korean natural speech. The implementation of the rules of prosodic information are proposed for an unlimited-vocabulary Korean speech synthesis system. It is shown that the quality of synthesized speech is improved by implementing the rules of prosodic information on the unlimited-vocabulary Korean speech synthesis system.

I. Introduction

Two criteria of performance for an unlimited-vocabulary speech synthesis system are naturalness and intelligibility (Papamichalis (1987)). The naturalness and the intelligibility of synthesized speech depend much on the presence or the absence of prosodic information in it. To improve naturalness and intelligibility of an unlimited-vocabulary Korean speech synthesis system, it is important to reflect prosodic information such as pitch, stress, and duration on synthesized speech. There have been researches in linguistics on prosodic information of the Korean spoken language which were focused on finding the characteristics of prosodic information. However, there have been only few researches on implementing prosodic information on Korean speech synthesis. The first step of implementing prosodic information on speech synthesis is to find the rules of prosodic information such as pitch, stress, and duration from Korean natural speech. Although there have been a few researches on finding the rules of prosodic information from Korean natural speech, most of them dealt with limited prosodic information.

Language Research, Volume 27, Number 2, June 1991. 0254-4474 / 211-217

In this paper the rules of prosodic information and their implementation are proposed on an unlimited-vocabulary Korean speech synthesis system. From Korean natural speech we extract prosodic information in various types of Korean sentences. We also examine how prosodic information varies in a clause in sentences depending on various factors. By examining prosodic information the rules of prosodic information are obtained which is to be implemented on an unlimited-vocabulary Korean speech synthesis system. In section II we show the types of prosodic information and the methods to implement prosodic information on an unlimited-vocabulary Korean speech synthesis system. In section III we show the rules of prosodic information which are obtained by examining prosodic information with Korean natural speech. In section IV we draw conclusions.

II. Prosodic Information and its Implementation Methods

Examining Korean natural speech, we find the prosodic information in three types of Korean sentences : a declarative, an interrogative, and an imperative sentences. Prosodic information also varies according to an inflection, a postposition, a punctuation mark, and a conjunction in clauses in a sentence. Prosodic information of Korean language consists of pitch, stress, and duration. It was claimed that each of pitch, stress, and duration is of different importance to characterizing prosodic information in Korean natural speech (Lee (1987), Heo (1987)).

Pitch is the tone of a voice. The pitch in a sentence varies depending on the type of a sentence which is typically characterized by concluding inflection and punctuation mark. The pitch in a sentence varies further depending on the type of a clause which is typically characterized by a connecting inflection between two clauses.

Stress is the degree of force with which a syllable is pronounced. The stress is independent of the type of a sentence or a clause. However, the stress is dependent on the structure of a word. The stress in a word depends on the existence of a 'jongseong' and the number of syllables in a word. There are at most two stresses in a word. When there are two stresses in a word each of stresses is called the first or the second stress depending on the strength of the stress.

Duration is the length of a voice. Duration is divided into two types : that of a syllable and that of a pause. The duration of a syllable is the length for which a syllable is pronounced. The duration of a pause is the length of a silence. The duration of a syllable varies depending on the number of syllables in a word and the existence of a stress in a syllable. The duration of a pause varies depending on its location.

Now we consider methods to implement prosodic information on an unlimited-vocabulary Korean speech synthesis system in which MPLPC (multi-pulse linear predictive coding) is used as speech coding (Cheon (1989), Ozawa et al. (1986)). The block diagram of the system is shown in Figure 1. The speech synthesis system consists of five parts : a phonetic variation part, a prosodic rules part, a speech synthesizer, a speech database, and a prosodic control part. The phonetic variation part applies phonetic variation rules to Korean input text. The prosodic rules part applies rules of prosodic information to Korean text which is processed by the phonetic variation part. The speech database consists of speech data files for Korean demisyllables. The speech synthesizer converts Korean text into a speech signal taking speech database files corresponding to demisyllables in Korean text. The prosodic control part applies the rules of prosodic information to the speech signal. The prosodic control part applies rules of duration, stress, and pitch separately.

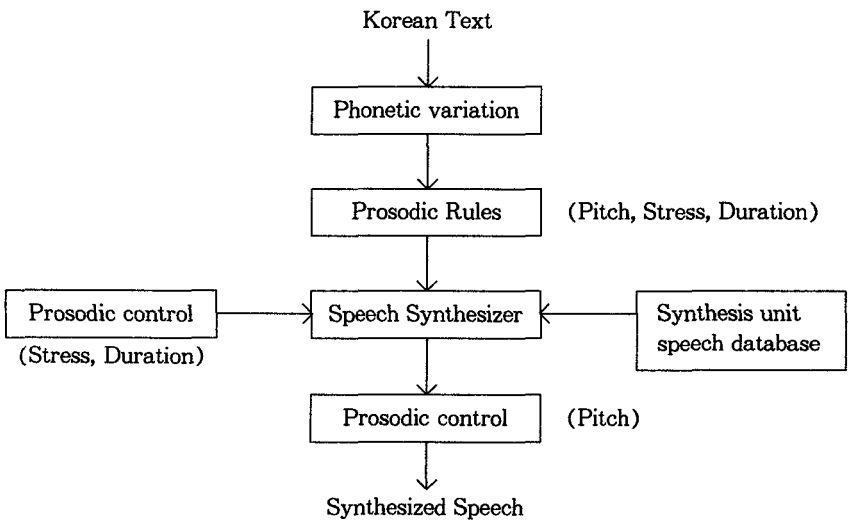

Fig. 1. Block Diagram of an Unlimited-Vocabulary Korean Speech Synthesis System.

Pitch is controlled, after the synthesizer outputs a speech signal. The pitch is reflected on a synthesized speech signal by adjusting the fundamental frequency of a synthesized speech signal in the time domain. To control the pitch in a synthesized speech signal the synthesized signal is first decomposed into a sequence of short-term signals which are obtained by multiplying the synthesized signal by a sequence of overlapping analysis windows. The analysis windows are centered at successive instants which is synchronous to the pitch of a synthesized signal during a voiced portion. Then the short-term signals are simply overlapped and added at different successive instants adjusted by the rules of prosodic information. This PSOLA (pitch synchronous overlap add) method results in a speech signal having an adjusted pitch (Hamon et al. (1989), Charpentier & Stella (1986)).

Stress is controlled inside the synthesizer. The stress is reflected on Korean speech synthesis system by adjusting the amplitude of a synthesized signal. A synthesized signal has a larger amplitude in a syllable with a stress than in that without a stress. Stresses are divided into two types depending on their amplitude in a synthesized signal : a first and a second stresses. The amplitude of a synthesized signal is multiplied by a factor depending on the presence and the type of a stress. Through a subjective quality test for synthesized speech the multiplying factors are obtained.

Duration is controlled inside the synthesizer. The duration is reflected on a Korean speech synthesis system by controlling the length of a voiced portion of a synthesized signal. Durations are divided into two types depending on their length of a voiced portion in a synthesized signal or length of a pause : duration of a syllable and duration of a pause. The duration of a syllable is controlled by repeating the period of a voiced portion of the syllable. The duration of a pause is controlled by shortening and lengthening zero signal depending on the position of the pause.

III. Experimental Results

By examining signals of Korean natural speech, we obtained prosodic information which consists of pitch, stress, and duration. Pitch is observed through a spectrograph. Stress is examined by measuring the amplitude of a speech signal. Duration is examined by measuring the length of a speech

signal. By analyzing these prosodic information the rules of prosodic information are extracted for pitch, stress, and duration as follows.

Pitch gradually declines in a clause. A clause is considered as a unit of pitch variation, because each clause has an independent pitch in a sentence having one or more clauses. The reference declination of pitch in a clause is set depending on its length, because the pitches at the beginning and the end of a clause are not influenced by the length of a Korean clause. The variation of pitch consists of the reference declination of pitch in a clause and pitch variation at an inflection and a postposition which overlaps the reference declination (Sagisaka (1990)).

The pitch of a concluding inflection in a declarative clause varies more in an honorific expression than in a nonhonorific expression. Pitch of a concluding inflection in an interrogative clause varies less in an honorific expression than in a nonhonorific expression. An interrogative clause is divided into two types : a Wh-interrogative and a yes/no interrogative clauses. The pitch of a concluding inflection in a Wh-interrogative clause is falling or flat, while the pitch of a concluding inflection in a yes/no interrogative clause is rising. The pitch of a concluding inflection in an imperative clause varies very similarly to that in a declarative clause, however the pitch varies less in the former than in the latter. The pitch of a connecting inflection in all types of a clause varies very similarly to that of a concluding inflection in a declarative. The pitch of a negative connecting inflection in all types of a clause is not falling but flat. The pitch of a postposition varies less than that of a concluding or a connecting inflection in all types of a clause.

Stress in a word shows strong dependence on the number of syllables and presence of a jongseong. A word with a single syllable has always the first stress. A word with multiple syllables is divided into two pieces so that the first and the second stresses. The first piece contains first two syllables and the second piece contains the rest. The former has a first stress and the latter has a second stress. In each piece the first syllable having a jongseong has stress. In each piece if no syllable has a jongseong, the last syllable of the piece has stress.

Duration of a syllable shows strong dependence on stress and the number of syllables in a word. The duration of a syllable in a single syllable word is longer than that in a multiple syllable word. The more the number of sylla-

bles in a word is, the shorter the duration of each syllable in the word is. In a word with more than two syllables if there is no syllable having a jongseong, the last syllable in the word is longer than any other syllables. In a concluding or connecting inflection the last syllable in the word is longer than any other syllables.

Duration of a pause depends on its position. When a pause is placed between two paragraphs, the average duration of the pause is 1500 msec. When a pause is placed between two sentences, the average duration is 700 msec. When a pause is placed between two clauses, the average duration is 450 msec. When a pause is placed between two words, the average duration is 250 msec. When a pause is placed in front of a syllable which has a fortis as a 'choseong', the average duration is 90 msec. When a pause is placed behind a syllable which has ㄱ, ㄷ, or ㅂ as a jongseong, the average duration is 75 msec.

The rules of prosodic information are implemented on an unlimited-vocabulary Korean speech synthesis system. From an unlimited-vocabulary Korean speech synthesis system the intelligibility and naturalness of synthesized speech are improved by applying the rules of prosodic information to the synthesized speech. The improvement is verified by comparing synthesized speeches with and without applying the rules of prosodic information.

Ⅳ. Conclusions

By examining signals of Korean natural speech, we have obtained prosodic information which consists of pitch, stress, and duration. By analyzing these prosodic information the rules of prosodic information are extracted for pitch, stress, and duration. The rules of prosodic information are implemented on an unlimited-vocabulary Korean speech synthesis system to improve the intelligibility and naturalness. We have verified the improvement of the quality in synthesized speech by comparing synthesized speeches with and without applying the rules of prosodic information.

References

Akers, G. & M. Lenning (1985) 'Intonation in Text-to-Speech Synthesis: Evaluation of Algorithms,' *J. Acust. Soc. Am.*, vol. 77, No. 6, pp. 2157-2165.

Charpentier, F. J. & M. G. Stella (1986) 'Diphone Synthesis Using an Overlap-added Technique for Speech Waveforms Concatenation,' *Proc. ICASSP*, pp. 2015-2018, Tokyo, Japan.

Cheon, Kang Sik (1989) *On Multi-pulse LPC Speech Coding for Unlimited-Vocabulary Speech Synthesis*, Seoul National University.

Hanmon, C., F. Moulines and F. Charpentier (1989) 'A Diphone Synthesis System Based on Time-domain Prosodic Modifications of Speech,' *Proc. ICASSP*, pp. 238-241, Glasgow, Scotland.

Heo, Woong (1987) *Korean Phonology*, Jeongumsa: Seoul, Korea.

Lee, Hyun Bok (1987) 'Korean Prosody: Speech Rhythm and Intonation,' *Korean Journal*, vol. 27, No. 2, pp. 43-70.

Middle School Korean Text (1989).

Ozawa, K., S. Ono and T. Araseki (1986) 'A Study on Pulse Search Algorithms for Multipulse Excited Speech Coder Realization,' *IEEE J. Selected Areas in Comm.*, vol. SAC-4, No. 1, pp. 133-141.

O'Shaughnessy (1987) *Speech Communication*, Addison-Wesley.

Papamichalis, Panos E. (1987) *Practical Approaches to Speech Coding*, Prentice-Hall.

Sagisaka Yoshinori (1990) 'Speech Synthesis from Text,' *IEEE Communn*, Mag., pp. 35-41.

Department of Electronics Engineering
Seoul National University
56-1 Shillim-dong Kwanak-ku
Seoul 151-742
Korea