KOREAN TEXT-TO-SPEECH SYSTEM USING TIME DOMAIN-PITCH SYNCHRONOUS OVERLAP AND ADD METHOD

Sang-Hun Kim, Jung-Chul Lee

Electronics and Telecommunication Research Institute
P.O.BOX 106 Yu-Seong Post Office, Taejeon, KOREA

ABSTRACT - We developed an advanced Korean text-to-speech conversion system using TD-PSOLA(Time Domain-Pitch Synchronous Overlap and Add) technique. Our system consists of language processing module, prosodic processing module, and synthetic speech generation module. This paper mainly describes the prosodic processing on text-to-speech system. To derive the segmental duration and intonation model, we selected appropriate sentences containing a variety of phrase structure. The prepared prosodic database, read by a female announcer, is composed of 38 sentences and 1021 syllables. The prosodic processing calculates segmental duration and *F0* contour from the rules extracted through the analysis of prosodic database. Finally, we applied the phrase level macro prosody using the syntactic and positional information of prosodic phrases in a sentence. The syllable level micro prosody was set up using the phonetic context and the position of syllable in a prosodic phrase. The advanced Korean text-to-speech conversion system applying prosodic processing shows more naturalness.

INTRODUCTION

Prosody conveys linguistic information, speaker's intention and emotion. Especially in speech synthesis, to generate natural synthetic speech, prosodic information must imitate a human speech as closely as possible. The prosodic information is composed of *F0* contour, segmental duration, and energy. Recently, many researches have been worked on prosodic processing for TTS system. For Japanese, the segmental duration rules were extracted by statistically analyzing speech data from four speakers(Kaiki *et al.*, 1992). In the statistic study based on speech data, it was found that many factors are involved in durational phenomena in Japanese, such as position in phrase, phrase group, the difference between content words and function words, and geminated consonants. However, prosodic processing has not been intensively studied for Korean yet. Therefore, we tried to analyze statistically the prosodic database, and then developed segmental duration and *F0* contour generation model for Korean. For the first attempt, the influence of neighboring segments and the prosodic boundary were studied based on the basic prosodic database. The results of statistic analysis were used to formulate the control rule for duration of segments in TTS system.

In this paper, we first overview the *GeulsoriII* system. Secondly, we show the experimental results of vowel duration affected on two factors: phonetic environments, and prosodic phrase boundaries. Also, we introduce the formula used in modeling duration and *F0* contour. Finally, simulation results are shown.

SYSTEM OVERVIEW

Our system consists of language processing module, prosodic processing module, and synthetic speech generation module(Kim, *et al.*, 1993). Language processing module performs text preprocessing, Korean functional word analysis, parser, prosody marker generation, and grapheme-to-phoneme conversion. Text preprocessing converts numerals, symbols into Korean. Korean functional word analysis decomposes Korean particles, suffix inflections, adverbs, and conjunctions, and then assigns analyzed word one of 48 attributes. Parser builds phrase level syntax structure using the attributes. Prosody marker generation uses syntax information for each phrase to specify one of 13 prosody markers that influence the spoken output. Exceptional word dictionary provides pronunciations for exceptional words, and 26 Korean phonological rules converts grapheme to phoneme. Prosody module calculates the duration of phonemes and the contour of fundamental frequency by rules.

The synthesis module selects the synthesis units(total 1226) from phoneme string, modifies and concatenates the synthesis units. Each unit contains time domain data, pitch, and segment information necessary for TD-PSOLA application. The synthesis units include the main variations of Korean phonemes. To decrease the concatenation defects, we use acoustic phonetic knowledge of Korean for spectral match. The prosodic parameters, such as duration, $F\emptyset$ and amplitude are directly scaled on the time-domain.

EFFECTS OF NEIGHBORING SEGMENTS AND PROSODIC PHRASE BOUNDARIES

Phonetic Environments

Each phonetic segment duration is affected by its neighboring phonemes. Furthermore, vowel duration has more variability than consonants. According to these facts, we investigated the variation of vowel duration with regard to its neighboring phonemes. We analyzed the combined effects of the consonantal classes preceding and following vowels.

$$Consonants(\text{or } Pause) + Vowels + Consonants(\text{or } Pause)$$

Consonants used in this experiment were classified into plosives, fricatives, affricates, nasals, and liquids. In case of vowels, we used 8 vowels: i,e,ae,a,eo,o,u,eu. We also considered pause effect. Table 1 shows the frequency of each vowel occurrence between the consonants preceding and following vowels in the prepared prosodic database.

	Plo.	Fri.	Aff.	Nas.	Liq.	Pau.
Plo.	39	32	14	51	45	69
Fri.	16	3	5	67	22	7
Aff.	4	4	10	48	5	6
Nas.	43	8	1	98	30	26
Liq.	7	4	3	35	9	18
Pau.	5	2	1	13	20	1

Table 1. Frequency of each vowel occurrence neighbored by consonantal class
(† row: consonantal class following vowels, column: consonantal class preceding vowels,
† Plo.: Plosives, Fri.: Fricatives, Aff.: Affricates, Nas.: Nasals, Liq.: Liquids, Pau.: Pause)

Since phoneme has its intrinsic duration, the effects of its intrinsic duration must be eliminated. Hence we found mean duration and MAD (mean absolute deviation$=\frac{1}{N}\sum_{i=1}^{N}|Dur_i - \bar{Dur}|$) of each vowel from the prepared database.

To inspect the tendency of durational variation, we computed the ratio of real duration of each vowel to the mean duration of that vowel, and then averaged these ratios. Fig.1 shows the tendency of durational variation of vowels.

The observed tendencies of durational variation on effects of consonantal class following vowels are as follows, where abbreviated words are consonantal class following vowels, and vowel duration is more shortened to the right direction.

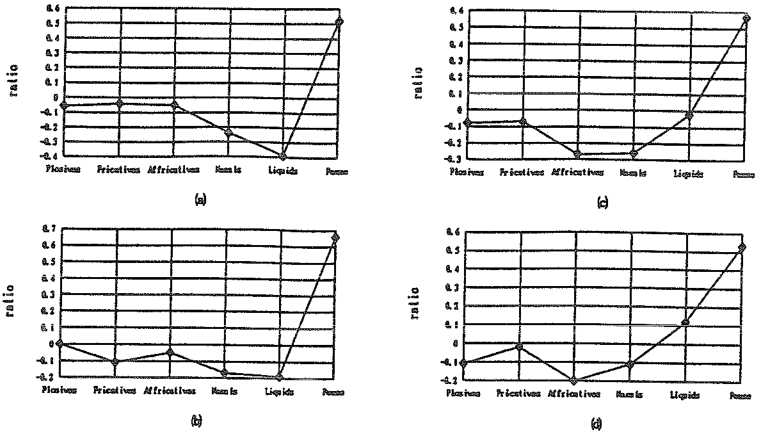


Figure 1. Tendencies of vowel durational variation affected by following consonants when preceding consonants is (a) Plosives (b) Fricatives (c) Affricates (d) Nasals

1. when consonantal class preceding vowels is plosives.
   Fri. > *Aff.* > *Plo.* > *Nas.* > *Liq.*
2. when consonantal class preceding vowels is fricatives.
   Liq. > *Fri.* > *Plo.* > *Nas.* > *Aff.*
3. when consonantal class preceding vowels is affricates.
   Liq. > *Fri.* > *Plo.* > *Nas.* > *Aff.*
4. when consonantal class preceding vowels is nasals.
   Plo. > *Aff.* > *Fri.* > *Nas.* > *Liq.*

In addition, we obtained the effects on pause, and plosives which was final consonants. In our database, each vowel duration preceding pause was lengthened by 30%-100% compared with mean vowel duration. When plosives was final consonant of syllable, each vowel duration was shortened by 10%-40%.

Prosodic Phrase Boundaries

Prosodic phrase boundary indicates a chunk of meaning during utterance, which enables a hearer to understand speaker's utterance more easily. Furthermore, it has much correlation with syntactic information of sentences, so it can be inferred from syntactic structure of sentences(Michelle *et al.*, 1992). So, we tried to investigate prosodic changes in prosodic phrase boundaries. To detect prosodic phrase boundary, we used three factors(Dur_{Ratio}, $F\emptyset_{Ratio}$, $and Pause$) in final syllable in each word, and then chose the candidate of the prosodic phrase boundary based on degree of prosodic changes.

- $Dur_{Ratio} = \frac{\text{duration of syllable}}{\text{mean duration of syllable}}$
- $F\emptyset_{Ratio} = \frac{\text{mean } F\emptyset \text{ of syllable nuclei}}{\text{mean } F\emptyset \text{ of word}}$

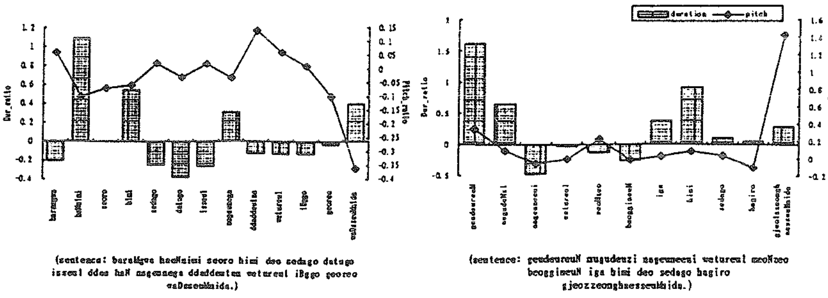


Figure 2. Dur_{Ratio}, $F\emptyset_{Ratio}$ in sentences

- Pause following words

We found three factors in prepared prosodic database which was composed of 38 sentences. Fig.2 shows two example sentences of the prosodic changes of each word final syllable. According to the experiment results, we found a few tendencies of prosodic changes at prosodic boundaries:

1. A long pause occurs.
2. Tendency of declination in $F\emptyset$ contour
3. Lengthening of the final syllable at prosodic phrase
4. Shortening of word final syllable in prosodic phrase
5. $F\emptyset$ rising and leveling between word boundaries in prosodic phrase

Generation Rule for Segmental Duration

Since the syllable duration is highly dependent on the number of syllables in a word, phrase, clause and sentence(Allen, *et al.*, 1987, Lee, 1987). However, the duration of a phoneme is not proportional to that of syllable because each phoneme has its own minimum and inherent duration. So we used phoneme as a unit to predict the duration though our system adopts synthesis units like $V, CV, VC, VCV, VCCV$. First, we calculated the duration of each syllable according to the number of syllables. The $SYLdur$, duration of syllable, is given by

$$SYLdur = RFdur \times \frac{[A \cdot (N-1)+1]}{N} \qquad (1)$$

where $RFdur$ is a standard duration of a syllable, N is the number of syllables in word, and A is a constant. And then, the duration of each phoneme is calculated using segmental duration model.

$$PHONdur = MEANdur \times (1 + \sum_i Prcnt(i)) \qquad (2)$$

where $MEANdur$ is the mean duration of each phoneme obtained from prosodic database, and $Prcnt(i)$ is the degree of durational variation due to the phonemic environment and the boundary of word, phrase, clause and sentence.

GENERATION RULE FOR $F\emptyset$ CONTOUR

Korean has a characteristics of $F\emptyset$ declination like other languages and of "fall and rise" phenomena at phrase and clause boundaries. In our study, $F\emptyset$ contour was generated as follows:

1. Global contour as sentence type:

$$F\emptyset_G(t) = A \cdot exp^{(\frac{-(t-B)^2}{C})}, \quad (3)$$

where A,B,C are constants given as sentence type.

2. Phrase level contour as phrase structure:

$$F\emptyset_P(t) = a \cdot exp^{(\frac{-(t-b)^2}{c})}, \quad (4)$$

where a,b,c are constants given as phrase structure.

3. Local $F\emptyset$ variation due to neighboring phonemes.

4. Sentence final, and phrase final $F\emptyset$ contour

RESULTS AND DISCUSSION

In this paper, we tried to analyze statistically the prosodic database, and then developed segmental duration and $F\emptyset$ contour generation model. For the first attempt, the influence of neighboring segments and the prosodic boundary were studied based on the basic prosodic database. The results of statistic analysis were used to formulate the control rule for duration of segments for TTS system.

The prosodic processing module calculates segmental duration and $F\emptyset$ contour using above mentioned rules. Fig.3 shows one example of the segmental duration and $F\emptyset$ contour generated by the prosodic rules. By perception test, TTS system applying prosodic rules shows more naturalness.

We are building up more prosodic database to obtain meaningful statistical model. In the future, we will consider more linguistic factors affecting the segmental duration and the $F\emptyset$ contour.

REFERENCES

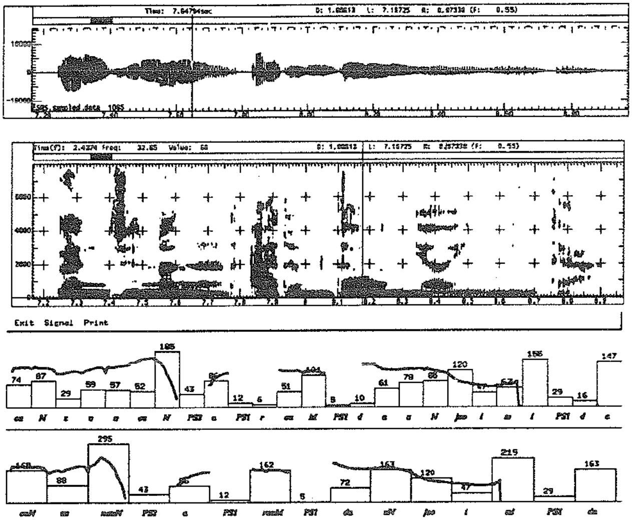
Hyun Bok Lee (1987) 'Korean prosody: Speech rhythm and intonation', *Korea Journal* Vol.27, No. 2.

J. Allen, M. S. Hunnicutt, and D. Klatt (1987) in *From text to speech : The MITalk system*, Cambridge University Press, USA.

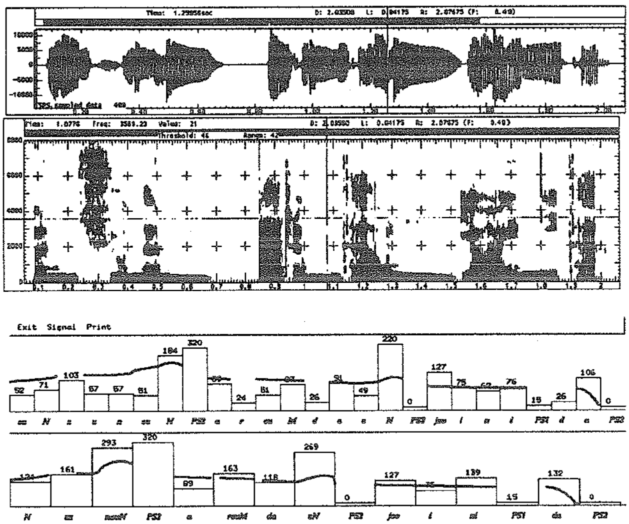
Sang-Hun Kim, Minje Zhi, and Un-Cheon Choi (1993) 'Application of TD-PSOLA Technique to Korean TTS Conversion', in *Internation Workshop on Speech Processing*, Tokyo, Japan, 83–88.

Nobuyshi Kaiki, Kazuya Takeda, and Yoshinori Sagisaka (1992), 'Linguistic properties in the control of segmental duration for speech synthesis', in *Talking Machines:Theories, Models, and Designs*, Elsevier Science Publishers,Netherlands, 255–263.

Michelle Q. Wang and Julia Hirschberg (1992) 'Automatic classification of intonational phrase boundaries', in *Computer Speech and Language*, Academic Press, 175–196.



(a)



(b)

Figure 3. Top: waveform, middle: spectrogram, bottom: segmental duration and $F0$ contour

(a) original speech (b) synthesized speech