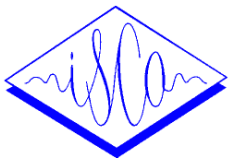

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

THE RULES IN A KOREAN TEXT-TO-SPEECH SYSTEM

Seung-Kwon ***AHN***

Advanced Technology Lab 4
GoldStar Central Research Laboratory, 16
Woomyeon-D Seocho-G, Seoul 137-140, Korea

Koeng-Mo ***SUNG***

Department of Electronics Engineering
Seoul National University , San 56-1,
Shillim-D Kwanak-G,Seoul 151-742, Korea

ABSTRACT

The rules for a Korean text-to-speech system based on the formant synthesis method are proposed. This paper describes all the aspect of the synthesis rules for Korean language. The rules are required for two parts. Firstly, for a linguistic process which converts the input text into phonetic unit. Secondly, for a synthesis process which generates and concatenates the synthesis unit to form a data package for spoken output message.

All these rules have been implemented on a personal computer with a digital signal processing board which is designed for real time processing.

I. INTRODUCTION

Text-to-speech technology for various languages has been developed to generate more natural voice with a smaller database.[1-3] Recent studies on this technology can be classified into two groups. The first group is based on an analysis-synthesis method using speech coding techniques such as LPC, LSP. this group has the merits of easy realization and impementation while it has the difficulty of the speed control, large database, and unnatural sound generation when two phonetic elements are concatenated. The second group is the formant synthesis method, in which every formant is extracted from the phonetic elements, and some rules are provided for the formant transition between two phonetic elements. Therefore, the second group does not have the problems of the first group, whereas it has the difficulty of making rules for formant transition.

We propose a method for the Korean text-to-speech system based on the formant synthesis method which uses a demisyllable element as the synthesis unit. Therefore, it is not necessary to make rules for formant transition by concatenating two phonemes. High quality voice can be synthesized using a relatively small database by the following two ways. Firstly, as a database, we have used only the boundary information of the linearized formant transition segments. Secondly, the phonemes which have a similar articulation position are grouped after carefully considering the phonetic characteristics of Korean. In order to enhance the naturalness of the synthesized speech, the stress, rhythm, and intonation of the Korean language, which are required for the prosody control, have been realized by making several rules for energy, pitch, duration, and pause insertion. Simplified Klatt synthesizer has been used after modification in order to fit the characteristics of Korean.

II. OVERALL STRUCTURE OF THE SYSTEM

Fig.1 shows the overall structure and flow of the proposed synthesis system. It consists of a linguistic process, a speech synthesis process, a Korean database, and a synthesizer. In the linguistic process, the special characters, symbols, and numerical digits in the input text are converted into the Korean characters. If there are long duration vowel in the input text, long duration marks are indicated. Noun dictionary is provided for this processing. Next, a stress mark is indicated and then the whole text is converted into phonetic symbols. An irregular conjugation dictionary is also provided for the exception processing in this step. In the speech synthesis process, prosody controls such as pitch, energy, and syllable duration control are carried out. Speech template information such as formant frequency, bandwidth, and segment duration in each phonetic element is read from the database, and then modified and concatenated according to the phonology of Korean. The output of the above process is converted into the synthesis parameters and forms data packages. Finally, the synthesizer can generate the speech according to the data packages.

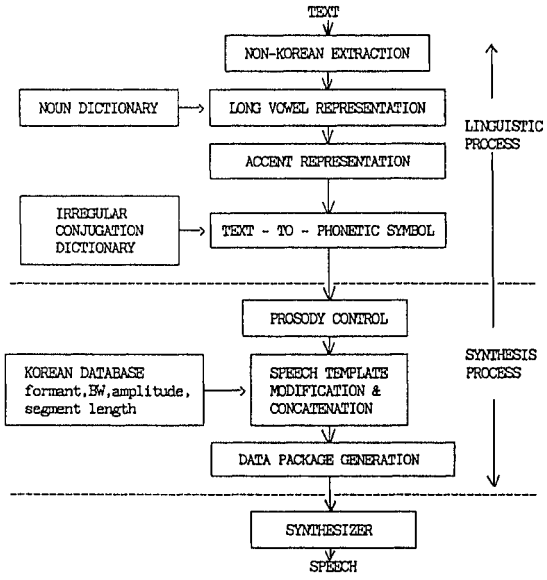

Figure 1: *Overall Structure of The System*

III. LINGUISTIC PROCESS

All the linguistic processes are carried out by the sentence unit. The procedures and rules for these processes are described as follow.

III.I Stress Representation

Any syllable having a stress has the characteristics of a louder sound, a higher pitch, a longer duration than normal syllables. Rules for the stress in a phrase are shown below.[4]

(a) A phrase with one syllable always has a stress.
(b) A syllable with a long duration vowel always has a stress.
(c) In case of a phrase with two or three syllables, the first syllable having a final consonant has a stress. but if there are no final consonant in any syllable, the last syllable has a stress.
(d) In case of a phrase having more than four syllables, it generally follows the same rules of the case (a) and the case (b) , and the first or the second syllable has a stress.
(e) In case of a phrase having more than three syllables, the second stress occurs, but never in succession. The above rules can be applied for this case.
(f) In case of a fortis and a aspirate , it follows above rules on the assumption that the preceding syllable has a final sound.

III.II Conversion to Phonetic Symbols

When phrases or syllables are constructed by concatenating any consonant and vowel, the phonological structure would be changed by mutual interactions. The rules for these are listed below, and details of each rule are described in the reference [5].

(a) Exclusion of '/h/', and the aspiratization
(b) The substitution, fortisization, and lenisization.
(c) Consonant variation with a juncture (pass 1)
(d) consonant variation with a juncture (pass 2)
(e) palatalization

IV. SPEECH SYNTHESIS PROCESS

Speech synthesis processes are carried out by the phrase unit. The procedures and rules for these processes are described as follow.

IV.I Prosody Control

IV.I.I Pitch

Pitch information is composed of pitch envelope and pitch level. This information has been obtained by the following step.

(step 1) pitch envelope

Pitch envelope information represents pitch variation in a syllable. One syllable is divided into 5 sections by the pitch variation. PL1, PL2, PL3, and PL4 are the indices of the sections and P1, P2, and DEF are the pitch values. Eq.1. shows a pitch envelope (EPITCH in the equation) of a section.

```
EPITCH = P1       ; syllable start < i < PL1
EPITCH = P1+(1 - DEF )*i/(PL2 -PL1) ; PL1<i<PL2
EPITCH = DEF      ; PL2 < i < PL3                (1)
EPITCH = DEF+(P2 -1)*i/(PL4-PL3)    ; PL3<i<PL4
EPITCH = P2       ;PL4 < i < syllable end,
     where i is sample index,and DEF is 9.5msec.
```

(step 2) addition of the pitch level

As the next step, the level information is added to control the total pitch level of each syllable. Consequently, pitch information could be represented as Eq.2.

```
PITCH = EPITCH + LEVEL.                          (2)
```

The pitch level tends to increase in one utterence unit, however, it is difficult to find out the utterence unit exactly. Therefore, in order to overcome the difficulty, we find out the endings within the 5 phrases or the subjective, objective, and concatenative endings using a dictionary. Then we could consider them as a utterence unit.

IV.I.II Energy

Energy determines the amplitude of the voice, and influnces the stress implementation. This information has been obtained by the following step.

(step 1) energy envelope

The energy envelope represents energy variation in one syllable. One syllable is divided into 5 sections by the energy variation. EL1, EL2, EL3, and EL4 are the indices of the sections and E1, E2 are the energy values. Eq.3 shows the energy envelope (EENERGY in the equation) of each section.

```
EENERGY = E1          ; syllable start < i < EL1
EENERGY = E1(1-E1)*i/(EL2-EL1); EL1< i< EL2
EENERGY = 1           ; EL2 < i < EL3            (3)
EENERGY = 1+(E2-1)*i/(EL4-EL3); EL3< i <EL4
EENERGY = E2          ; EL4 < i < syllable end,
            where i is sample index.
```

(step2) multiplication of the peak envelope

Abrupt energy change could occur in some consonants, especially for the plosive, but it can't be represented by Eq.3. Peak envelope information (EPENERGY in Eq.4) covers these changes of energy. In Eq.4, KL1 and KL2 represent section indices, and K means degree of the peak.

```
EPENERGY = EENERGY * K  ; KL1 < i < KL2          (4)
EPENERGY = EENERGY * 1  ; otherwise,
      where i is sample index.
```

(step 3) multiplication of the energy level

For the next step, the level information is multiplied to control the total energy level of each syllable. Consequently, the energy information could be represented as Eq.5.

```
ENERGY = EPENERGY * LEVEL.                       (5)
```

Because the energy level is affected by the stress, "LEVEL" values of the syllables having 1st,2nd, and no stresses have been taken as 1, 0.8, and 0.7, respectively.

IV.I.III Syllable Duration

Duration of a vowel has been controlled for the implementation of the syllable duration according to the speaking speed, number of syllable in a phrase, and stress. The duration of a vowel,defined L in the equations below, which is the sum of the transition region and stable region duration.

```
L = Ltr1+ Lst+Ltr2 ; in case single vowel
L = Ltr1+(Ltrd+Lst)+Ltr2 ;in case dipthong,
```

where Ltr1, Ltr2 ; transition region in a vowel when CV are connected, and Ltrd ; transition region in the dipthong.

Ltr1, Ltr2 are varied with the concatenating consonants, and Ltrd is constant. Lst is varied with the speaking speed or the following rules.

(a) Lst decreases proportionally to the number of

syllables. But, if there are too many syllables (over 3)in one phrase, approximate value has been taken.

```
if{Number of syllable in one phrase (NSP) < 4 }
then Lst = Lst - NSP * 1.5 ms
else Lst = Lst - 3 * 1.5ms
```

(b) Lst could be varied with the stress, approximate value also has been taken.

```
if ( 1st stress )
then Lst = Lst + 10 ms
else if ( 2nd stress ) Lst = Lst + 5 ms
```

(c) Lst of the last syllable could be longer than the others, approximate value has been taken.

```
if ( the last syllable )
then Lst = Lst + 40 ms
```

IV.II Phonological Process and Database

Phonological processes, which modifying and concatenating the speech template using the Korean database. The Speech template consists of the formant information and segment length. Therefore, formant slope should be calculated. Speech template read from database should be modified whenever the phonetic characteristic is changed by connecting phonetic elements.

IV.II.I Structure of The Korean Database

Database in this system means an elementary information sets to synthesize a desired speech signal. In this paper, we suggest a database composed of the demisyllable elements, but this kind of database also requires large memory capacity. We could reduce the size of the database endurably by grouping several phonemes which have same formant characteristics.[6]

Table 1 shows a grouping of the Korean phoneme by the formant characteristics. Phonemes in the same group have the same formant characteristics. The left group and the right group in Table 1 are symmetric in the transition characteristics. The formant characteristic information which covers all the combinations of the consonant and vowels in each group are stored in the database.

Table 1: Grouping of The Korean Phonemes By The Formant Characteristics (V ; vowel)

Group	First sound	Final sound
1	/g/,/k/,/k^h/, + V	V + /g/, /g/
2	/n/,/d/,/t/,/t^h/ + V	V + /n/,/d/
3	/l/ + V	V + /l/
4	/r/ + V	
5	/m/,/b/,/p/,/p^h/ +V	V + /m/,/b/
6	/s^h/,/s/ + V	
7	/J/,/c/,/c^h/ +V	
8	/h/ + V	

Each phoneme or segmental transition area is divided into several linear transition segments of the formant. An example is shown in Fig. 2. Following rules have been made to prepare a synthesis unit for syllable dividing.

(a) One syllable is divided into two parts.

- part 1 ; the initial sound + transition area + stable area
- part 2 ; stable area + transition area + the final sound

(b) Each of above part is divided into arbitrary number of segments.
(c) Formant in one segment has a linear transition characteristic.

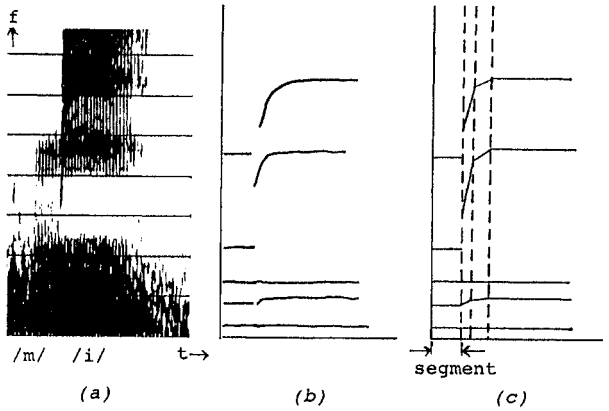

Figure2: An Example of The Formant Segmentation
(a) spectrogram of a demisyllable "/mi/",
(b) formant trajectory of (a),
(c) Formant Segmentation Model of (a)

IV.II.II Formant Concatenation

Formant information of each start point consists of formant frequency(F) and bandwidth(BW). When the formant is connected to the next segment, formant slope (FSLOPE & BWSLOPE) has been calculated by Eq.7.

$$FSLOPE(s)=\{F(s+1)-F(s)\}/LENGTH(s) \quad (7)$$
$$BWSLOPE(s)=\{BW(s+1)-BW(s)\}/LENGTH(s),$$

where s, LENGTH ;segment index , segment length.

After the slope had been calculated, formant information of each sample could be obtained by Eq.8.

$$F(j) = F(0) + FSLOPE * j / LENGTH \quad (8)$$
$$BW(j) = BW(0) + BWSLOPE * j / LENGTH,$$

where j is sample index.

IV.II.III Formant Modification

Formant could be varied with the concatenating phonetic elements. Some phonologies applied to cover these variations are shown below.

(a) Nasalization

Vowel with a preceding or a following nasal could be nasalized. These rules are following.

- Center frequency of the first formant (F1) has been increased.

Fn1 = F1 + 80 Hz (Fn1 ; nasalized F1)

- Nasal pole(Fnp & BWnp) and nasal zero(Fnz & BWnz) have been added.

Fnp = 250, BWnp = 100 ; nasal pole
Fnz = (F1 + Fnp)/2, BWnz = 100 ; nasal zero

(b) Vocalization

If there is a voiceless consonant between voiced vowels, it would be vocalized except for aspirate.[6] These rules are shown below.

- Voice source have been substituted by the periodic function.

- Energy have been decreased by 70 ~ 80 %.

IV.III Data Package Generation

Data packages have been generated from all the above-mentioned parameters in a predetermined order for the synthesizer. They consist of the information based on the syllable unit and the segment unit.[5] The former covers the information on the voice source, initial pitch, energy envelope , pitch envelope, energy level, and pitch level. The later covers the information on the voice/ voiceless, parallel/cascade synthesizer, abrupt change of energy envelope, formant, segment length.

V. SYNTHESIZER & IMPLEMENTATION

Klatt software synthesizer [7] which is modified to fit the characteristics of Korean has been used. Voiced sound and aspiration are synthesized by a cascade synthesizer and fricative and plosive are synthesized by a parallel synthesizer.

A LF - model[8] has been used as a voice source for a voiced sound, and a random signal with a Gaussian distribution has been used for a voiceless sound.

The hardware of the proposed system is composed of a personal computer (IBM PC's compatible) and an in-house board with a TMS320C25 digital signal processor. Linguistic process requires a huge memory for the dictionary, and can use a pause interval between sentences because this process is executed by the sentence unit. Therefore, the linguistic process is programmed by C language and executed on the personal computer. Speech synthesis process requires a real time processing. Thus, the speech synthesis process is programmed by assembly language of TMS320C25 and executed on the digital signal processing board.

In order to construct a database of Korean, all possible Korean syllable data pronunciated by a well trained male are recorded twice on an audiotape. The recorded data is sampled with sampling rate 10KHz and stored into the hard disk of the personal computer. Formant information is extracted from these data using a sona graph and an ILS(Interactive Laboratory System) package. Speech synthesis is carried out using all the above information, and frequency characteristics of these synthesized speech signals have been compared with the original ones, then these signals are D/A converted and generated as a voiced message.

As a proof of the quality of the synthesized speech, spectrograms of a natural and a synthesized speech are shown in Fig.3 for a phrase: /nanunganda/ which is "I go"in English.

VI. CONCLUSION

This paper describes the rules for a Korean text-to-speech system based on the formant synthesis method. It proposes a Korean demisyllable database and its reducing method. We have used only the boundary information of the linearized formant transition segments, and grouped the phonemes of the same characteristics to reduce the size of the database. By suggesting the rules for the stress, rhythm, and intonation, we could obtained a articulative and naturalized sound.

We are going to improve the present prosody rules in order to enhance a naturalness of the syntatic speech.

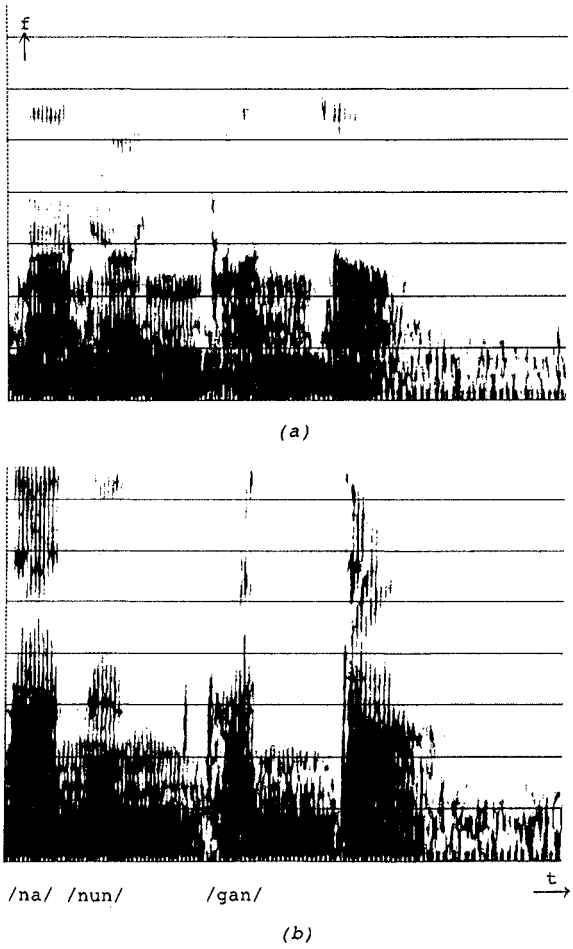

Figure 3: Spectrograms of a Natural and Synthesized Speech (a) Natural, (b) Synthesized

REFERENCES

[1] Special Issue on Man-Machine Communication by Speech, Proc. **IEEE**, Vo.73, 1985.

[2] *L.S.Lee, C.Y.Tseng, M.O.Young,* "The Synthsis Rules in a Chinese Text-to-Speech System, "**IEEE** Trans. Acoust., Speech, Signal Processing, ASSP-37, pp.1309 - 1319, 1989.

[3] *Tohru Shimizu, Seiichi Yamamoto, Norio Higuchi,* "The control of the prosodic features for the Japanese speech synthesis system by rule with editing functions," J. **Acoust. Soc. Jpn.**(J) Vo. 43,pp. 434 - 445, 1989 (in Japanese).

[4] *S.B.Lim, J.H.Lee, K.K.Choi* " A study on the Korean Synthesis by Rule, "Proc.Korean Inst.Tele. Elec. Fall Conf.Vo.10,pp.845-847,1987(in Korean).

[5] *Y.K.Lee, et.al.,* "On the Korean speech synthesis by rule using formant synthesis method:Database structure and prosody," Proc. **Acoust. Soc. Korea** Fall Conf. pp. 219-223,1989 (in Korean).

[6] *W.Hur,* The Korean Phonology (**Chung Yeum Sa** Press, Seoul, 1984),pp13 - 45 (in Korean).

[7] *Dennis H. Klatt,* "Review of text-to-speech conversion for English," J. **Acoust. Soc. Am.** Vo.82, pp. 737 - 793, 1987.

[8] *Rolf Carlson, et.al.,* "Voice source rules for Text-to-Speech Synthesis," Proc., **ICASSP**, pp.223 -226, 1989.