A Study on the Implementation of Korean Synthesis-By-Rule System Using Formant Synthesis Method

(포만트합성법을 이용한 한국어 규칙합성시스템의 구현에 관한 연구)

Cheol-Woo Jo,* Tae-Won Rhee**

조 철 우,* 이 태 원**

ABSTRACT

This paper suggests an example of implementing synthesis-by-rule system using formant synthesis method. At first we suggest an alphabetic phoneme description rules and then prepares phoneme parameter database from the natural speech materials. Then we propose the concatenation rules and synthesize speeches by concatenating phoneme units. A realtime speech synthesizer is constructed and used to perform synthesis-by-rule. Then synthesized speech is evaluated by listening tests.

요 약

본 연구에서는 포만트 합성법을 이용하여 규칙합성시스템을 구현한 일례를 제시한다. 먼저 음소의 입력을 위한 영문 알파벳과 음소의 대응관계를 설정한 뒤 수집된 자연음성으로부터 포만트 합성을 위한 특징 파라미터를 추출하여 데이터베이스를 작성한다. 그 다음 이러한 데이터베이스를 이용하여 제시된 음소간을 연결하는 규칙을 제안하고 음소단위의 합성을 행한다. 합성에는 신호처리 프로세서를 사용한 실시간 포만트 음성합성기를 구현하여 사용하였다. 합성결과 단독음소와 연결음소에 대하여 합성음성을 얻고 이를 평가하였다.

I. INTRODUCTION

Text-to-Speech system has been thought as an ultimate end of speech synthesis technology among researchers. And already some commercial systems can be found in the market for the languages such as English and Japanese. Of course their performances are different from product to product. But, as for Korean no commercial systems are available now and some laboratory level systems are announced through papers. There are many types of

* Dept. of Control and Inst. Eng., Chang Won Univ.
** Dept. of Elec. and Computer Eng. Korea Univ.

synthesis units or synthesis methods. We selected formant synthesis method for our text-to-speech system. This method is the most difficult one to implement, but gives the most flexibility if implemented well. Also text-to-speech technology requires the knowledges of various fields. And a large amount of speech materials should be processed to extract the characteristic parameters of phonemes for a language. As for Korean, the analytic experimental results about speech sound is yet not enough. Also we need the rule description method to express the rules from the language. Besides many problems remain intact to be studied.

Until now some experimental Korean text-to-speech systems are announced, but they used phoneme concatenation method utilizing LPC coded or MPLPC(Multipulse LPC) coded units. And their phonemic synthesis units were syllables or demisyllables not phonemes[1][2][3].

In this paper, we suggest a case of implementing formant based synthesis-by-rule system giving main focuses to the generation of phoneme database and formant track generation rules.

Ⅱ. Phoneme transcription Rules

To input phoneme texts to the synthesizer, we used English alphabetical characters similar to the phoneme alphabets from the IPA. Of course we developed a transcription software to convert normal Korean texts into phonemes[4]. But in this experiments we use alphabetical symbols to verify the exact matched speech[5]. That's because the transcription percentage of the software is not verified yet and the complete word dictionaries that do not fit to the transcription rules are not prepared yet, so to reduce the false results from the transcription errors we keyed-in phoneme symbols directly from the terminal.

Ⅲ. Analysis of Speech Materials and Structure of Database

To prepare the speech parameter database, analysis for the speech materials are conducted. To be used as an speech synthesis purpose, a thorough analysis for a one speaker is enough. So we collected speech material from one male speaker. And the material consists of 200 CV, VC monosyllables consisting of 18 initial consonants, 8 vowels and 7 final sounds. And 66 word materials are used to examine the variation of synthesis parameters during the boundaries of phonemes.

Analysis of speech materials for parameter database is conducted by the signal processing and statistical methods and partly by manual works. At first, linear predictive singal analysis method is applied for all the speech materials and then formant informations are extracted from the results by the formant tracking using parabolic interpolation method.

Ⅳ. Design of Synthesis Rules and Data Base[6]

In the design of rules and database, we limited the range to formant trajectory information. From the previous research reports, it is known that pitch and amplitude contour is less dependent on the formant informations[7]. And formant informations are unique to each phoneme. So we used only formant informations to make database. Pitch and Amplitude rules will be studied in the later papers.

To prepare phoneme database, we first analyzed the formant trajectory informations for each phonemes from the methods stated in the previous section. But the problem is about which information do we store to the database. For example, we can store whole the series of formant trajectories into the database, but it requires too much storage.

So we simplified the structure of the phoneme database by segmenting a phoneme into several parts. Segmentation can be done in two aspects. At first we assumed that all consonants consist of steady state part and transient part. Although this can be tought too simplified, we verified the propriety by the analysis of real speech data track. Then from the observations from the analyzed resuslts we categorized the form of formant trajectory that there are several kind of concatenating patterns between a phoneme and the other phonemes. These concatenating patterns are classified rather from the characterisitcs that a phoneme is voiced or unvoiced and plosive or non-plosive than from the mutual relations of phonemes. And they are different from the position of phonemes, namely whether a phoneme is located in front of or back of other phoneme. And also different formant parameters are assigned according to the following vowels because the contents of formant informations can be varied according to the kind of vowels that follows or precedes it. In this paper, we classified formant values according to the following vowels into 3 or 4 kinds to reduce database size. In case of Klattalk and MITalk they used some linear equations relating consonants and vowels from many measured data.[8]

To make rules about database and formant track shapes, we assumed the followings based on the analysis of real speech and references about general phonetics.[9]

—There must be transient parts at the starting and ending part of the vowels. From the analysis of real data we observed that even single vowel have transient part at the starting and ending part.

—Silent part can be inserted when vowel part is apart from the front phoneme.

—Normally formant track consists of "transient part+steady state part".

—The starting point of the transient parts of formant track is dependent on the previous phoneme kind.

—There are transient parts between voiced vowels.

—There are no transient parts between voiced phoneme and unvoiced phoneme.

—There must be transient part between vowels.

—Silence part can be inserted between consonant and preceeding phoneme according to the consonant kind.

Based on these assumptions each phoneme requires 2 to 4 kinds of data as follows:

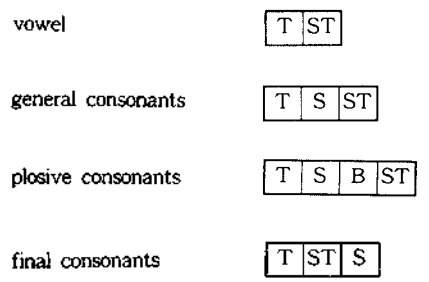


Fig. 1. Classification of parameter tracks by the kind of phonemes

Following the previously mentioned databases, we classified the kinds of formant track into 6 kinds as in table 1. Then the conditions to assign a specific track is as follows.

In general consonants : c1 c : c1=sonorant, c=nonsonorant track1
c1=sonorant, c=sonorant track2
c1=nonsonorant, c=nonsonorant track3
vc : track2
#c : track 3

plosive consonants : c1 c : c1=sonorant track4
c1=nonsonorant track5

final consonants : track6
(c : current phoneme, # c : initial consonant)

Table 1. Kinds of parameter track

kind number	structure
1	T S ST
2	T ST
3	S ST
4	T S B ST
5	S B ST
6	T ST S

The effect of phonetic environments are also considered. Generally the characteristic of a phoneme varies according to the its environments. And also that is required in synthesizing speech. Variation rules are determined using discriminative features of each phoneme. In this paper, we used 9 rules concentrating to the duration and amplitude variation. The variation rules are based on the reference[10] and we adopted proper rules for Korean from the general variation characterisitics from the reference. Rules mostly reflects the variation at the initial or final phonemes than at the complex environments. We used trial and error method to decide the variation factors of each parameter because the rules stems from the conceptual one without through experiments. The assumptions adopted are as follows:

- Nonsonorant plosives are aspirated at the word start.
- Duration increases when a consonant are at the word end.
- Duration increases when a vowel is located in front of the nonosonorant consonant.
- Duration increases when a nonosonorant sound is at the word end.
- Duration increases when a sonorant sound is at the word end.
- A vowel takes nonsonorant characterisitic in front of the nasals.
- Nonsonorant sound is vocalized in between two sonorant sounds.
- The energy of nasals is weakened at the word start.

The seaquence that rules are applied is as follows. At first the shape of the track is determined, then variation rules are applied. Only one rule for a condition can be applied.

To describe rules, we proposed rule description methods and Turbo-PASCAL is used for coding[6]. According to the proposed description rules, for example if parameter Ah(aspiration amplitude) of current phoneme increases by +5 when next phoneme belongs to the unvoiced and stop consonants, the rule can be specified by PASCAL as follows:

"if CurrentPhoneme in unvoiced and Current Phoneme in stops then Modify(Ah, +5)".

Futher coding examples can be found in ref.[6]

V. Real-time Speech Synthesizer[8]

We constructed a real-time formant speech synthesizer to synthesize speech. Synthesizer consists of two microprocessors(Z-80, TMS32010 Digital Signal Processor) and some accompanying ICs and it is designed for its operation to be controlled by IBM-PC through the connections by

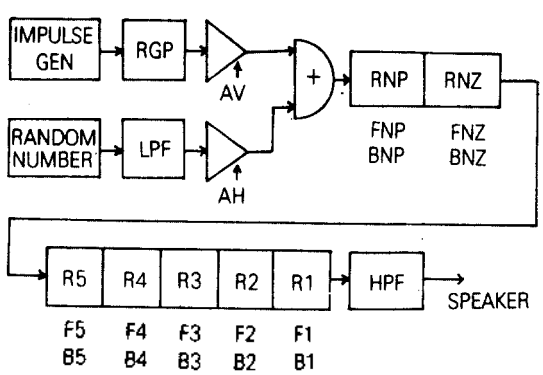


Fig. 2. Structure of the Fomant Synthesizer

its slot interface. Fig. 2 and 3. -- shows the structure of the formant synthesizer, the schematic diagram of the formant synthesizer hardware module respectively. And Table 2. shows the performance listings of the synthesizer.

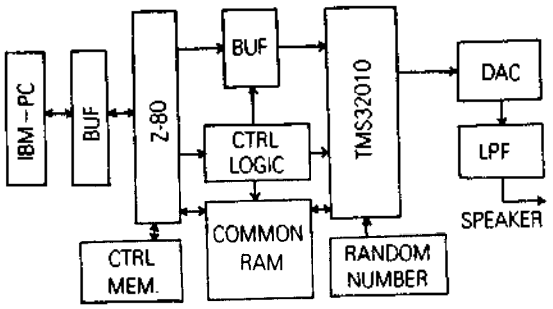


Fig. 3. The Schematic diagram of Real-Time Formant Synthesizer Module

Table. 2. Required time to synthesize a speech sample

module name	time(micro-second)
excitation	29.2
vocal tract filter	48.4
data store	7.2
data out / receive	9.4
other control	2.0
total	96.2

Ⅵ. Experimental Results and Discussion

We synthesized randomely selected 21 continuously spoken 4 digit numbers and 47 words or short sentences consisting of 1 to 7 syllables(2 to 16 phonemes). Then we tested their intelligibility. The 21 continuously spoken numbers are selected in random order to contain 10 numbers in similar probability. And we selected 18 male and 2 female subjects aged from 21-24 for the recognition tests. In the case of continuouse numbers, it is pre-notified to the subjects that 4 continous numbers will be spoken. But in the test for the words nothing is told to the subjects about the contents of the words. The words contains numbers, words and short greetings. All the synthetic speech is presented to the subjects with portable tape recorder in the compomparatively silent laboratory environment. Test resultls are shown in the table 3.

Table 3. Recognition Test Results in Random Suggestion Case

Contents	Recognition Rates(%)
4 continuous numbers (total 4 number)	13.8%
4 continuous numbers (each number)	52.4%
multi phoneme word or sentences	21.8%
high recognition rate words	
안녕히 가십시오 ("Goodbye" in Korean)	95%
안녕하십니까 ("How do you do" in Korean)	75%
어서오십시오 ("Welcome" in Korean)	70%
서울 (Seoul" : Korean Capital)	70%

As shown in table 3 recognition rate is generally low, but some showed high recognition rate. Generally muti-syllable words have higher recogniti on rate than 1 or 2 syllable words. In this paper no further systematic analysis of the synthesis system is not conducted because of the lack of the developed test methodology for Korean.

From the analysis for the short recognition tests, we can estimate and evaluate the performance of our synthesis system in spite of much lackness. Vowel sounds are recognized well in the words or numbers. But in consonants some have problems from case to case. This means that there are margins to develop new algorithms in track generation for the consonants and to analiz e real consonant speech more thoroughly. Errors mainly occurred in voiced consonants /l/, /r/, /m/, /n/ and in the conjunction of semivowel /j/s, /w/s and vowels and also in final /t/s

and /p/s. There are also needs to collect more statistical data about speech materials. And from the fact that long words have comparatively great recognitioon rate we can expect more recognition rate if this system is used to generate sentence sounds except that some consonants still have false recognitions.

Ⅶ. Concluding Remarks

We implemented a Korean synthesis-by-rule system using formant synthesis method. As a result we synthesized 21 continuously spoken numbers and 47 words and short sentences. The resulting synthetic speech is verified through recognition tests. The recognition result is not so satisfiable, but some commonly used words showed high recognition rates in spite of no prior comments about the contents of the suggested words before recognition tests. Also the recognition test result shows that the recognition rates can be greatly reduced if the domain of the words are controlled properly. To reduce error rate further analysis of natural speech and improvement of synthesis rule structure is going on.

References

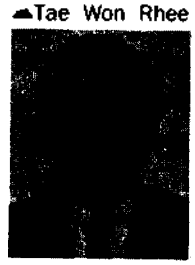
1. B.S. Kim, G.S. Yun, S.H.Park, "The Korean Text-to-Speech Using Syllable Units", Journal of the KITE, Vol.27, No.1, pp.143-150, 1990.
2. S.J.Lee, J.G.Ha, K.S.Chun, Y.I.Kim, Y.H.Choi, J.H.Lee, K.M.Sung, "Unlimited Vocabulary Korean Speech Synthesis Using MPLPC Coded Demisyllables", Journal of KITE, Vol.27, No.9, pp.1-10, 1990.
3. G.S.Yun, S.H.Park, "A Study on the Korean Text-to-Speech Using Demisyllable Units", Journal of KITE, Vol.27, No.10, pp.138-145, 1990.
4. C.W.Jo et. al.,· "A Proposal to the Phonetic Text Generator", Summer Conference Proceedings of KITE (1), Vol.9, No.1, June, 1986.
5. C.W.Jo, T.W.Lee, W.S.Lee, G.H.Ree, J.A.Kim, G.I.Lim, T.W.Rhee, "A Study on the Phoneme Based Analysis of Korean Initial Plosives Using Statistical Method and Perception Tests", Journal of KITE, Vol.8, No. 5, pp.78-85, 1989.
6. C.W.Jo, "A Study on the Implementation of Concatenation Rules of Phonemes for the Synthesis-by-Rule of Korean Using Formant Synthesis", Trans. on Institute of Industry and Technology, CNU, Vol.4, pp.91-97, 1990.
7. Y.J.Lee, "A Study on thye Recognition of Korean Monosyllables and the Variations of Its Characteristics on the Sex and Age", Master Thesis, Korea University, 1987.
8. J. Allen, M.S. Hunnicutt, D.H. Klatt, "From text to speech: The Mitalk System", Cambridge University Press, 1987.
9. P.Ladegoged, "A Course in Phonetics", 2nd ed., Harcourt, N.Y., 1975
10. C.W.Jo, "On the Implementation of the Experimental Speech Sysnthesis System Using Real-time Formant Synthesizer", Trans. on Institute of Industry and Technology, CNU, Vol.3, pp.83-87, 1989.

▲Cheol Woo Jo was born in Pusan, Korea on

September 12, 1961. He received B.S., M.S., Ph.D degree in 1983, 1985, 1989 respectively from Korea University both in electronics engineering. From 1986 to 1987 he worked as a researcher in the area of speech signal processing at the signal processing section, ETRI (Electronics and Telecommunications Research Institute), Daejon, Korea. Since 1989 he has been with the department of control and instrumentation engineering at Changwon National University, Changwon, Korea.

His current research interests include synthesis-by-rule of speech, man-machine interface, applications of speech in biomedical engineering and other application areas of signal processing.

▲Tae Won Rhee was born in Korea on July 27,



1931. He received B.S. degree in communication engineering from Seoul National University in 1958 and M.S. and Ph.D degree in electronics engineering in 1960 and 1975 respectively from the same university. From 1963 to 1971 he was with Kwangwoon University, Seoul, Korea. From 1971 to 1976 he was a faculty member of Jungang University, Seoul,, Korea. Since 1977 he has been a professor in the department of electronics and computer engineering at Korea University, Seoul, Korea. Also he was a visiting professor at Cornell University from 1981 to 1982. He was a president of Korean Institute of Telematics and Electronics and is currently a head of the university computer center at Korea University.

His current research interests are speech and image signal processing and their applications.