EVALUATION OF THE SYNTHETIC SPEECH QUALITY BY THE TD-PCULI METHOD

Chan Hee Kang*, Yong Jo Shin*, Yun Seok Kim*,
Ki Hyung Kwon**, and Yong Ohk Chin**

*, Dept. of Electronics in Sangji junior college
**, Dept of Electronic Engineering in Kyunghee Univ.

<ABSTARCT>

In this paper we have evaluated the synthetic speech quality by the proposed TD-PCULI speech synthesis method[1]. For the synthesis we have extracted parameters from the Korean monosyllables through the analysis of speech waveforms in the time domain. We have constructed the Korean data format dictionary for the synthesis-by-rule depending upon the frequencies of the Korean pronunciation large vocabulary dictionary[2], in which V type syllables are 19, CV type's are 80, VC type's are 30 and CVC type's are 100. And using them we have synthesized various Korean monosyllables, words and sentences. We have tested each 10 syllables selected according to the 4 Korean syllable types with the objective MOS(Mean Opinion Score) evluation method[3)-5)] about the 4 items i.e., intelligibility, clearness, loudness, and naturality after selecting random group without the knowledge of them. And also we have tested the possibility to modify a duration and F_0 into another forms with changing a duration(i.e., 150msec, 300msec, 500msec, 700msec and 1sec) and a central fundamental frequency(i.e., 80Hz, 118Hz, 140Hz, 170Hz, and 200Hz). As the results of experiments the noises occurred in the course of synthesizing the speech by the rules are removed to be a very clear level and we can find that the prosodic elements can be controled as a good condition.

I . Introduction

The difficulties of the prosody control in the time domain are that if a random function is weighted on the original speech or it is modified into another form the pseudo-periodic characteristics are lost and synthetic speeches are distorted heavily. In general the synthetic speech quality by the synthesis method[6)-8)] in the frequency-domain such as LSP or FORMANT synthesizer is inferior to the time-domain method. The reason is that in the frequecy-domain methods due to the estimated source and the estimated vocal tract functions it have been occured a estimation error to be deteriorated in the synthetic speeches. To overcome the problems we had proposed a new speech synthesis method called a TD-PCULI[1),9)]. In this paper to test the validity of the TD-PCULI method we have suggested the results which had been evaluated the synthetic speech quality with subjective Mean Opinion Score(MOS) method[11)-12)]. In the experimentation we have tested the Korean monosyllables and multisyllabic words with the changing of the prosodic factors.

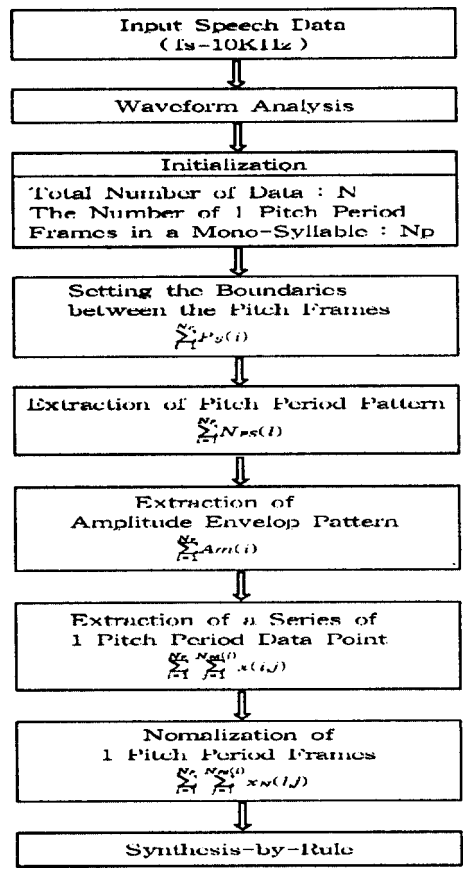

Fig. 1 The block diagram of extracting the parameters for synthesis-by-rule

Ⅱ. The Representation of Speech Signal in the TD-PCULI method[9),12)]

If we have defined a random data point of mono-syllable speech data by x(n) to extract the parameters for the sythesis-by-rule from the speech waveforms a series of speech data string will be defined by $\sum_{n=1}^{N} x(n)$. Here, the total number of speech data point is defined by N. If the total numbers of 1 pitch period frame in a mono-syllble are defined by N_p, the boundaries between the pitch frames are defined by P_{s1}, P_{s2}, P_{s3}, ····, P_{sNp}, the numbers of 1 pitch frame in a mono-syllable are defined by N_{Ps1}, N_{Ps2}, N_{Ps3}, ····,N_{PsNp}, and N_{Ps1}, N_{Ps2}, N_{Ps3}, ····,N_{PsNp}, are defined by 1-D array N_{Ps}( ), 1-Dimensional N-point speech data string $\sum_{n=1}^{N} x(n)$ will be represented as the summation of N_p small blocks which are consisted of 1 pitch frames. Then 1-Dimensional N-point speech data string $\sum_{n=1}^{N} x(n)$ will be as follows.

$$\sum_{i=1}^{N} x(n) = \sum_{n1=1}^{Np} \sum_{n2=1}^{Nps(i)} x(n1,n2) \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (1)$$

If we have defined each maximum values in 1 pitch frames by A_{m1}, A_{m2}, A_{m3}, ···· and the normalized speech data by $x_N(n)$, 2-Dimensional Array $\sum_{n1=1}^{Np} \sum_{n2=1}^{Nps(i)} x(n1,n2)$ will be as follows.

$$\sum_{n1=1}^{Np} \sum_{n2=1}^{Nps(i)} x(n1,n2) = \sum_{n1=1}^{Np} \sum_{n2=1}^{Nps(n1)} A_m(n1) \cdot x_N(n1,n2) \quad \cdots\cdots\cdots\cdots \quad (2)$$

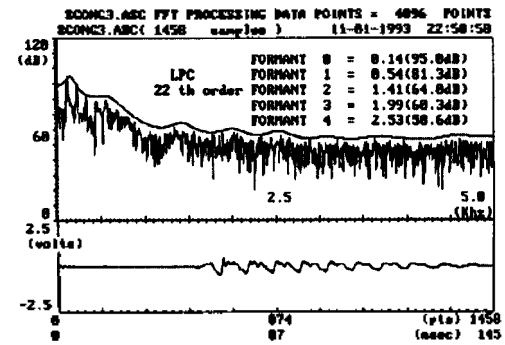

Fig. 2(a) The synthetic short-speech spectrum "공/gong/"

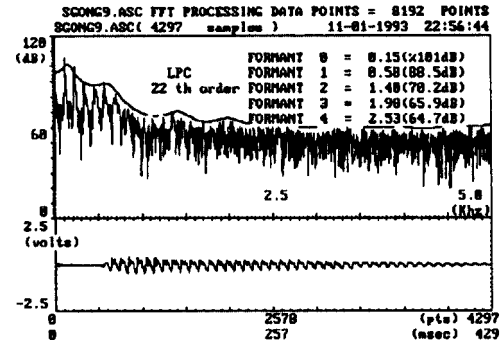

Fig. 2(b) The synthetic short-speech spectrum "공/gong/"

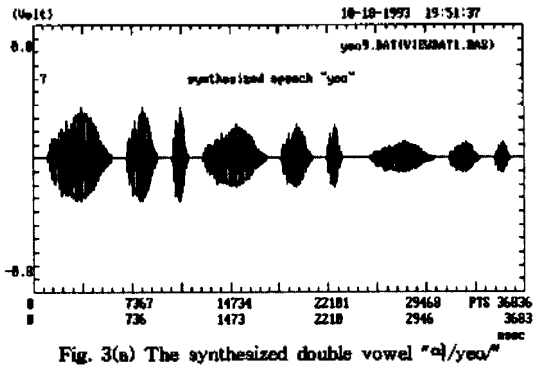

Fig. 3(a) The synthesized double vowel "여/yeo/"

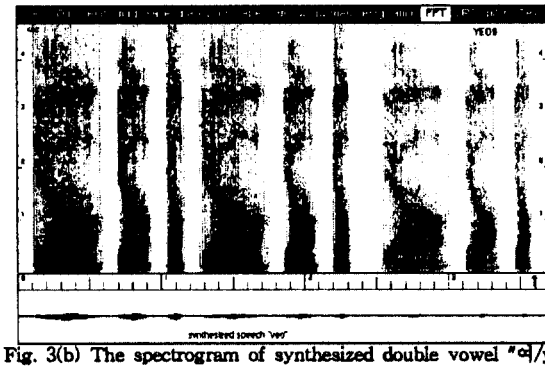

Fig. 3(b) The spectrogram of synthesized double vowel "여/yeo/"

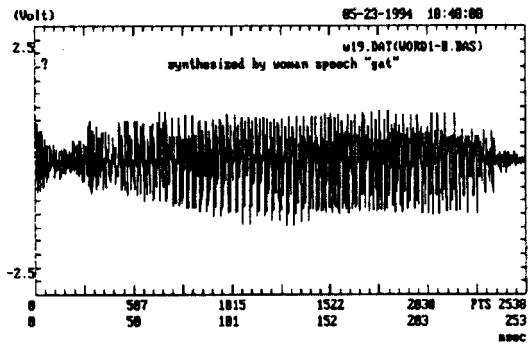

Fig. 4 Example of synthesized woman speech "갓/gat/"

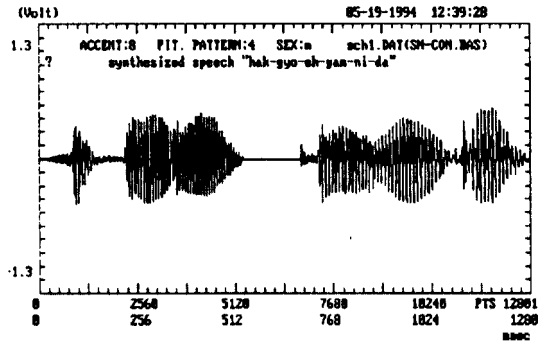

Fig. 5 Example of synthesized sentence "학교에 갑니다"

(Here, to be used The Korean Representation in Roman Characters)

Ⅲ. The Examples of Apllication to the Korean Synthesis-by-Rule

A below part in fig. 2(a) is a waveform plot of short synthetic speech "공/gong/" with a 0.146msec duration and a above part represents its spectrum plot. Fig 2(b) represents a long synthetic speech "공/gong/" with a 0.430msec duration. If we compare the spectrum and the formants given in fig. 2(a) and fig. 2(b) we can find that both are nearly to be same and there are not the noise componants to be deteriorated the quality due to the discontinuity in connection parts. Fig. 3(a) represents the synthetic speech of double vowel "여/yeo/" to be controled the duration and the stress and fig. 3(b) represents the spetrogram of fig. 3(a). Fig. 4 represents an example of the possibility to the synthesis-by-rule of a woman speech which is difficult to extract the parameters for the prosody control in the time-domain. The results of an experiment to another woman were almost the same to a man's. An example of sentence is given in fig. 5 which represents synthetic speech to be controled the prosody elements such as duration, stress, intonation, energy level and pause.

Table 1. The Experimentation Method

		음질평가실험		실험대상자	
		실험조	규칙합성음	A-그룹	B-그룹
단음절어	지속시간변화	S-1	150msec, 118Hz	김기홍	김은희
		S-2	300msec, 118Hz	신재연	이옥요
		S-3	500msec, 118Hz	안승현	윤명균
		S-4	700msec, 118Hz	안홍근	정종혁
		S-5	1sec, 118Hz	박현숙	홍의철
	피치변화	P-1	80Hz, 300msec	김호갑	류은하
		P-2	140Hz, 300msec	최윤정	한상윤
		P-3	170Hz, 300msec	박철	최명윤
		P-4	200Hz, 300msec	예윤경	신주철
2음절어	피치변화	D-1	80Hz	정민	김태순
		D-2	118Hz	원치선	김도경
		D-3	140Hz	김영민	조명훈
		D-4	170Hz	장기진	백종욱
		D-5	200Hz	신형섭	주효성
3음절어	피치변화	T-1	80Hz	민일홍	정욱화
		T-2	118Hz	장성태	임진영
		T-3	140Hz	원상구	성낙간
		T-4	170Hz	송용화	반문종
		T-5	200Hz	권오현	이재석

Table 2. The Korean Monosyllable used in the Experiment

KOREAN SYLLABLE TYPES	KOREAN MONOSYLLABLES									
	1	2	3	4	5	6	7	8	9	10
V-TYPE	이 /i/	유 /yu/	우 /u/	어 /eo/	오 /o/	어 /eo/	위 /ui/	아 /a/	의 /eui/	의 /ui/
CV-TYPE	다 /da/	리 /li/	세 /seh/	사 /sa/	끼 /ggi/	표 /pyo/	노 /no/	가 /ga/	과 /goa/	수 /su/
VC-TYPE	일 /il/	원 /won/	연 /ryeon/	용 /eum/	양 /ryang/	입 /eup/	약 /yak/	역 /yeok/	육 /yuk/	입 /ip/
CVC-TYPE	정 /jeong/	관 /gwan/	행 /haeng/	장 /jang/	발 /bal/	뻡 /ppeop/	목 /mok/	급 /geup/	속 /sok/	작 /jak/

(The Korean Representation in Roman Characters)

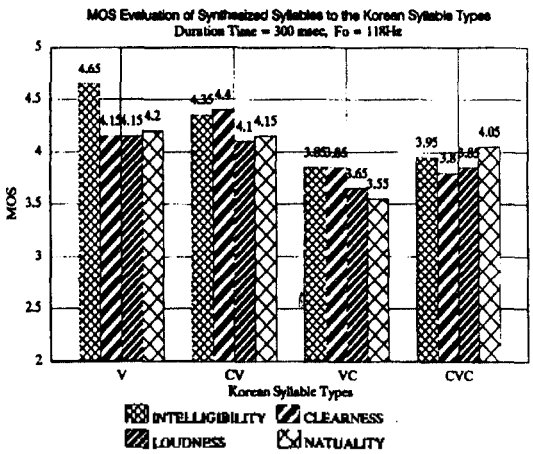

Fig. 6 The MOS Evaluation to the Korean Syllable Types

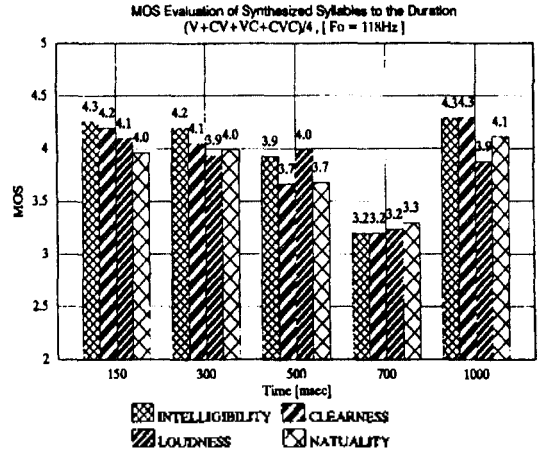

Fig. 7 The MOS Evaluation with the Change of the Durations

Ⅳ. The Subjective MOS Evaluation of a Synthetic Speech Quality

In this section we have presented the results of the evaluated synthetic speech quality in the Korean TTS(Text-to- Speech) system. First, we have constructed 229 syllables data format dictionary depending upon the frequencies of the Korean pronunciation large vocabulary dictionary[7], in which V type syllables are 19, CV type's are 80, VC type's are 30 and CVC type's are 100. And we have synthsized the Korean monosyllables with varying the durations and the fundamental frequencies in order to test a limit of TD-PCULI method. Then we also synthesized the Korean multisyllabic words in the same ways and measured the MOS scores to them.

1. The MOS Evaluation of Monosyllables with the change of the Durations and the F_0

Table 1 is to demonstrate the experimentation method presented in this paper. As shown in the table 1 we have synthesized various Korean mono-syllables which are different a duration(i.e., 150msec, 300msec, 500msec, 700msec and 1sec) and a central fundamental frequency(i.e., 80Hz, 118Hz,

Table 3. The Results of MOS Evaluation with the Change of F_0

(Duration : 300msec)

KOREAN SYLLABLE TYPES		80Hz				118Hz				140Hz				170Hz				200Hz			
		①	②	③	④	①	②	③	④	①	②	③	④	①	②	③	④	①	②	③	④
V-type	A	4.0	3.6	4.0	2.8	4.7	4.0	4.0	4.1	4.2	4.2	4.4	4.1	4.7	4.6	4.6	4.5	3.0	2.7	2.9	2.7
	B	2.8	2.7	2.7	2.7	4.6	4.3	4.3	4.3	4.0	3.7	3.9	3.5	4.6	4.5	4.4	4.0	4.2	4.0	4.2	3.9
CV-type	A	3.7	3.8	4.1	3.8	4.2	4.2	4.0	3.9	4.2	4.0	4.1	4.0	4.8	4.8	4.8	4.8	2.9	2.8	3.1	2.8
	B	4.2	4.0	4.2	4.4	4.5	4.6	4.2	4.4	4.0	3.9	4.0	3.7	3.6	3.6	3.6	3.6	4.9	4.8	4.8	4.3
VC-type	A	2.5	2.8	2.8	2.6	4.2	4.2	3.9	3.8	3.6	3.4	3.8	4.0	4.4	4.4	4.4	4.4	2.4	2.2	3.2	2.2
	B	3.3	3.4	3.9	3.4	3.5	3.5	3.4	3.3	4.4	4.3	4.3	3.8	4.2	4.2	4.2	4.1	4.3	4.1	4.3	3.6
CVC type	A	3.5	3.6	3.7	3.8	3.6	3.6	3.8	4.0	3.7	3.8	4.2	4.2	4.4	4.4	4.4	4.4	1.8	1.6	2.4	1.0
	B	4.3	3.9	4.3	4.2	4.3	4.0	3.9	4.1	4.3	4.3	4.3	3.9	4.2	4.2	3.6	3.7	3.2	3.4	3.6	2.7
Total		3.5	3.5	3.7	3.5	4.2	4.1	3.9	4.0	4.1	4.0	4.1	3.9	4.4	4.6	4.3	4.2	3.3	3.4	3.6	2.9

(Here, ①:INTELLIGIBILITY, ②:CLEARNESS, ③:LOUDNESS, ④:NATURALITY)

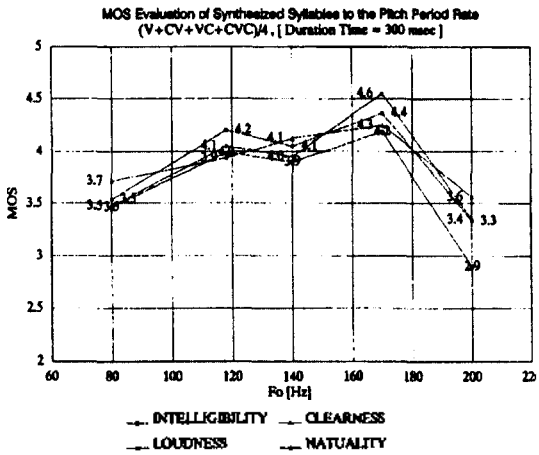

Fig. 8(a) The MOS Evaluation with the Change of the F_0

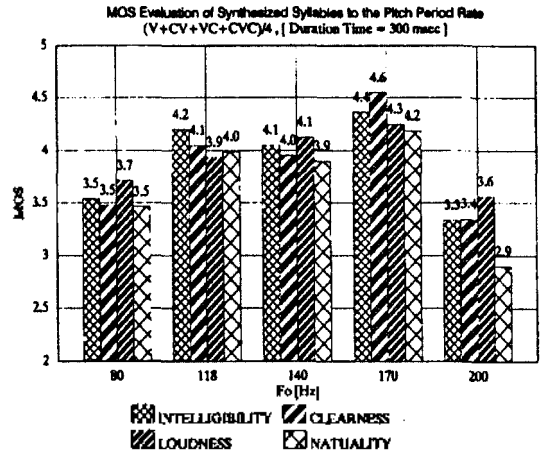

Fig. 8(b) The Histogram of Fig. 8(a)

140Hz, 170Hz, and 200Hz) in a mono-syllable using the extracted parameters such as a amplitude, a duration and a pitch period in each mono-syllables through the analysis of speech waveforms in the time domain. Table 2 is to demonstrate the Korean monosyllables used for the expeirments. We had tested a MOS evluation to each 10 syllables selected according to the 4 Korean syllable types about the 4 items i.e., intelligibility, clearness, loudness, and naturality after selecting random group without the knowledge of the synthetic speeches. Fig. 6 is the result of an experiment that according to the 4 Korean syllable types given in table 2 we have synthesized monosyllables with a duration in 300msec and tested a MOS evaluation about 4 items. As shown in fig. 6 V-type and CV-type are determined to be very good and VC-type and CVC-type are estimated to be good. Fig. 7 is the result of the MOS evaluation to an experiment that we have synthesized each syllables with 5 kinds of durations(i.e., 150msec, 300msec, 500msec, 700msec and 1sec) in order to test the limit of a duration in synthetic syllables. As shown in fig. 7 the results are estimated above a good level. Table 3 is the results of a MOS test to the possibility to modify a fundamental frequency F_0. And the results are also estimated to be above a good level as in fig.8(a) and fig.8(b).

Table 4. The Korean Words used in the Experiment

KOREAN WORDS	THE KOREAN WORDS USED IN THE EXPERIMENT														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2-SYLLABIC WORDS	안 녕	종 이	형 편	근 심	바 보	교 회	설 날	작 품	학 교	명 증	화 재	안 부	여 우	임 신	감 원
3-SYLLABIC WORDS	어 머 니	조 미 료	불 면 증	야 유 회	항 아 리	무 수 히	방 위 병	아 저 씨	용 왕 님	어 저 께	청 포 도	대 둘 보	공 경 심	아 버 지	안 녕 히

Table 5. The Results of MOS Evaluation to the Words with the Change of F_0

KOREAN WORDS		80Hz				118Hz				140Hz				170Hz				200Hz			
		①	②	③	④	①	②	③	④	①	②	③	④	①	②	③	④	①	②	③	④
2-SYLLABLE	A	4.1	3.9	3.9	3.9	3.7	3.7	4.0	3.5	3.9	3.7	3.5	3.8	2.9	3.1	3.5	3.1	3.9	3.4	3.3	3.5
	B	4.3	4.1	4.1	4.0	3.6	4.0	4.0	4.2	3.4	3.3	3.2	2.8	3.4	3.0	3.5	2.4	4.3	4.3	4.5	4.3
Total		4.2	4.0	4.0	4.0	3.7	3.9	4.0	3.9	3.7	3.5	3.4	3.3	3.2	3.1	3.5	2.8	4.1	3.9	3.9	4.0
3-SYLLABLE	A	3.3	2.8	3.6	2.6	3.7	3.6	3.3	3.3	3.9	3.7	3.9	3.5	2.8	2.9	4.1	2.9	2.4	2.9	3.4	2.1
	B	2.5	2.6	2.5	2.2	1.9	2.4	2.7	2.2	2.9	3.0	4.0	3.1	2.3	4.9	4.9	1.8	3.2	3.8	3.5	3.2
Total		2.9	2.7	3.1	2.4	2.8	3.0	3.0	2.8	3.4	3.4	3.9	3.3	2.5	3.9	4.5	2.3	2.8	3.3	3.4	2.7

(Here, ①:INTELLIGIBILITY, ②:CLEARNESS, ③:LOUDNESS, ④:NATURALITY)

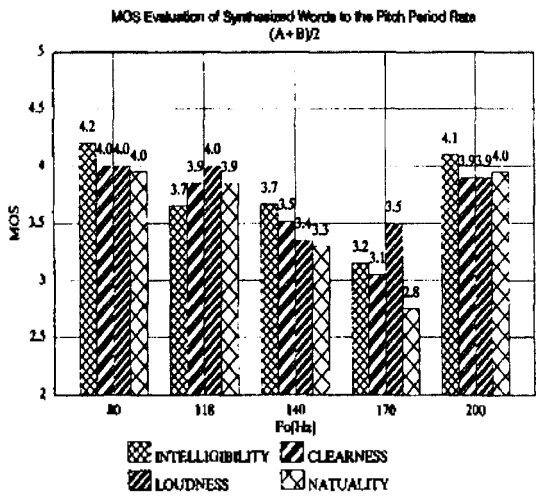

Fig. 9(a) The MOS Evaluation to the Dissyllabic Words with the Change of the F_0

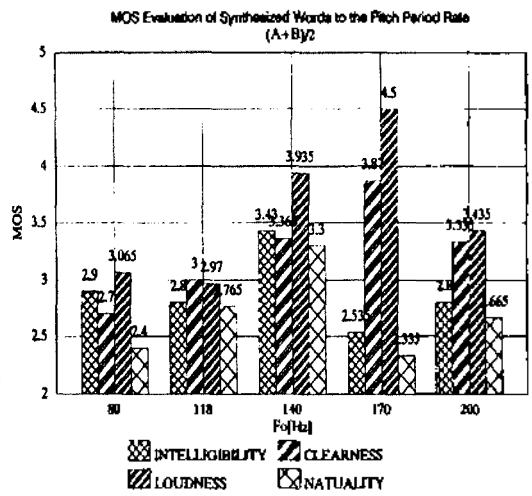

Fig. 9(b) The MOS Evaluation to the Trisyllabic Words with the Change of the F_0

2. The MOS Evaluation of Synthesized Multisyllabic Words with the Change of the F_0

Table 4 represents Korean words used in the MOS test to the dissyllabic and the trisyllabic words. We have measured the score with the changing of fundamental frequencies as shown x-axis in fig. 9(a) and fig. 9(b). The result is given in table 5. In the case of dissyllabic word the quality is estimated to be a good level but in trisyllabic words it is estimated to be a not bad level. The reason why trisyllabic words are inferior to dissyllabic words seems to be the inappropriate pitch patern. To remove the phenomenum we have been studing on making pitch patterns in Korean speech[9].

V. Conclusion

TD-PCULI is a new method revised to develop the Korean TTS system. In this paper we have tested each 10 syllables selected according to the 4 Korean syllable types with the objective MOS(Mean Opinion Score) evluation method about the 4 items i.e., intelligibility, clearness, loudness, and naturality after selecting random group without the knowledge of them. As the results of that V-type and CV-type are measured to be very good and VC-type and CVC-type are estimated to be good. And we have the MOS evaluation to an experiment that we have synthesized each syllables with 5 kinds of durations(i.e., 150msec, 300msec, 500msec, 700msec and 1sec) in order to test the limit of a duration in synthetic syllables. And also we have tested the possibility to modify F_0 into another forms with changing a central fundamental frequency(i.e., 80Hz, 118Hz, 140Hz, 170Hz, and 200Hz). As the results of that the clearness is improved and aslo prosody can be controled to a level as the synthesis methods in the frequency domain.

<REFERENCES>

1, C. H. Kang, Y. O. Chin, others, "Speech Synthesis In the Time Domain by Pitch Control using Lagrange Interpolation(TD-PCULI)," WESTPRAC V proc., 1994. 8

2. The Korea Broadcasting Corporation, The Standard Korean Pronunciation Large Vocaburary Dictionary, Eo Mun Gak Publishing, 1993

3. Nobuhiko Kitawaki, Hiromi Nagabuchi, "Quality Assessment of Speech Coding and Speech Synthesis System," IEEE Comm., 1988. Vol26. No.10

4. Toshiro Watanabe, "規則合成音の自然性評價法の檢討", 電子情報通信學 會論文誌, A Vol. J74-A No.4, 1991

5. J. H. Kim, S. H. Kang, "A Study on the Standardization of Subjective Assessment of Speech Quality," Electronic Communication Movement Analysis, 1990.7

6. Jonathan Allen, M. Sharon Hunnicutt and Dennis Klatt, From Text to Speech : The MITalk system, Cambridge Univ. Press, 1987

7. Shuzo Saito, Fundamentals of Speech Signal Processing, Academic Press, 1981

8. G. Rigoll, "The DECtalk system for German : A study of the modification of a text-to-speech converter for a foreign language," IEEE Proc. ICASSP '87, 1987

9. C. H. Kang, "A Study on the Korean Speech Syntehsis-by-Rule using Unit Pitch Frame Informations," Kyunghee Univ. Doctorate Thesis, 1994. 8

10. C. H. Kang, Y. O. Chin, "Speech Synthesis for the Korean large Vocabulary Through the Waveform Analysis in Time Domains and Evauation of Synthesized Speech Quality," The Journal of the Acoustical Society of Korea, Vol. 13, No. 1, 1994.

11. C. H. Kang, J. H. Lee, J. K. An, K. H. Kwon, S. T. Sea., Y. O. Chin, "The Evaluation of Speech Quality Synthesized by Rule to Korean Syllable Types," 1993 Conference of the Acoustical Society of Korea, Vol.12, No.1, 1993. 12.

12. C. H. Kang, Y. O. Chin, "Development of Speech Synthesizer in Korean TTS System," The Journal of the Acoustical Society of Korea, Vol. 12, No. 2, 1993.2