Algorithm for Concatenating Multiple Phonemic Units for Small Size Korean TTS Using RE-PSOLA Method*

Il-Suh Bak** · Cheol-Woo Jo**

ABSTRACT

In this paper an algorithm to reduce the size of Text-to-Speech database is proposed. The algorithm is based on the characteristics of Korean phonemic units. From the initial database, a reduced phoneme unit set is induced by articulatory similarity of concatenating phonemes. Speech data is read by one female announcer for 1000 phonetically balanced sentences. All the recorded speech is then segmented by phoneticians. Total size of the original speech data is about 640 MB including laryngograph signal.

To synthesize wave, RE-PSOLA (Residual-Excited Pitch Synchronous Overlap and Add Method) was used. The voice quality of synthesized speech was compared with original speech in terms of spectrographic informations and objective tests. The quality of the synthesized speech is not much degraded when the size of synthesis DB was reduced from 320 MB to 82 MB.

Keywords: Speech, Database, Search Algorithm, RE-PSOLA

1. Introduction

Recently there are many Korean Text-to-Speech(TTS) systems which are commercialized at market and widely used on various applications. A corpus-based system is considered to be a good method to achieve good voice quality of synthetic speech[1]. But a corpus-based synthesis system requires a large amount of database and memory to find an optimum synthesis unit for the given sentence.

But the need of small size TTS is increasing following the widespread use of portable computing devices such as PDA etc. For such devices synthesizers with mid-size DB or small-size DB are required and a method to reduce the size of the database and the algorithm is essential.

In this paper we propose a method to reduce the size of synthesis DB by limiting the kinds of phonemic units based on the statistical characteristics of mother corpus and

* This study was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (project No. 02-PJ1-PG3-31402-0004).

** SASPL, School of Mechatronics, Changwon National University

applying RE-PSOLA method to limited before and behind environment diphone. And an algorithm to find an optimal unit sets is suggested. The synthesized speech is tested by a listening test.

2. Statistical Characteristics of Synthesis Unit

The size of the synthesis database depends on the choice of the kind of synthesis units. The following is a comparison of the general characteristics of the different synthesis units in terms of quality, complexity of algorithm, size of DB.

(1) Phoneme: The number of units is smallest, but very difficult to obtain good quality of speech due to the difficulty of finding a nice rule to concatenating the units.
(2) Diphone: It contains a transient part of speech and the size of DB is intermediate.
(3) Triphone: It contains a transient part between phonemes and is easy to get good quality speech, but size of database is biggest.

First we tried to find proper synthetic units based on the frequency of each unit[3]. The statistics are found from the speech corpus which contains 1,000 sentences. Table 1 shows the frequency of each synthesis unit. Diphone unit includes CV, CC, VC, VV, triphone unit includes CVC, VCV, CVV, VCC, CCV, VVC.

Table 1. The frequency of synthesis unit

synthesis unit	description	frequency	kinds
2phoneme	CV	29,007	255 (63%)
	VC	28,423	274 (72%)
	CC	9,301	122 (34%)
	VV	3,863	198 (50%)
3phoneme	CVC	24,915	3,553 (49%)
	CVV	3,323	1,172 (15%)
	CCV	9,301	919 (12%)
	VCV	18,943	4,800 (63%)
	VCC	9,301	902 (12%)
	VVC	3,303	1,214 (15%)

From Table 1, the number of diphone is 70,594 and frequency of CV, VC is 81% of total diphone. Meanwhile there are 69,086 triphones and CVC and VCV are 63% of the

total triphone. Numbers in the parenthesis mean the percentage of the phonetic environment which is included in the statistics.

Frequencies of diphones and triphones in the corpus is shown in table 2. Considering frequencies of diphones and triphones, frequencies of the specific synthesis unit is sought for CV, VC, VV, CC, CVC, VCV. CVV, VCC, CCV, VVC showed a relatively less number of unit kinds compared to the class frequence and are not considered as a candidate here.

From Table 2, 125 and 106 units are missing in the class CV, VC respectively. But they are found to be an un-realistic combination of phonemes, which is not possible for real speech. Those consist of units with strong consonants or double vowels. In case of CC and VV, the number of identifiable synthesis unit is restricted because of the strong consonants or double vowels. Based on this fact, diphone unit is enough to be used as synthesis units.

Table 2. Frequencies of Each Phonemic Unit

description rate	CV	CC	VC	VV	CVC	VCV
0	125	239	106	202	3,667	2,800
1~20	85	51	106	136	2,077	1,692
21~40	31	16	45	39	157	135
41~60	26	12	25	12	50	45
61~80	17	6	16	3	34	26
81~100	18	10	12	2	9	6
101~120	9	2	7	2	12	6
121~140	10	4	12	1	3	2
141~160	6	3	5	3	1	4
161~	53	18	46	0	8	4

From the triphone units, CVC or VCV whose kinds are large, missed 3,367 and 2,800 of the kinds respectively. From CVC and VCV , whose numbers are large, 3,367 and 2,800 units are not observed. Most of them are units which are un-realizable. Because the frequency of the units are not even, it is required to design the corpus text carefully and the database becomes oversized in this case. Based on the considerations above, we tried to devise an algorithm to find the optimal units considering phonemic environments.

3. Search Algorithm

3.1 Text-to-Speech System

The structure of the TTS system, which is used in this experiment, is shown in Fig 1. Total system consists of 3 parts, which are the linguistic processing part, prosody processing part, and speech signal processing part.

The linguistic part analyses the input sentence and generates various linguistic information. In this stage, it analyses the input text in the form of Korean code and substitutes symbols which corresponds to the phoneme sequences. Also a prosody or an accent information is generated from the result of the analysis of sentences. The results are used to control the prosody of the synthesized speech.

The signal processing part generates the encoded speech wave. Synthesis units consist of diphone classes (CV, VC, CC, VV) and phonetic informations are used in the process. Because the synthesis units are decided into diphones and phonemes, all the speech corpus is segmented and labeled. The constructed database is encoded into REPSOLA method.

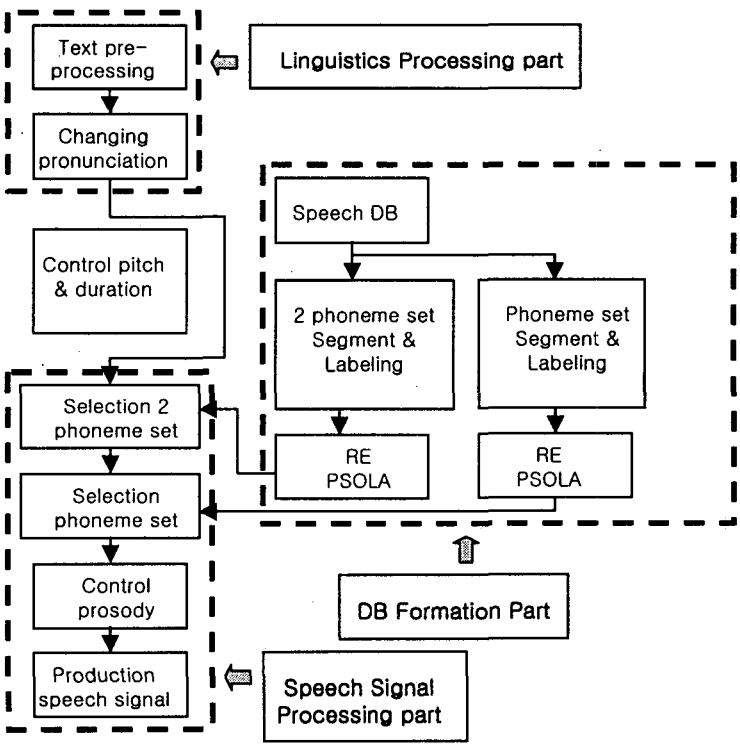

Figure 1. The structure of TTS system

3.2 Search Algorithm

In current corpus based synthesizers, it was possible to select synthesis units with enough number of different phonetic environments. But for synthesis systems, which do not have a large enough database, a new algorithm is required to find the sub-optimal units from the small database.

A suggested search algorithm in this paper, mapping of CV unit comes first. The CV unit has the most frequently observed unit class. And later considering the relations between the preceding phoneme and the current unit and the relations between the current unit and the following unit. After that CC class and VV class if searched. By doing so not all of the unit class is sought and we can reduce the number of search paths. Figure 2 shows a flow of search algorithm.

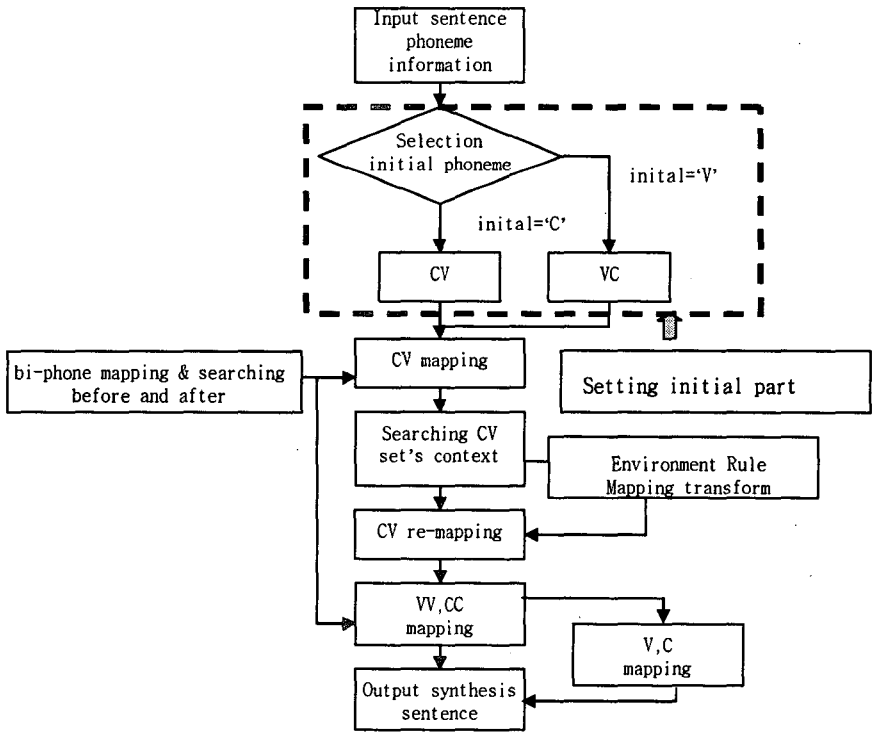

Figure 2. The structure of searching algorithm

To make the transient part at diphone concatenation more natural, preceding and following phonemes information are written together. The class is divided into 5 kinds according to the position of the tongue, and into 4 kinds according to the position of the tongue and 1 silence. To find the precedence sequence, preceeding and following environmental informations according to each phoneme are used. They are silence

interval, articulatory positions, duration of each phoneme etc. Based on such environmental information the most probable phoneme is selected from the database.

Figure 3 shows the similarity and precedence table of consonants and figure 4 is for the vowels.

Table 3. Similarity Table of Consonants

1	ㅁ= ㄴ(= ㄹ) > ㅇ	
2	ㄱ > ㅋ = ㄲ	
3	ㄷ > ㅌ = ㄸ	
4	ㅅ > ㅈ > ㅆ (= ㅊ) >ㅉ	
5	ㅎ > ㅂ > ㅍ	

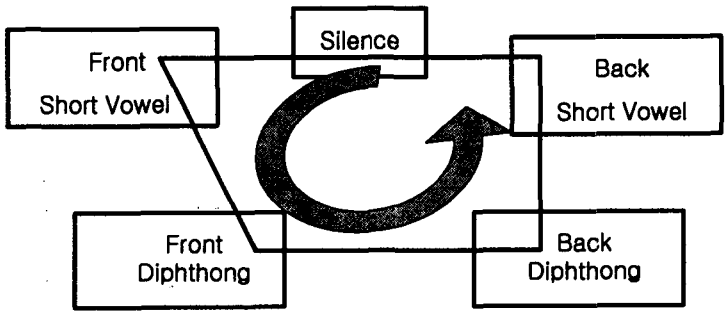

Figure 3. Search Priorities of Vowels

4. Results and Discussions

Applying the suggested reduction algorithm, the total size of database reduced from 320 MB to 82 MB. To verify the validity of the search algorithm, 82 Mbyte of speech corpus which consist of 15,000 diphones and phonemes with the sampling rate of 11,025 kHz and the resolution of 8 bit. Then the search process is performed in the order of CV, CC, VC, VV, C, V.

The following is the list of sentences which are used for the evaluation of TTS.

Sentences which are included in the original database.

-반사되는 빛의 스펙트럼을 시간에 대한 함수로 측정함으로써 광섬유 위치에 따른 온도나 인장 강도의 변화를 측정할 수 있다.

-오이는 소금에 살짝 절여야 짜지 않고 모양이 늘어지지 않는다.

-스카이라운지 20% 할인

-소련의 비협조 때문이었다.

Sentences which are not included in the original database.

-서울 시가지 교통정보입니다.

-창원대학교 음성 및 음향 신호처리 실험실입니다.

The former 4 sentences are chosen from the original database and the latter 2 are randomly selected ones. Each synthesized speech is analyzed and compared with the original speech or with the speech from different stages of synthesizer. It was shown that the synthesized speech is heard natural except on the unnatural preceeding silence.

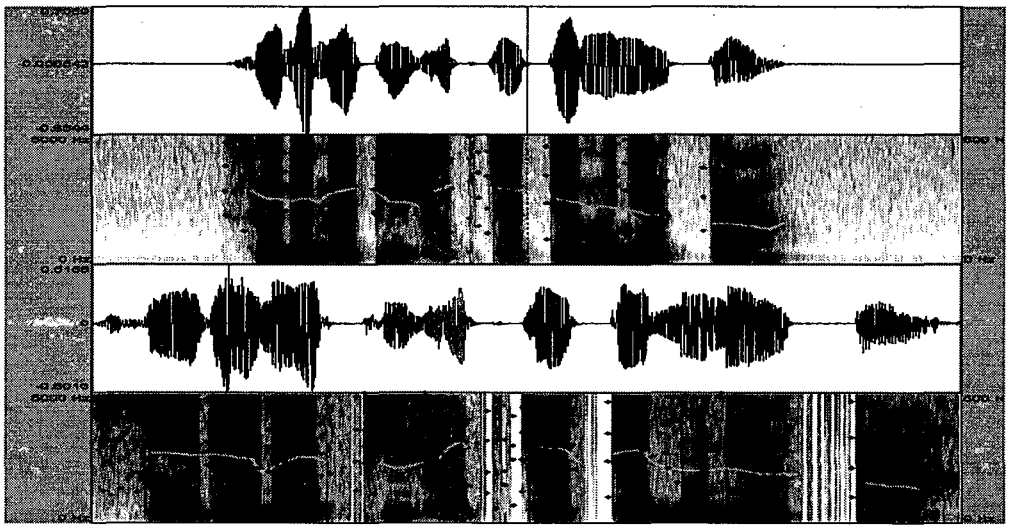

Figure 4. Original Speech vs. Synthesized Speech

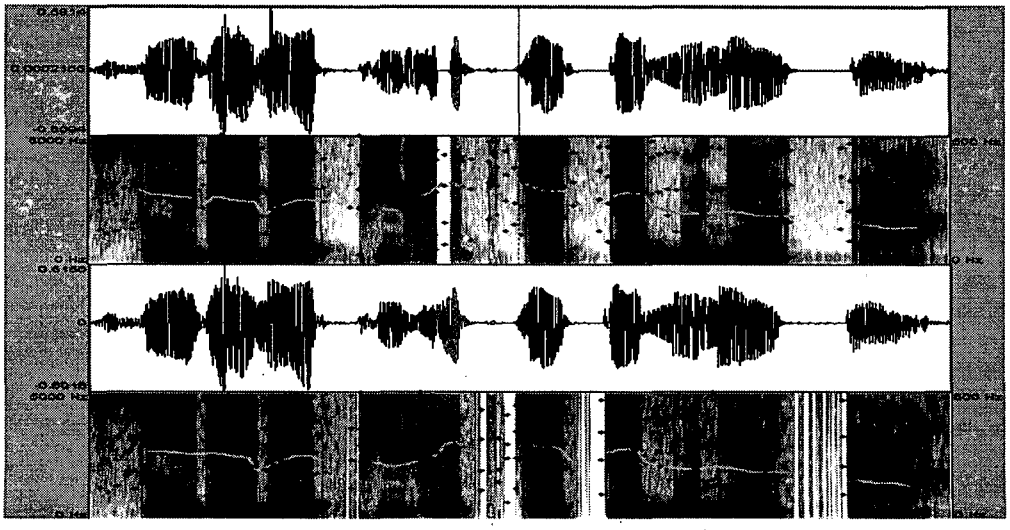

Figure 5. Concatenated Speech vs. Encoded and Synthesized Speech

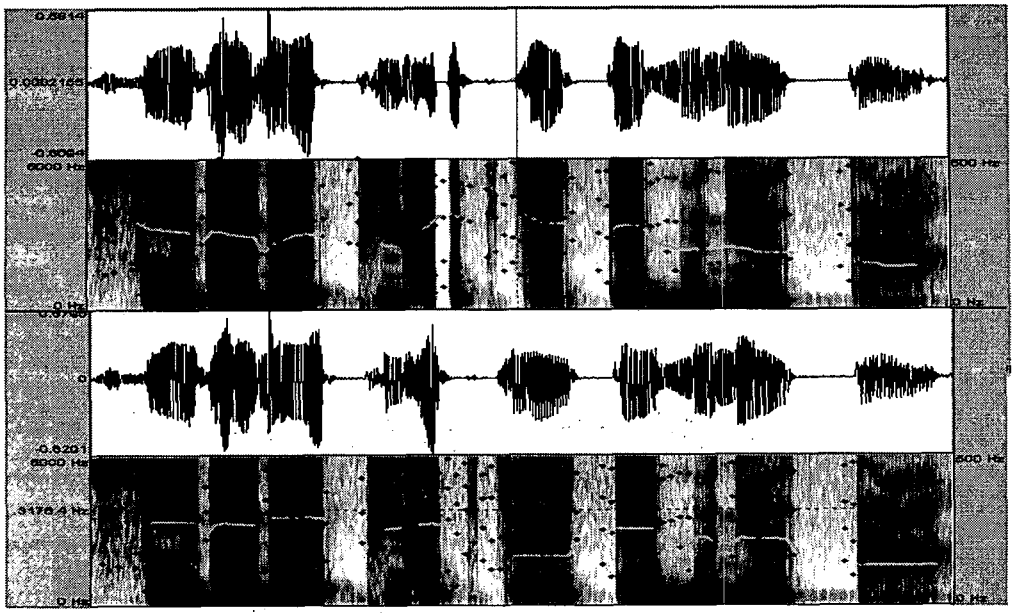

Figure 6. Encoded and Synthesized Speech vs. Speech with Prosody Info

Figure 7 shows the spectrographic analysis result for the sentence which is not included in the original DB. The sentence is '서울 시가지 교통 정보입니다.' Because there is no original speech, comparison is done on the basis of spectrographic domain.

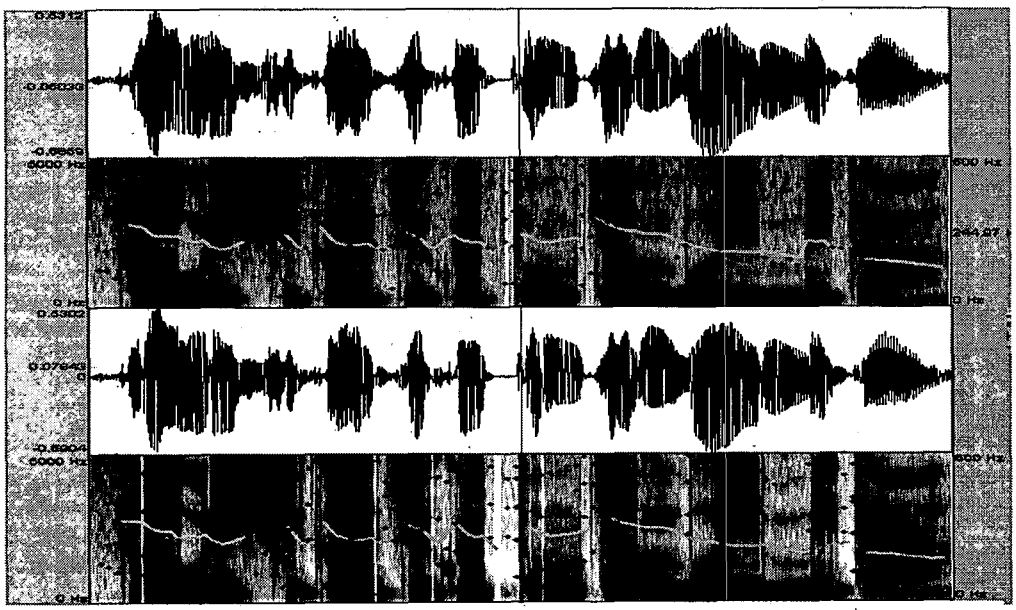

Figure 7. Original Speech vs. Synthesized Speech for the Sentence which is not in the Original DB

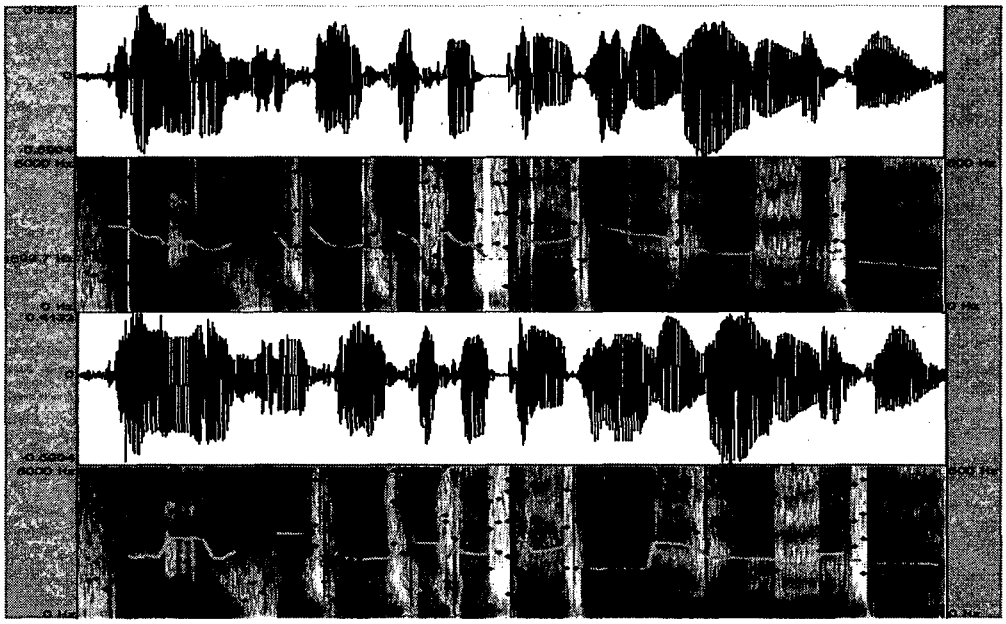

Figure 8. Encoded Speech vs. Speech with Prosody Info for the Sentence which is not in the Original DB

By comparing the synthesized speech and the natural speech, it was observed that the quality of the synthesized speech is not much degraded after reduction. The test text is chosen among the corpus database but the chosen units are not the same as the original one. By comparing the subjective listening test, some unnatural sound is found at the transient region. And additional correction of database is going on.

5. Conclusion

In this paper we proposed an algorithm to search a proper synthesis units considering phonetic environments. The unit class is chosen from the statistics of the speech corpus. And the synthesized sample sentences are used to compare the quality between natural speech and synthetic speech. By using the search orders based on statistical characteristics, it was possible to reduce the unnecessary searches. And the reduction of speech DB can make it applicable to the small database systems such as PDA. Currently additional processing to reduce the size of the speech DB is going on without losing much of the quality.

References

[1] Nakajima, S. & H. Hamada. 1981. "Automation generation of synthesis units based on context oriented clustering." *IEEE International Conference on Acoustics, Speech and Processing*, New York, 659-662.

[2] Black, A. W. & N. Campbell. 1995. "Optimizing selection of units from speech database for concatenate synthesis." *Proceeding of EUROSPEECH'95*, Spain, 573-576.

[3] Kim, S. & J. C. Lee. 1994. "Korean Text-To-Speech system using TD-PSOLA." *Proc. of SST94*, 587-592.

[4] Abe, Y. & S. Imai. 1981. "Speech synthesis from CV-syllable cepstral parameters." *Trans. IECEJ64-D*, 861-868.

[5] Valbret, H., J. P. Moulines Tubach. 1991. "Voice transformation using PSOLA Technique." *EUROSPEECH 91*, 345-348.

Received: January 24, 2003
Accepted: February 27, 2003

▲ Il-Suh Bak
SASPL, School of Mechatronics, Changwon National University
9 Sarim-dong, Changwon, Kyongnam, 641-773, Korea
Tel: +82-55-279-7559 Fax: +82-55-262-5064
H/P: 011-9310-9741
E-mail: ilsuh@korea.com

▲ Cheol-Woo Jo
SASPL, School of Mechatronics, Changwon National University
9 Sarim-dong, Changwon, Kyongnam, 641-773, Korea
Tel: +82-55-279-7552 Fax: +82-55-262-5064
E-mail: cwjo@sarim.changwon.ac.kr