시간영역에서의 파형분석에 의한 무제한 어휘 합성 및 음절 유형별 규칙합성음 음질평가

Speech Synthesis for the Korean large Vocabulary Through the Waveform Analysis in Time Domains and Evauation of Synthesized Speech Quality

姜 贊 熙*, 陳 庸 玉**

(Chan Hee Kang* and Yong Ohk Chin**)

요 약

본 논문은 한국어 문어변환(TTS : Text-to-Speech) 시스템내에서의 음성합성시 음질 및 자연성 개선을 위한 연구 결과이다. 합성방법으로는 단음절단위의 파형을 시간영역에서 분석(표1)하여 규칙합성에 필요한 매개변수(표2)를 추출하여 규칙합성시켰다. 실험에 사용된 음절은 한국어 발음 대사전의 빈도순위에 따라 V형 19개, CV형 80개, VC형 30개, CVC형 100개등 총 229음절을 선정하여 규칙합성시켰다. 규칙합성음의 평가방법으로는 229개의 규칙합성음중 음절 유형별로 15개씩 무작위로 추출한 합성음을 사전지식이 없는 임의의 그룹을 선정하여 이해도, 명료도, 잡음감, 자연성등 4가지 항목에 대하여 주관적인 오피니온 평가를 수행하였다. 실험결과, 합성음의 음질은 대단히 명료한 수준이었으며, 운율요소의 제어결과는 지속시간(장단)과 악센트(강약)의 제어(그림 9, 그림 10)가 가능하였으며, 피치주기(억양)의 제어도 Lagrange 보간법을 사용함으로써 가능하였다(그림 11, 그림 12).

ABSTRACT

This paper deals with the improvement of the synthesized speech quality and naturality in the Korean TTS (Text-to-Speech) system. We had extracted the parameters(table2) such as its amplitude, duration and pitch period in a syllable through the analysis of speech waveforms(table1) in the time domain and synthesized syllables using them. To the frequencies of the Korean pronunciation large vocabulary dictionary we had synthesized speeches selected 229 syllables such as V types are 19, CV types are 80, VC types are 30 and CVC types are 100. According to the 4 Korean syllable types from the data format dictionary(table3) we had tested each 15 syllables with the objective MOS(Mean Opinion Score) evaluation 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 experiments the qualities of them are very clear and we can control the prosodic elements such as durations, accents and pitch periods(fig9, 10, 11, 12).

* 상지대학교 병설 전문대학 전자과
Dept. of Electronics in Sangji junior college
** 경희대학교 전자공학과
Dept. of Electronic Engineering in Kyunghee Univ.

접수일자 : 1993년 11월 5일

Ⅰ. 서 론

언어는 인간과의 의사전달 수단으로써 발성기관을 통하여 생성되어진 음성을 매체로 사용되는 청각적인 정보전달 수단의 한 방편이며, 문자는 시각적인 의사전달 수단이다. 따라서 음성합성의 최소한의 목적은 정보전달에 있으며, 음성합성시 정보전달의 능률을 극대화시키기 위하여는 반드시 자연성을 고려하여야 한다. 음성합성에서의 자연성이란 인간의 다양한 감정이 음성에 표출되어 전달되므로 강약과 장단과 쉼과 억양등과 같은 운율요소를 인위적인 조작으로 합성음에 부여시켰을 때 원음에 일치한 정도라 할 수 있으며, 이들요소는 서로 분리되기 어려운 통합된 상관관계[1),2),3)]를 지니고 있다.

일반적으로 음성합성방식은 3가지 부류로 분류할 수 있는데 첫번째는 조음기관의 물리적 특성을 전기적 등가회로로 구성하여 조음기관(articulator)을 제작시키는 방법이 있으나, 이는 실제의 조음기관과의 수학적 오차가 너무 클 뿐만아니라 효율적인 제어가 힘들어 최근에는 연구가 이루어지지 않고 있다. 두번째는 DSP(Digital Signal Processor) 칩을 이용한 주파수영역에서의 합성방식[4),5),6)]을 들 수 있다. 이는 인간의 물리적인 조음모델을 주파수영역에서 형상화시켜 합성하는 방법으로써 최근에 까지도 이 방법이 주류를 이루고 있으며, 그 방법 또 한 다양하게 존재한다. 이 방법이 지니는 장점은 시간영역에서의 합성방식에서는 극히 어려운 운율요소의 제어가 가능하고 시스템 구현이 간단하여 문어변환시스템 내에서의 음성합성기로서는 주파수영역에서의 음성합성방이 주류를 이루고 있다. 텍스트 합성장치의 거의 대부분은 주파수영역에서의 LSP, PARCOR, FORMANT 방식등을 이용하고 있다. 반면에 시간영역에서의 합성방식은 음질은 좋으나 운율요소의 제어가 어려워 최근까지도 녹음재생기능을 지닌 자동응답장치에 주로 국한되어 사용되어졌다[7)]. 주파수영역에서의 합성방법 또한 인간의 물리적인 조음기관을 완벽하게 구현시키기에는 한계가 있어 시간영역에서 합성시킨 음질 보다는 떨어진다.

시간영역에서의 운율제어의 어려움이란 원 파형에 임의의 함수를 가하거나 조작 변경시키면 피치주기성을 상실하기 쉬울 뿐만아니라, 합성하고자 하는 파형이 자연 생성되어진 파형보다 심하게 왜곡되어 음질이 저하되거나 변질되는데 문제가 있다. 본 논문에서는 이러한 시간영역에서의 합성방식이 지니고 있는 운율제어의 한계성을 극복하여 양질의 규칙합성음을 발생시키고자 하였다. 합성방법으로는 파형분석과정을 거쳐 합성시키고자 하는 파형의 진폭, 지속시간과 피치주기간격등과 같은 성분을 합성용 매개변수로 추출시켜 규칙합성용 데이타 포멧 사전을 구성하여 합성하였다. 데이타 포멧 사전에 등록된 음절은 한국어 발음 대사전[8)]의 빈도순위에 따라 V형 19개, CV형 80개, VC형 30개, CVC형 100개등 총 229음절이다. 규칙합성음의 평가방법[9),10),11),12)]으로는 합성음을 음절 유형별로 15개씩 선정하여 20 초간 3회 반복음의 녹음 테이프를 작성한 후 합성음에 대하여 사전지식이 없는 임의의 그룹을 선정하여 이해도, 명료도, 잡음감, 자연성등 4 가지 항목에 대하여 주관적 평가법에 의한 오피니온 평가를 수행하였다.

Ⅱ. 음성합성 알고리즘

시간영역에서의 규칙합성시 텍스트로 부터 변환된 음소 기호열에 따라서 저장된 음성 데이타를 액세스하여 합성시킬 경우에는 운율요소의 제어가 용이하지 못하여 서론에서와 같이 자연스런 합성음을 발생시킬 수 없다. 본 논문에서는 이와같은 문제점을 해결하기 위하여 표1에 제시된 파형분석과정을 거쳐 합성시키고자 하는 파형의 진폭, 지속시간과 피치주기간격등과 같은 성분을 합성용 매개변수로 추출시켜 규칙합성용 데이타 포멧 사전을 작성(표2)하여 합성에 이용하였다. 그림 1은 규칙합성에 사용되는 운율제어용 매개변수를 추출하기 위한 블록도이며, 그 과정을 설명하면 다음과 같다.

음성파형으로 부터 규칙합성용 매개변수를 추출하기 위하여 임의의 음성데이타를 x(n), 단음절의 데이타 갯수를 N, 단음절내에서의 1 피치주기의 프레임 갯수를 N_P로 각각 정의하면 단음절의 음성 데이타 열은 $\sum_{n=1}^{N} x(n)$으로 표기된다. 이 때 각각의 피치 프레임 구간의 경계를 P_{s1}, P_{s2}, P_{s3}, … 등으로 나타내고, 각 피치 프레임 구간에서의 데이타 갯수를 N_{Ps1}, N_{Ps2}, N_{Ps3}, … 등을 배열 $N_{Ps}(\ )$로 표기하면, N 개의 음성 데이타 열 $\sum_{n=1}^{N} x(n)$을 1 차원 배열인 1 피치 프레임 단위의 N_P개 소블록의 합으로 표기 가능하므로 이를 2 차원 배열로 표시하면,

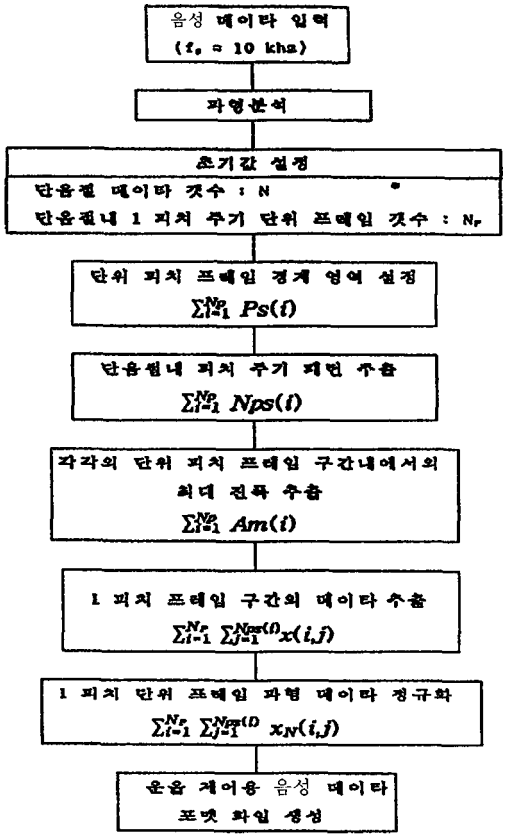


그림 1. 규칙합성용 매개변수 추출 흐름도

Fig 1. Extraction flowchart of parameters for synthesis by rule

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

(단, $n = n1 + \sum_{n1}(n1-1) \cdot N_{ps}(n1-1)$, $N_{ps(0)} = 0$ 임.)

와 같다. 여기서, x(5, 10)은 5번째 단위피치 구간의 10번째 데이타를 의미한다. 즉, x(n1, n2)는 x(음절내 단위피치 구간 번호, 단위피치 구간내 데이타 번호)를 의미한다. 또한, 단위 피치 프레임 구간내에서의 음성 데이타 열의 최대진폭의 절대치를 각각 A_{m1}, A_{m2}, A_{m3}, … 등으로 정의하고, 각각의 피치 프레임 단위 구간내의 데이타 열을 일정한 크기로 정규화시킨 임의의 음성 데이타를 $x_N(n)$로 정의하면, 2 차원 블록화 배열로 표시된 음성 데이타 $x(n) = \sum_{n1=1}^{Np} \sum_{n2=1}^{Nps(i)} x(n1, n2)$은

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

로 표시된다. 윗 식들로 부터 추출·저장하여 데이타 포맷 사전에 작성된 매개변수는 총 5개로써 이를 정리하여 보면, 식(1)과 식(2)에서 추출된 매개변수는

1) 단음절내 전체 데이타 갯수 정보 : N(2 바이트)
2) 단음절내 피치 갯수 정보 : N_P(1 바이트)
3) 각 피치 프레임 구간에서의 데이타 갯수 정보 : $\sum_{i=1}^{Np} N_{psi}$(N_P 바이트)
4) 단위피치 구간내에서의 최대진폭 정보 : $\sum_{i=1}^{Np} Am(i)$ (N_P 바이트)
5) 단위피치별로 정규화된 음성 데이타 : $\sum_{n1=1}^{Np} \sum_{n2=1}^{Nps(n1)} x_N(n1, n2)$(N바이트) 등이다.

Ⅲ. 음성합성예

그림 2는 본 논문에서 사용한 음성합성의 개략적인 과정을 블록도로 표시한 것이다. 이 그림에서와 같이 단음절 단위의 파형이 입력되면 맨 먼저 파형분석과정을 거쳐 합성에 필요한 파형정보를 추출하는 과정을 거친다. 그림 3에 표시된 CVC형 파형 "공"을 한 예로 들어 설명하면 3장의 식(1)에서 1음절어의 전체 음성데이타는 2,434 샘플 포인트이다. 이를 1 피치주기간격으로 분할하여 표시하면 표 1에서와 같이 23개의 구간으로 분할하여 표시할 수 있으며, 표2의 19번째 규칙합성용 포멧표에 표시된 1 음절내에 존재하는 1 피치주기의 구간갯수 NP는 23으로 주어진다. 피치구간갯수 NP는 1피치주기의 경계구간을 검출하여 구하였다. 이 때 각각의 피치주기를 정확히 검출하지 못하면 단음과 장음합성시 2가지의 중요한 잡음이 발생된다. 그림 4은 그림 3의 파형을 장음으로 규칙합성시켰을 때 이웃간의 파형의 접합부에서 불연속점이 발생하여 찌글거리는 잡음을 유발시키는 한 예를 표시한 것이다. 또 한가지 잡음은 위상왜곡으로써 추정된 피치주기간격이 짧아지거나 길어지면 장음이나 단음 합성시 원래의 파형이 지나고 있는 주기성이 흐트러지고 이로 인하여 위상왜곡이 발생하는 요인이 된다. 따라서 1 피치주기의 경계구간을 검출하기 위하여는 상당한 주의를 요하여야 한다. 표1에서 좌측 DATA POINT항은 파형상단부에서의 그 피치주기열 내의 최대값의 위치를 검색하여 표시한 것이며,

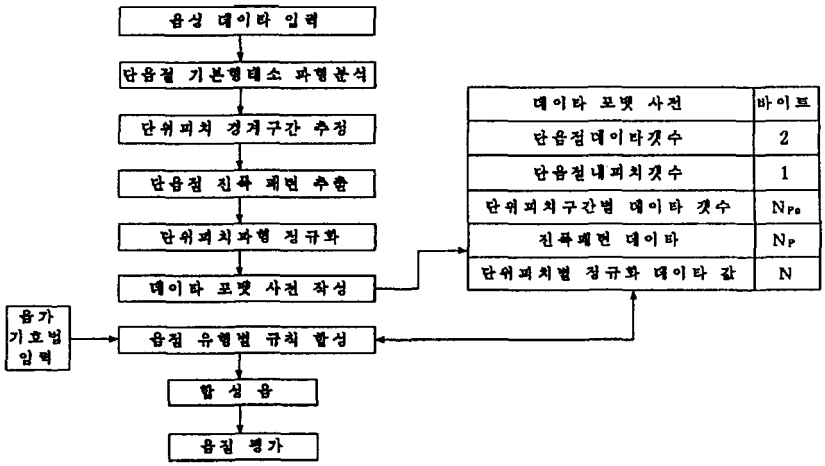


그림 2. 규칙합성 분록도
Fig 2. Block diagram of synthesis by rule

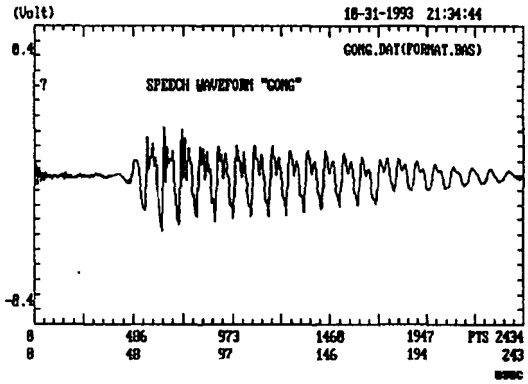


그림 3. 음성파형도 "공"
Fig 3. speech waveform of "gong"

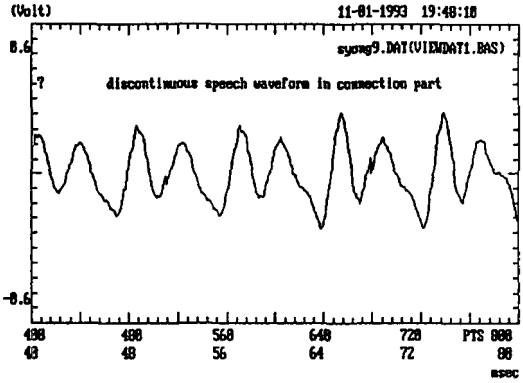


그림 4. 접합면에서의 불연속 잡음
Fig 4. discontinuity in connection part

표 1. 파형분석표
Table 1. waveform analysis table

SPEECH SIGNAL (GONG. DAT) ANALYSIS
(FORMAT. BAS)

No.	DATA POINT (PTS)	MAX (mV)	INTERVAL (PTS)	DATA POINT (PTS)	MIN (mV)	INTERVAL (PTS)	No. ZERO CROSSING	MAX RATE	MIN RATE
1	496	55	95	457	−23	70	1	0.30	−0.12
2	554	132	70	543	−118	85	2	0.71	−0.63
3	635	173	83	626	−187	84	4	0.93	−1.00
4	721	151	88	711	−165	86	6	0.81	−0.88
5	810	97	90	798	−139	88	4	0.52	−0.74
6	901	96	90	887	−158	88	3	0.51	−0.84
7	990	103	89	975	−145	88	4	0.55	−0.78
8	1080	101	90	1063	−141	88	2	0.55	−0.75

9	1169	99	89	1151	−138	89	2	0.53	−0.74
10	1258	85	89	1240	−134	89	2	0.46	−0.71
11	1348	82	90	1330	−127	90	2	0.44	−0.68
12	1437	82	89	1419	−114	89	2	0.44	−0.61
13	1527	75	90	1509	−106	90	2	0.40	−0.56
14	1617	63	108	1599	−99	90	2	0.34	−0.53
15	1743	57	91	1688	−98	90	1	0.31	−0.52
16	1798	58	73	1779	−54	91	2	0.31	−0.29
17	1889	50	94	1870	−49	87	2	0.27	−0.26
18	1986	39	95	1953	−47	85	2	0.21	−0.25
19	2080	30	93	2041	−42	91	2	0.16	−0.22
20	2172	27	93	2134	−31	92	2	0.15	−0.17
21	2265	24	94	2225	−26	90	2	0.13	−0.14
22	2360	17	50	2315	−18	93	1	0.09	−0.09
23	2366	15	71	2411	−10	71	2	0.08	−0.05

표 2. 데이타 포맷 예

Table 2. examples of data format

No.	FORMAT NAME	DS	NC	NTOTAL	NP	MAX	1 PERIOD	AMP. RATIO
1	I.PRM	0	0	2099	22	912	128 99 ... 87	76 .2 .5 ... .2 .1
2	DA.FRM	1	186	1674	16	2304	114 89 ... 93	73 .6 .5 ... .1 .1
3	IN.FRM	0	0	2488	32	2348	95 79 ... 69	32 .3 .4 ... .1 .0
4	JEONG.FRM	1	434	2392	22	2716	106 80 ... 89	49 .3 .8 ... .1 .1
5	LT.FRM	0	0	2000	22	2060	101 88 ... 92	82 .1 .2 ... .1 .1
6	WUEON.FRM	0	0	2796	33	3320	111 86 ... 72	66 .2 .4 ... .1 .1
7	DONG.FRM	1	287	2144	21	2672	88 78 ... 85	74 .3 .8 ... .1 .1
8	IL.FRM	0	0	2390	30	2464	92 77 ... 79	81 .2 .4 ... .1 .1
9	SANG.FRM	1	500	2940	27	2956	114 79 ... 89	75 .3 .8 ... .1 .1
10	GI.FRM	1	497	1884	15	1884	99 87 ... 92	77 .6 .7 ... .1 .1
11	SA.FRM	1	680	2350	19	3048	109 84 ... 89	80 .6 .8 ... .1 .1
12	JI.FRM	1	530	2142	18	1836	92 85 ... 90	85 .3 .5 ... .1 .1
13	AN.FRM	0	0	2298	28	1528	106 80 ... 74	63 .2 .6 ... .2 .1
14	JEON.FRM	1	475	2614	24	3052	113 83 ... 84	67 .5 .9 ... .1 .1
15	O.FRM	0	0	1642	17	2176	125 91 ... 99	70 .1 .3 ... .1 .1
16	EUM.FRM	0	0	2742	34	1872	109 82 ... 78	76 .1 .1 ... .1 .0
17	MUL.FRM	0	0	1920	22	2644	73 80 ... 85	70 .1 .3 ... .2 .1
18	EO.FRM	0	0	1654	18	2152	104 88 ... 85	74 .2 .4 ... .1 .1
19	GONG.FRM	1	430	2434	23	2448	70 85 ... 93	71 .1 .6 ... .1 .1
20	JA.FRM	1	424	2112	19	2328	98 79 ... 84	37 .2 .8 ... .1 .0
21	GA.FRM	1	310	2024	19	3088	103 90 ... 82	37 .4 .6 ... .1 .0
22	MUN.FRM	0	0	2540	28	3300	137 92 ... 91	51 .1 .3 ... .1 .0
23	A.FRM	0	0	1592	18	2488	99 57 ... 91	72 .1 .3 ... .1 .1
24	HA.FRM	1	535	2086	18	3468	109 77 ... 81	48 .2 .5 ... .1 .0
25	SEON.PRN	1	694	2592	22	3612	110 81 ... 54	17 .5 .9 ... .0 .0
26	EUI.FRM	0	0	2244	24	1512	108 88 ... 87	70 .2 .4 ... .2 .1
27	SU.FRM	1	1512	2932	16	2240	91 79 ... 90	115 .4 .6 ... .2 .1
28	EOP.FRM	0	0	1090	13	2696	107 84 ... 89	55 .1 .2 ... .2 .0
29	BU.FRM	1	329	1862	17	3204	129 84 ... 90	42 .5 .8 ... .1 .0
30	IN.FRM	0	0	2478	31	2048	102 80 ... 75	87 .2 .4 ... .2 .1

우측항의 DATA POINT항은 파형 하단부에서의 최대값 위치를 표시한것이다. MAX와 MIN항은 최대값과 최소값의 크기를 구한 것으로써 규칙합성시 진폭패턴을 제어할 수 있도록 1 음절어내에서의 최대값에 대한 상대비로 변환하여 표2의 데이타 포맷표(AMP RATIO)에 저장하여 규칙합성에 이용하였다. 표1에서의 INTERVAL항과 표2에서의 1PERIOD항은 각각의 1피치주기열의 데이타 수를 추정하여 표기한 것이며, 표2에서의 D$은 무성자음이 초성구간에 존재하는 유무를 표시한 것이다. 그리고 표2에서의

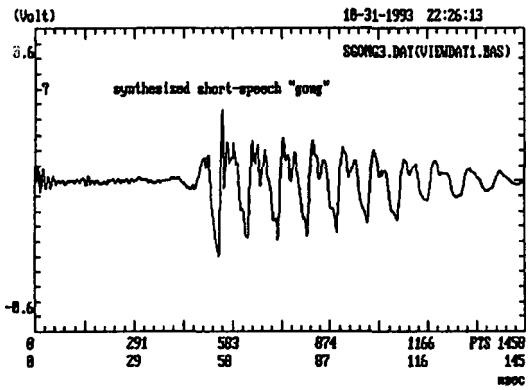


그림 5. CV형 단음 규칙합성 예
Fig 5. example of synthesized short-speech by rule

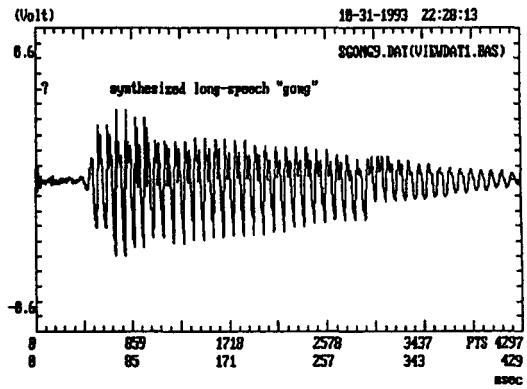


그림 6. CV형 장음 규칙합성음 예
fig Fig 6. example of synthesized long-speech by rule

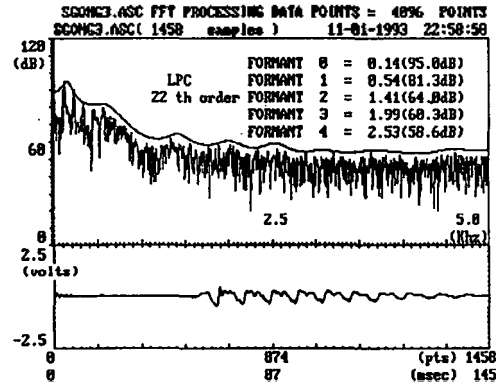


그림 7. 그림 5의 포르만트 추정도
Fig 7. estimated formant of fig 5.

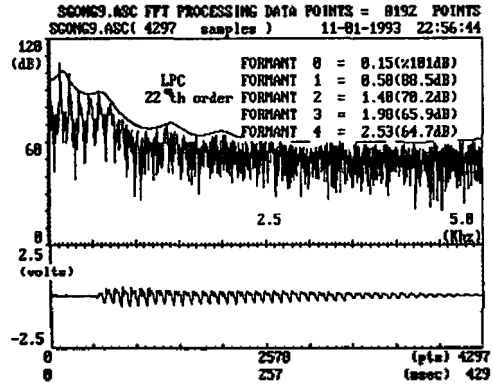


그림 8. 그림 6의 포르만트 추정도
Fig 8. estimated formant of fig 6.

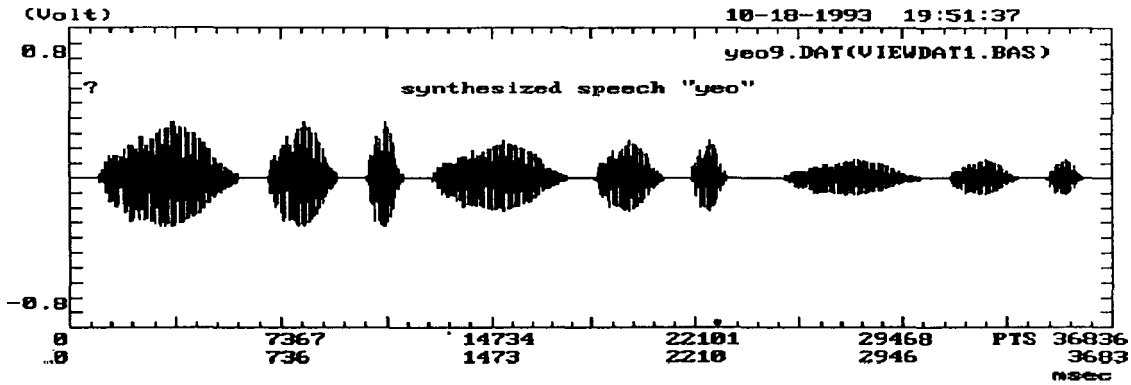


그림 9. 음절어에 대한 지속시간과 강약성분 제어 예(단음절어 : 여)
Fig 9. examples of synthesized speech "yeo for the control of duration and stress"

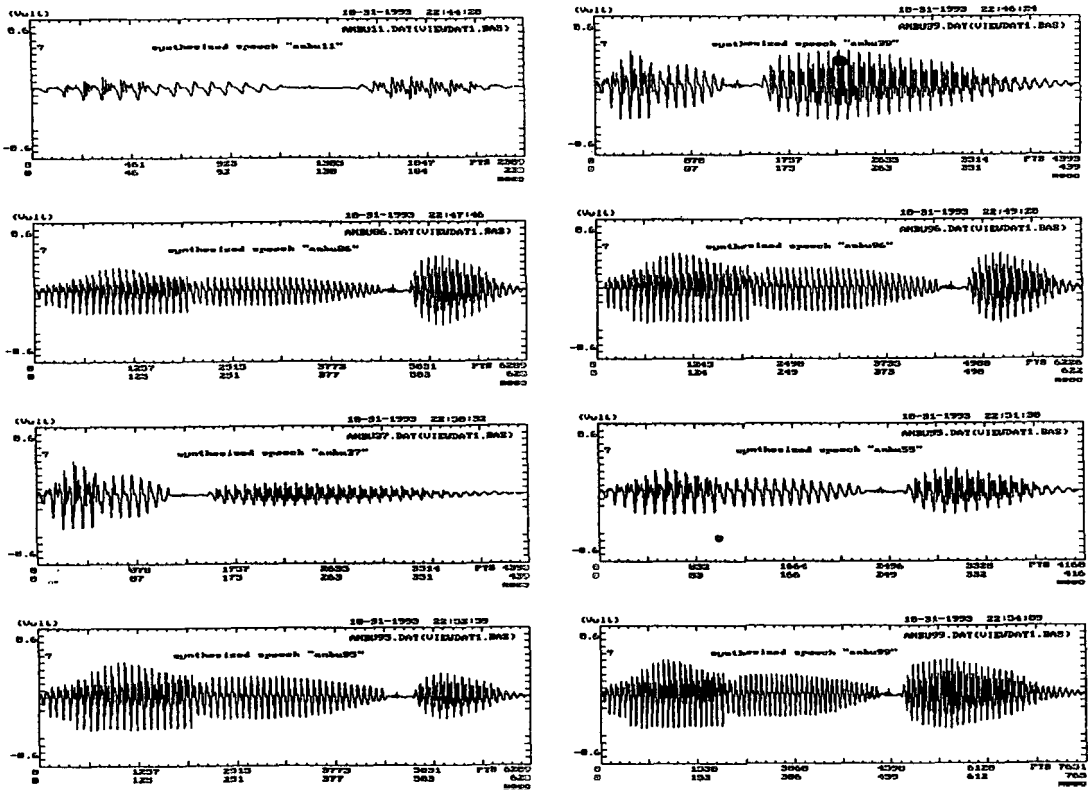
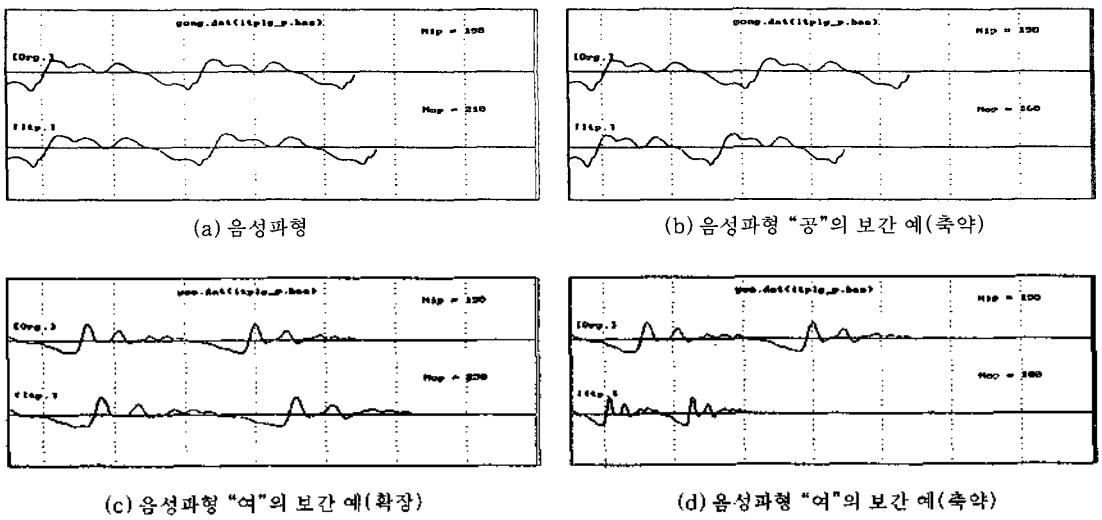


그림 10. 2음절어(안부)에 대한 지속시간과 강약성분 제어 예
Fig 10. examples of synthesized speech "anbu" for the control of duration and stress



(a) 음성파형 (b) 음성파형 "공"의 보간 예(축약)

(c) 음성파형 "여"의 보간 예(확장) (d) 음성파형 "여"의 보간 예(축약)

그림 11. Lagrange 보간법에 의한 피치변경 예(축약과 확장)
Fig 11. example of pitch control by the Lagrange interpolation(shrinking and expansion)

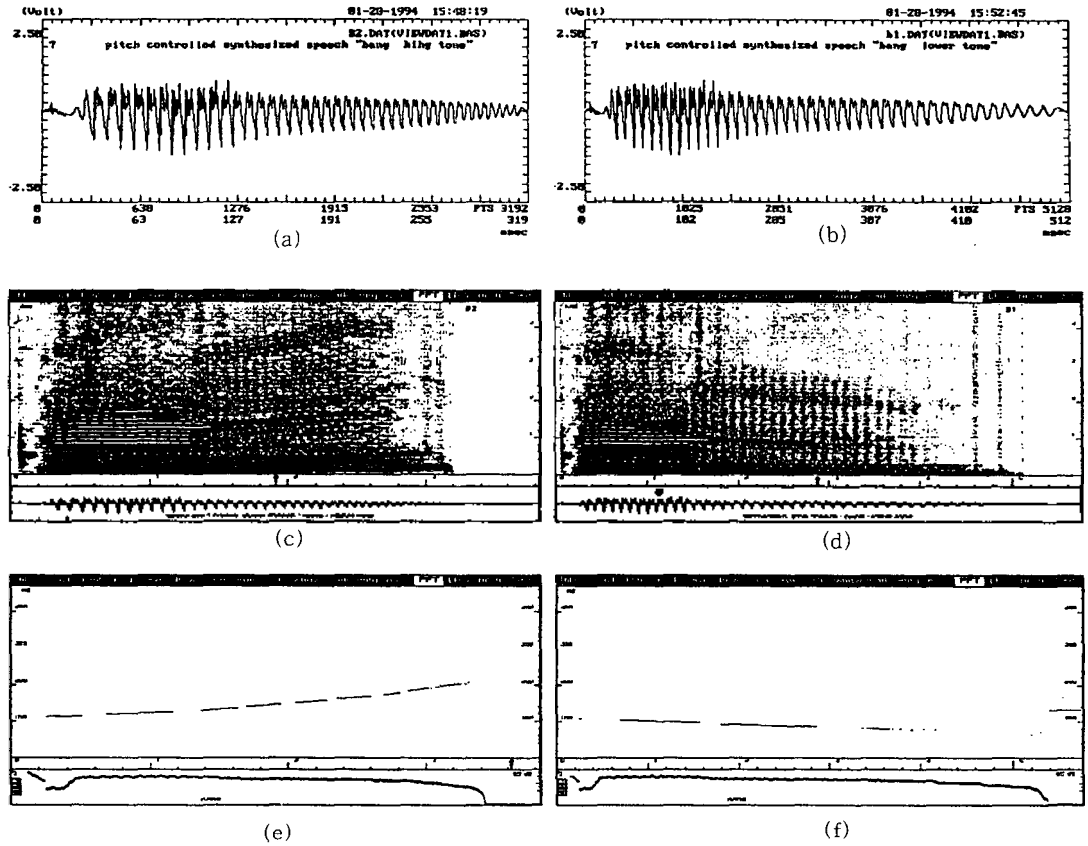


그림 12. Lagrange 보간법에 의한 피치제어 예(단음절어 : 방-고음, 저음)

Fig 12. example of pitch control by the Lagrange interpolation(syllable : bang-higher and lower tone)

NC는 자음소의 데이타 갯수를 표시한 것이며, NTOTAL은 전체 데이타 갯수를 표시한 것이다. MAX는 1음절 데이타중 최대진폭을 나타낸 것으로써 이를 이용하여 강약음의 정도를 규칙합성시키기 위한 파라메타로 사용하였다. 3장의 식(2)에 표시된 우측항의 A_m은 표2에서의 AMP RATIO항에 표시된 값을 나타내며, N_{ps}는 1PERIOD항에 표시된 1피치주기열내의 데이타 갯수를 의미하며, N_p는 1음절어내의 단위피치주기 갯수를 표시한 것이다. 표2의 데이타 포맷은 표1에 제시된 파형분석결과로 부터 구한 값이다. 이와같이 파형분석과정과 포멧사전이 구축된 후에는 규칙합성단계를 거쳐 합성음을 생성하게 된다. 그림 5와 6은 이와같이 합성된 파형을 나타낸 것이며, 그림 7과 8은 합성음의 포르만트 성분을 추정비교한 것이다. 그림 9은 강세순으로 단음은 1부터 3까지, 정상음은 4 부터 6까지, 장음은 7 부터 9까지 9개 등급으로 분류하여 규칙합성음을 발생시킨 예이며, 그림 10은 2 음절어 "안부"를 여러가지 형태로 규칙합성시킨 결과를 나타낸 것이며, 그림 11과 그림 12는 Lagrange 보간법을 사용하여 피치주기를 제어시킨 결과이다.

Ⅳ. 합성음 평가

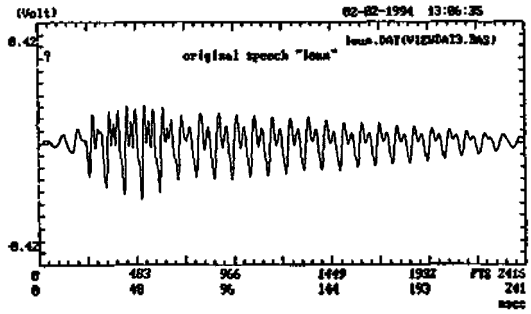
1. 객관적 평가

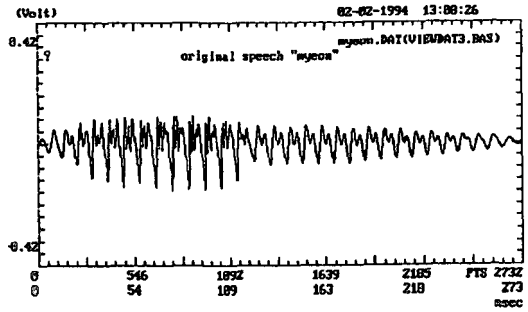
규칙합성음에 대한 음질평가는 객관적인 평가와 주관적인 평가를 병행하여 실시하였다. 객관적인 평가로는 첫번째로 음절 유형별로 단음절어를 2개씩 무작위로 2개씩 선정(V형 : 외, 우, CV형 : 초, 표, VC형 : 입, 악, CVC형 : 름, 면)하여 단음과 장음을 규칙합성

시킨 후 원음과 합성음에 대한 파형을 비교하였으며, 두번째로 이들을 각각 도입부, 정상부, 쇠퇴부등 3부분으로 분할하여 FFT시킨 후 스펙트럼 포락선과 포르만트 성분을 추정하여 원음과 합성음에 대한 스펙트럼 포락선과 포르만트 성분을 비교 제시하였다.

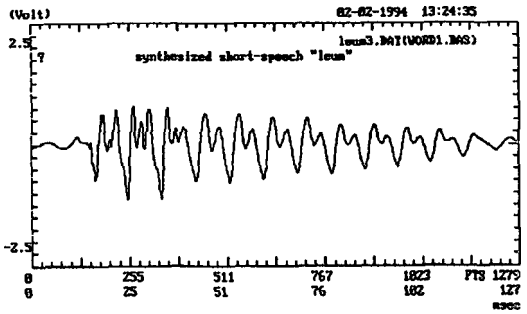
1.1 원음과 규칙합성음 파형 비교

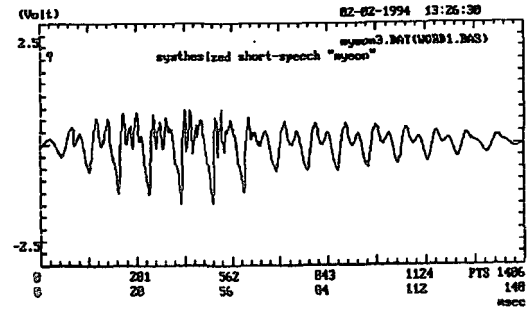
1) 원음 파형(름, 면)



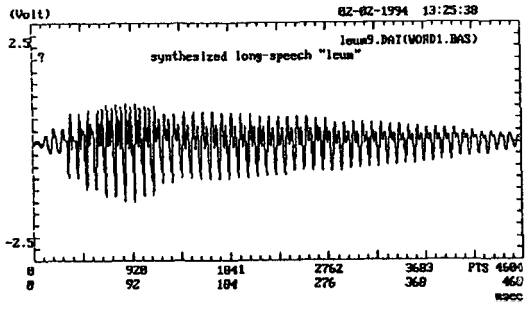


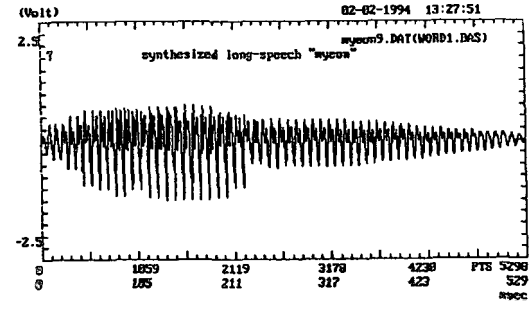
2) 합성음 파형(단음 : 름, 면)





3) 합성음 파형(장음 : 름, 면)

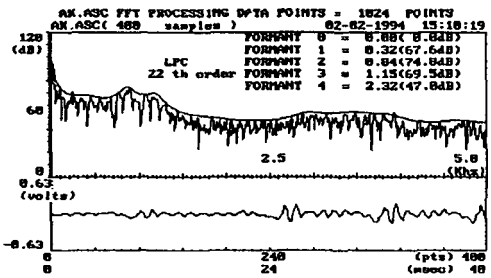




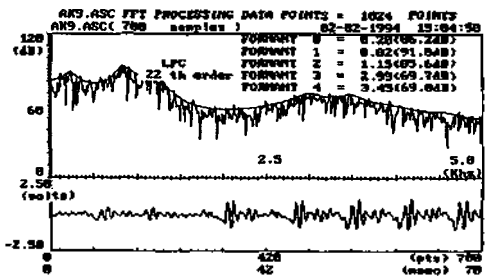
1.2 원음과 합성음의 스펙트럼 포락선 및 포르만트 성분 비교

1) 도입부

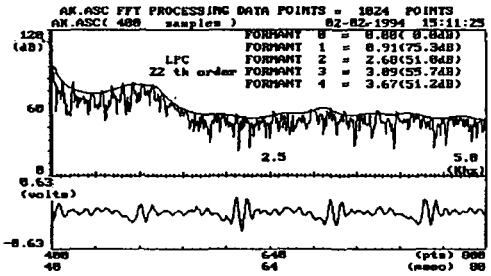
VC형 (원음:악)

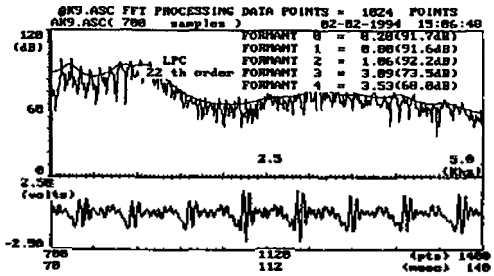


VC형(합성음 장음:악)

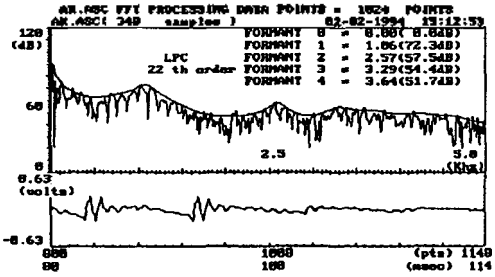


2) 정상부





3) 쇠퇴부



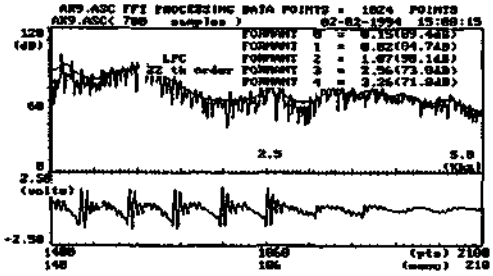


표 3. 데이타 포맷 사전에 등록된 음절표
Table 3. syllable table listed in the data format

음절 유형	데이타 포맷 작성 음절(빈도수 순위 : 한국어 발음대사전)															음절 갯수
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
V형	이	유	우	여	오	어	위	아	의	외	야	요	에	으	어	19 음절
	애	워	와	예												
CV형	다	리	기	사	지	대	자	가	하	수	무	부	고	기	화	80 음절
	따	구	도	주	거	소	치	서	비	조	미	마	재	보	게	
	나	때	내	그	호	훼	개	저	회	교	모	해	타	바	노	
	껴	재	차	니	세	까	콰	쌰	카	매	체	배	러	라	싸	
	처	추	두	과	초	르	찌	끼	드	후	포	로	표	피	새	
VC형	일	원	일	연	안	음	안	용	영	업	열	임	약	역	육	30 음절
	입	운	웬	은	울	염	암	왕	인	원	엄	온	언	옥	알	
CVC형	정	동	상	장	전	물	공	경	문	선	산	신	금	선	방	100 음절
	관	학	생	중	명	간	단	적	물	감	실	식	한	분	국	
	행	강	반	진	생	질	판	면	민	종	통	당	절	심	근	
	발	뗍	천	군	만	설	형	꿜	청	큼	등	들	각	말	철	
	난	목	년	병	견	남	망	삼	북	창	짱	석	변	출	송	
	본	건	평	항	속	광	품	풍	황	환	현	랑	술	살	쩍	
	범	합	글	급	달	담	작	녘	색	쌍						

2. 주관적 평가

합성음에 대한 주관적인 평가방법으로는 합성음에 대하여 사전 지식이 없는 남자 1명과 여자 2명을 선정하여 평가표를 작성하여 MOS(Mean Opinion Score)방법으로 구하였다. 평가시 스피치워크 스테이션 ver.2.1로 합성한 합성음을 소형 녹음기(AIWA : TP-26)로 녹음한 60개 음절(표 4)을, V, CV, VC, CVC형 순으로 1개 음절씩 번갈아 20초내에 3회 연속 반복음을 청취하여 작성토록 하였다. 평가항목은 이해도, 명료도, 잡음감, 자연성등 4가지 항목으로써 5개의 등급으로 분류하여 등급별로 가산점을 부여하여 평균을 취하였다. 평가시 이해도란은 2개의 항목으로 분류하여 첫번째 항목으로 청취음을 기입하게 하여 이질음화되는 음절을 분석하였으며, 평가시 잘못 청취한 6개의 음절(표 5)은 4가지 평가항목에서 가산점을 영으로 부여하여 평가하였다. 평가등급 항목에서 이해도란은 "5) 이해하기 아주 수월하다. 4) 이해하기 쉬운 편이다. 3) 보통이다. 2) 이해하기 어렵다. 1) 이해하기 아주 어렵다."로, 명료도는 "5) 음이 아주 명확하다. 4) 원음에 비하여 명확성이 조금 떨어진다. 3) 보통이다. 2) 나쁘다. 1) 아주 나쁘다."로, 잡음감은 "5) 잡음이 전혀 없다. 4) 좋은 편이다. 3) 보통이다. 2) 나쁘다. 1) 아주 나쁘다."로, 자연성은 "5) 아주 자연스럽다. 4) 자연스럽다. 3) 보통이다. 2) 어색하다. 1) 아주 어색하다."로 평가하였다. 그림 13와 그림 14은 표 4와 표 5를 그림으로 표시한 것이며 음절 유형별로는 V형이 4가지 유형중에서 가장 좋은 결과를 얻었으며, CV형과 CVC형인 경우에는

표 4. 음절 유형별 합성음 MOS 평가
Table 4. MOS test of synthesized speech for the korean syllable types

음절 유형	실험대상음절(음절유형별당 15음절 : 총 60음절)															음질평가			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	이해도	명료도	잡음감	자연성
V형	이	유	우	여	오	어	위	아	의	외	야	요	애	으	어	4.60	4.30	4.20	4.33
CV형	다	리	기	사	지	대	자	가	하	수	무	부	고	기	화	3.46	3.40	3.33	3.42
VC형	일	원	일	연	안	음	양	용	영	업	열	임	약	역	육	4.26	3.70	3.65	3.46
CVC형	정	동	상	장	전	물	공	경	문	선	산	신	금	성	방	3.66	3.30	3.26	3.41

표 5. 평가 항목별 MOS 산출표

Table 5. MOS evauation table according to the items

	이해도				명료도			잡음감			자연성		
	인식갯수	총점	평균	오인식율	갯수	총점	평균	갯수	총점	평균	갯수	총점	평균
수(5)	26	135	238/60 =3.96	6개 : 1. 사(자), 2. 상(장), 3. 산(잔), 4. 신(진), 5. 약(야), 6. 화(바)	15	75	3.60	5	25	3.61	11	55	3.65
우(4)	19	76			25	100		45	180		35	140	
미(3)	9	27			13	39		4	12		8	24	
양(2)	·	·			1	2		·	·			·	
총 계	54	238		10%	54	216		54	217		54	219	

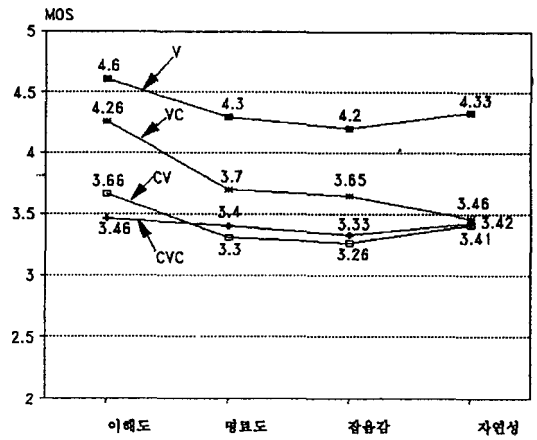


그림 13. 음절 유형별 합성음 MOS 평가도

Fig 13. MOS test plot of synthesized speech to the syllable types

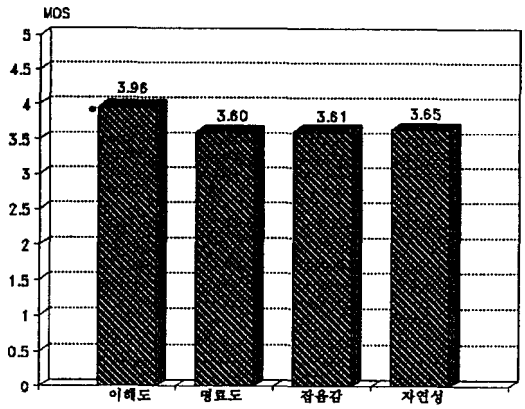


그림 14. 평가 항목별 MOS

Fig 14. MOS to the items of evaluation

잡음감과 명료성 항목에서 3.3내지 3.2정도로 평가 항목중 가장 낮은 평가를 받았다. 이질음화 되는 합성음도 6개로써 전체 음절 60개의 10%가 이질음화 현상을 보였으나, 대부분 "ㅅ"음을 "ㅈ"음으로 오인식(6개중 4개 : 사, 상, 산, 신)하였다(표5). 이질음화가 이루어지는 음절을 유형별로 살펴보면 CA형이 2음절("사"를 "자"로, "화"를 "바"로), CVC형이 3음절("상"을 "장"으로, "산"을 "잔"으로, "신"을 "진"으로), VC형이 1음절("약"를 "야"로) 존재하였다. 그림 14은 평가실험 음절 60개 전체에 대하여 이해도, 명료도, 잡음감, 자연성등을 평균한 값으로 음절 이해도는 3.96으로 비교적 합성음을 이해하는 데에는 어려움이 없으나, 명료도, 잡음감, 자연성등은 3.6 정도의 점수로 보통 수준의 결과로써 명료성, 잡음감, 자연성등은 좋은 편에 미달하는 수준으로 평가되었다.

V. 결 론

본 논문에서는 시간영역에서의 음성합성 알고리즘 개발에 관한 연구로써 이에 대한 타당성 검토로써 합성 음질과 자연성 측면에서 주로 고찰하였다. 특히 서론에서와 같이 시간영역에서의 합성방식 개발시 자연성에 영향을 미치는 운율요소의 효율적인 제어가 주요한 문제점으로 대두된다. LPC, PARCOR 혹은 LSP나 포르만트 합성기와 같은 주파수 영역에서의 합성방식은 성도의 극점을 구하여 필터로 재생시키는 방식으로써 소스와 성도 모델을 분리시켜 각각

을 추정한 후 합성에 이용하므로 소스와 성도모델 추정 오차로 인하여 단음절 단위의 합성음시 음질이 명료하지 못하나 문장단위 합성시에는 문장단위의 피치 패턴을 추정한 후 여기원의 임펄스 열의 간격을 조절함으로써 억양등과 같은 운율요소의 제어가 용이하다는 장점 때문에 음질의 열화에도 불구하고 시간영역에서의 합성방식 보다 자연성이 향상되어 주로 음성합성기에서는 이러한 방식들이 사용되고 있다. 따라서 본 논문에서는 이러한 시간영역 상에서의 문제점을 해결하기 위하여 음절단위의 음성파형을 입력시켜 단음절 단위로 파형을 분석하여 규칙합성용 매개변수를 추출하여 합성에 이용함으로써 지속시간, 강약 및 억양(피치주기)등과 같은 운율요소의 제어가 가능하였으며 음질도 매우 명료하였다. 주파수 영역에서의 합성 방식에서는 임펄스 열의 간격을 조절하여 피치를 제어시키나 본 방식에서는 파형 분석과정을 거쳐 1피치주기의 경계점을 구하여 추출한 단위주기 파형별로 보간을 시킴으로써 단음절 합성시 원하는 피치패턴의 제어가 가능하였다. 앞으로 문장단위의 합성시 자연성을 주파수영역에서의 합성방식 수준으로 향상시키기 위하여는 음절간 접속구간에서의 파형 윤곽선 제어 및 변이음 처리가 미흡하므로 이에 관한 연구가 이루어져야 할 것이다.

감사의 글

실험에 아낌없는 협조와 조언을 하여 주신 경희대학교 시스템공학연구소 강대수 박사, 이종현 박사, 권기형, 안정근, 서성태, 박상희 연구원에게 감사드립니다.

참 고 문 헌

1. 이현복, "현대 한국어의 악센트," 서울대학교 문리대학보 19권 합병호(통권28호), 1973.
2. 성철재, "표준 한국어 악센트의 실험 음성적 연구-청취테스트 및 음향분석," 서울대학교 대학원 석사학위 논문, 1991.
3. 이현복, 한국어의 표준발음, 교육과학사, 1989.
4. Jonathan Allen, M. Sharon Hunnicutt and Dennis Klatt, From Text to Speech : The MITalk system, Cambridge Univ. Press, 1987.
5. Shuzo Saito, Fundamentals of Speech Signal Processing, Academic Press, 1981.
6. 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.
7. 中津 良平, "音聲認識·合成技術の製品化および 市場動向," SP89-103, 1989.
8. 한국방송공사, 표준한국어 발음 대서전, 어문각, 1993.
9. Nobuhiko Kitawaki, Hiromi Nagabuchi, "Quality Assessment of Speech Goding and Speech Synthesis System," IEEE Comm., 1988. Vo.126, No.10.
10. 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.
11. 조철우, 김경태, 이용주, "무의미단어에 의한 규칙합성음의 평가 및 진단법에 관하여," 음성통신 및 신호처리 워크샵 논문집, 1993.8.
12. 김정환, 강성훈, "음성품질 주관법의 표준화에 관한 고찰," 전자통신 동향분석, 1990.7.
13. 강찬희, 진용옥, "한국어 문어변환 시스템 내에서의 음성합성기 개발," 한국음향학회 논문지, 1993.2.

▲강 찬 희(Chan Hee Kang) 1958년 3월 6일생

1980년 2월 : 경희대학교 전자공학과 졸업
1982년 2월 : 경희대학교 대학원 전자공학과 졸업
1983년 7월~1985년 7월 : 해군사관학교 교수부 전자과
1985년 8월~1986년 8월 : 삼성통신연구소
1989년 8월 : 경희대학교 대학원 박사과정 수료
1989년 3월~현재 : 삼지대학교 병설 전문대학 전자과

▲진 용 옥(Yong Ohk Chin) 1943년 3월 21일생

1968년 2월 : 연세대학교 공과대학 전기공학과 졸업
1975년 2월 : 연세대학교 대학원 전자공학과 졸업
1981년 8월 : 연세대학교 대학원 전자공학과(공학박사)
1980년 : 통신기술사
1976년~현재 : 경희대학교 공과대학 전자공학과 교수
현재 : 한국음향학회 회장