特輯

음성 합성 기술 분야

李 潤 根, 安 承 權
金星中央研究所 基礎 4 研究室

Ⅰ. 서론

현대 정보사회에서 인간은 많은 정보들을 다양한 매체를 통하여 받아들인다. 이와같은 정보수집을 보다 편리하고 신속하게 하기 위하여 다양한 매체의 통합인 멀티 미디어 환경과 인간과 기계의 정보교환을 편리하게 하기 위한 Man-Machine Interface 기술이 중요하게 대두되고 있다.

음성은 인간의 자연스러운 의사소통의 도구이므로 음성을 통한 기계와의 정보교환 기술은 매우 중요하다. 음성 처리 기술로는 인간의 음성을 기계가 인식할 수 있게 하는 음성인식 분야, 기계가 인간의 음성으로 말을 할 수 있게 하는 음성합성 분야, 음성을 저장및 송수신할 수 있게 하는 음성 코딩 및 통신 분야등이 있는데 본 기고에서는 이들중 음성합성 분야의 개요 및 국내외 연구동향을 살펴보도록 한다.

인간의 말은 성대의 울림에 의한 공기의 진동이 성도를 통해 입밖으로 나옴으로써 생성된다. 말은 정보를 가지고 있으며 정보는 말의 변별적 특성에 의해 표현된다. 이 변별적 특성은 입과 혀의 모양 및 위치에 따라 성도의 모양이 달라짐으로써 성대의 울림에서 발생한 공기 진동이 특정 주파수에서 공진을 일으킴으로써 발생하거나 또는 공기의 흐름을 막거나 틔움으로써 발생한다. 음성합성이란 이와같은 인간의 발성 기관을 인공적으로 만듬으로써 인위적 음성을 만들어 내는 것이다.

역사상으로 최초의 음성합성기는 1779년에 만들어졌다고 전해지며 그림1은 1791년에 von Kempelen에 의해 만들어진 기계식 음성합성기의 구조이다. 최초의 전기적 음성 합성기는 1922년에 J.Q. Stewart에 의해 만들어졌는데 전기적 공진회로에 의해서

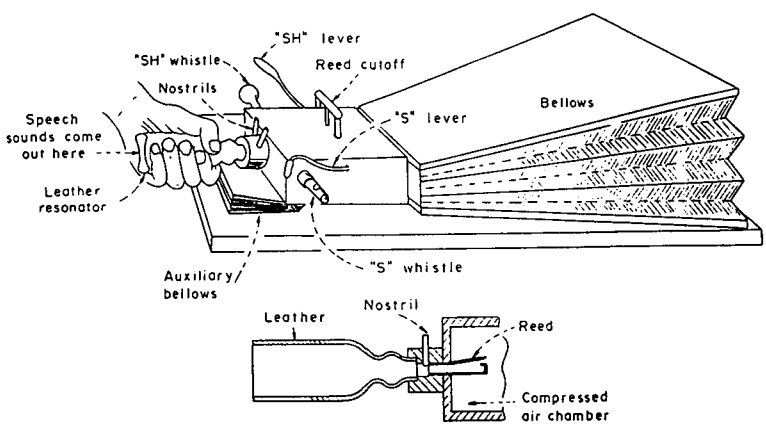


그림 1. Von Kempelen에 의해 만들어진 기계식 음성합성기

모음을 생성해 냈다. 최초의 연속음성을 합성할 수 있는 음성합성기는 voder라 불리는 합성기로써 1939년에 H. Dudley에 의해 개발되었는데 발판 및 10개의 건반을 이용해 기본 주파수 및 공진 필터의 특성을 제어하여 음성을 합성하도록 되어있다.

1960년 Fant에 의해 음성생성의 음향학적 원리 및 음성 생성의 디지탈 모델이 발표되고 최근들어 디지탈 신호처리 기술 및 컴퓨터의 발달로 인하여 본격적인 음성합성의 연구가 진행되고 있다. 음성합성은 합성 대상 어휘 및 합성 방법에 따라 여러가지 분류가 가능하므로 다음장부터 그 내용에 대해 상술토록 한다.

Ⅱ. 음성 합성 기술의 분류

음성 합성은 합성 대상 어휘에 따라 제한 어휘 합성과 무제한 어휘 합성으로 분류되며 합성 방법에 따라 파형 코딩법과 음원 코딩법, 혼합 방법등으로 분류할 수 있다. 그림 2는 음성 합성 방법의 분류도이다.

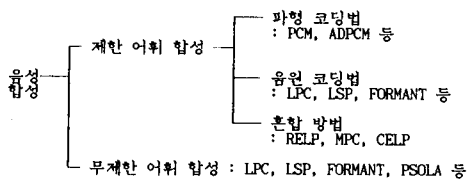


그림 2. 음성 합성 방법의 분류

제한 어휘 합성은 합성하고자 하는 어휘들을 미리 분석 하였다가 이들의 조합에 의해 말을 합성하는 방법으로써 합성 대상 어휘가 제한된다. 주로 단어 또는 소문장 단위의 음편들을 연결하여 말을 합성하는데 현재 지하철 안내방송, ARS(Audio Response System)등에 이용되고 있다. 구현이 용이하며 무제한 어휘합성에 비하여 높은 음질을 얻을 수 있으나 음편들의 연결 부분이 부자연스러우며 합성 대상 어휘가 바뀔 때마다 다시 녹음, 분석하여야 하는 단점이 있다. 반면에 무제한 어휘 합성은 언어의 기본 단위인 음소, 음절등의 조합에 의해 말을 합성해 내므로 합성 대상 어휘에 제한이 없으며 주로 TTS (Text-to-Speech) 장치 및 CTS(Context-to-Speech) 장치 등에 적용된다. 그러나 음소, 음절등의 연결시 상호 조음현상의 처리 및 자연스러운 운율처리등이 아직 미흡하여 현재까지는 제한 어휘 합성 방법에 비하여 음질이 떨어지는 실정이다.

합성 방법에 의한 분류에 있어서 파형 코딩법은 구현이 용이하고 음질이 좋은 반면 저장해야 할 데이타 양이 많아 제한 어휘 합성기에 주로 쓰인다. 음원 코딩법은 인간의 성도 특성을 모델링하여 특징 파라메터의 시간적 변화 정보에 의해 음성을 합성한다. 파형 코딩법에 비해 연산량이 많고 음질도 떨어지나 데이타 압축률이 높고, 특징 파라메터의 변환에 따라 말의 속도, 음높이, 스펙트럼 변환 등이 용이하여 주로 무제한 어휘 합성에 응용된다.

1. 파형 코딩법에 의한 합성

파형 코딩법은 음성 신호를 시간축 상에서 표본화하여 코딩한 후 저장하였다가 합성시 디코딩하여 음성신호를 재생하는 방법이다. 보통 음성신호는 4~6kHz내에 주요 정보가 거의 다 포함되므로 표본화 주파수는 Nyquist 법칙에 의해 8~12kHz가 된다. 이를 일정한 bit으로 양자화 하여 저장하는데, 예를 들어 8kHz 표본화된 음성 신호를 8bit로 양자화 할 경우 64kbps의 데이타가 필요하다. 그러나 음성신호 표본간의 상관성(correlation) 제거 및 적응 양자화 기법등을 이용해 데이타 압축이 가능하다. 대표적인 데이타 코딩 기법으로는 PCM, ADPCM, ADM 등

표 1. 파형 쿠딩법에 의한 음성합성 chip

품명	방식	DataRate (Kbps)	표본화 주파수(KHz)	최대 단어	내장ROM (bit)	내장 DAC	제조회사
MSM5202	ADPCM	12-32	4/6/8	-	없음	10bit	OKI
MSM0218	ADPCM	12-32	4/6/8	-	없음	10bit	OKI
MSM6202	ADPCM	7-24.6		124	144k	10bit	OKI
MSM6212	ADPCM	7-24.6	5.5/8.2	124	288k	10bit	OKI
MSM5248	ADPCM	16.5/21.6	5.5/8.2	7	18k	10bit	OKI
MSM6243	ADPCM	10-32.8	5.5/8.2	124	192k	10bit	OKI
μPD7751C	ADPCM	14-24	4/5/6		없음	없음	NEC
T6931	ADM	5.5-16	-	61	64k	10bit	TOSHIBA
LR3681	파형소편	1.6-3.2	-	256	32k	8bit	SHARP
μPD1771	파형소편	불명			506x16	8bit	NEC
MC3417	CVSB	24-32					MOTOROLA
μPD7730	ADPCM	24-32					NEC
TC8830F	ADM	8-32		16	없음		TOSHIBA
TC8801F	ADM	11-32		64	256k		TOSHIBA
μPD7755	ADPCM, PCM		4/5/6/8		96k		NEC
μPD7756	ADPCM, PCM		4/5/6/8		256k		NEC
μPD7757	ADPCM, PCM		4/5/6/8		512k		NEC

이 있으며 대략 24kbps~64kbps의 정보량을 갖는다. 파형 코딩법에 의한 음성합성은 단일 chip으로 개발되어 많은 제품에 응용되고 있는데 표1에는 음성합성 chip의 품명과 주요 사양이 나타나있다.

2. 음원 코딩법에 의한 합성

음원 코딩법은 음성 발성 기관을 모델링하여 특징 파라메터에 의해 음성을 합성해 내는 방법이다. 대표적인 합성 방법으로는 포만트 합성법, LPC 합성법등이 있다.

1) 포만트 합성법

성도는 여러가지 단면적을 갖는 관의 연결로 모델링할 수 있는데, 각각의 관의 단면적에 따라 고유의 공진 주파수를 가지며 이는 입모양, 혀의 위치 등에 따라 변하므로 이에 따라 발음의 변별적 특성이 생긴다. 성도의 공진 주파수를 포만트(formant)라 하며 3 ~ 5개의 포만트로써 성도 특성을 나타낼 수 있다. 각각의 공진 특성은 2차 디지탈 공진회로로 구현가능하며 이를 직렬 또는 병렬로 연결하여 성도를 모델링한다. 공진회로의 구성 및 주파수 특성은 그림3에 나타나있다. 그림3의 공진회로의 입출력 관련식은 식(1)과 같다

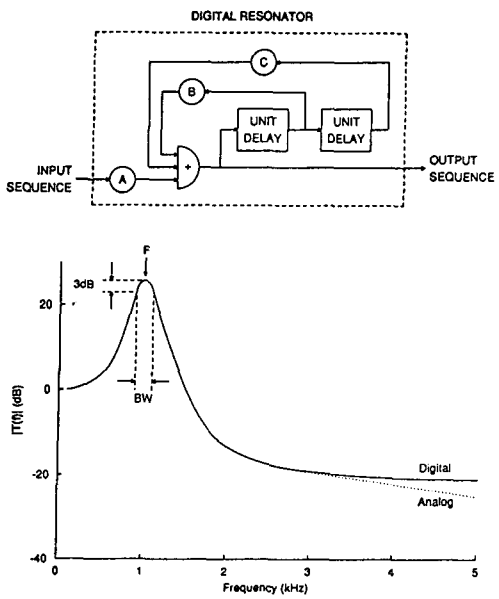


그림 3. 공진회로의 구성 및 주파수 특성

$$y(nT) = Ax(nT) + By(nT - T) + Cy(nT - 2T) \qquad (1)$$

$$C = -\exp(-2\pi BWT)$$

$$B = 2\exp(-\pi BWT)\cos(2\pi fT)$$

$$A = 1 - B - C$$

T:표본화 주기, f:공진주파수, BW:대역폭

그림 4는 포만트를 이용한 무제한 어휘 음성합성기의 하나인 MITalk의 구성을 나타낸다. 유성음을 합성하기 위한 5개의 직렬형 공진회로와 비음을 합성하기 위한 공진-반공진 쌍회로, 무성음을 합성하기 위한 병렬형 공진회로 등으로 성도 모델링을 하였다. 음원으로는 유성음의 경우 그림5와 같은 펄스열을, 무성음의 경우 백색 잡음을 이용하였다.

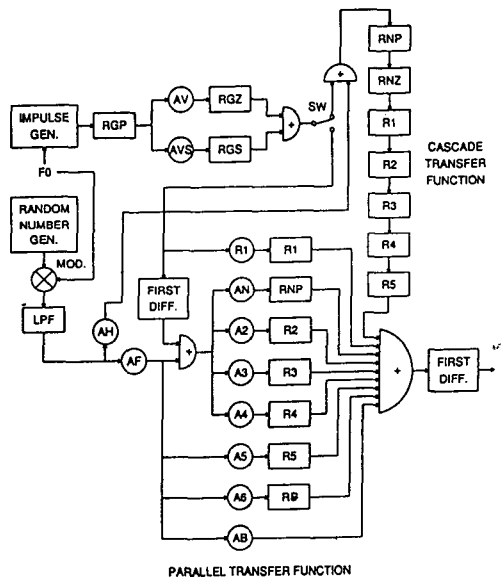


그림 4. 포만트 합성기의 구성

포만트 합성기는 음성의 기본 특성인 포만트 정보만으로 음성을 합성하므로 필요 데이타 양이 매우 적으며 파라메터 제어에 따른 음성 특성 및 운율 변화가 용이하나, 정확한 파라메터 추출이 어렵고, 무제한 어휘 합성시 다양한 조음현상에 따른 포만트의 변화를 정확히 규칙화 하기 위하여 방대한 음성신호의 분석이 필요하므로 구현이 어렵다. 표2에는 포만트 합성기에 필요한 제어 파라메터들과 이들의 대략적인

표 2. 포만트 합성기의 제어 파라메터

Symbol	Name	Min.	Max.	Typ.
F1	First formant frequency (Hz)	150	900	500
F2	Second formant frequency (Hz)	500	2500	1500
F3	Third Formant frequency (Hz)	1300	3500	2500
F4	Fourth formant frequency (Hz)	2500	4500	3300
F5	Fifth formant frequency (Hz)	3500	4900	3850
F6	Sixth formant frequency (Hz)	4000	4999	3850
FNP	Nasal pole frequency (Hz)	200	500	250
FNZ	Nasal zero frequency (Hz)	200	700	250
FGP	Glottal resonator 1 frequency (Hz)	0	600	0
FGZ	Glottal zero frequency (Hz)	0	5000	1500
B1	First formant bandwidth (Hz)	40	500	50
B2	Second formant bandwidth (Hz)	40	500	70
B3	Third formant bandwidth (Hz)	40	500	110
B4	Fourth formant bandwidth (Hz)	100	500	250
B5	Fifth formant bandwidth (Hz)	150	700	200
B6	Sixth formant bandwidth (Hz)	200	2000	1000
BNP	Nasal pole bandwidth (Hz)	50	500	100
BNZ	Nasal zero bandwidth (Hz)	50	500	100
BGP	Glottal resonator 1 bandwidth (Hz)	100	2000	100
BGS	Glottal resonator 2 bandwidth (Hz)	100	1000	200
BGZ	Glottal zero bandwidth (Hz)	100	9000	6000
A1	First formant amplitude (dB)	0	80	0
A2	Second formant amplitude (dB)	0	80	0
A3	Third formant amplitude (dB)	0	80	0
A4	Fourth formant amplitude (dB)	0	80	0
A5	Fifth formant amplitude (dB)	0	80	0
A6	Sixth formant amplitude (dB)	0	80	0
AN	Nasal formant amplitude (dB)	0	80	0
AB	Bypass path amplitude (dB)	0	80	0
AV	Amplitude of voicing (dB)	0	80	0
AF	Amplitude of friction (dB)	0	80	0
AH	Amplitude of aspiration (dB)	0	80	0
AVS	Amplitude of sinusoidal voicing (dB)	0	80	0
SW	Cascade/parallel switch	c	p	c
SR	Sampling rate (Hz)	5000	20000	10000
FO	Fundamental freq. of voicing (Hz)	0	500	0
NWS	Number of waveform samples per chunk	1	200	50
GO	Overall gain control (dB)	0	80	48
NFC	Number of cascaded formants	4	6	5

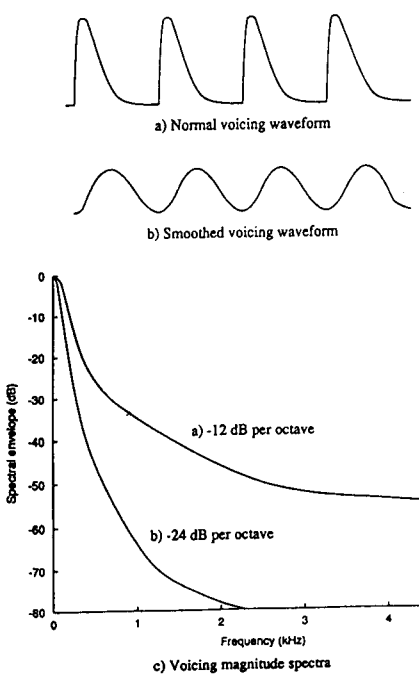


그림 5. 유성음의 음원 신호

값이 나타나있다.

2) 선형 예측(LPC) 합성법

음성신호는 표본간의 상관성이 높으므로 과거의 n개의 표본으로 현재의 표본값을 예측할 수 있다. 이 예측 계수를 이용하여 all-pole 성도모델 필터를 구성하고 음원 신호를 필터링 하면 음성 신호를 합성해 낼 수 있다. 음원 신호로는 유성음의 경우 펄스열을, 무성음의 경우 백색 잡음신호를 이용한다. 음성 신호를 all-pole 모델로 합성하므로 zero특성이 나타나는 비음의 처리가 미흡하다. 선형 예측 계수를 a(k), 이득 계수를 G라 하면 p차 선형 예측에 의한 성도 전달 함수 H(z)는 식(2)와 같으며 선형 예측 계수를 이용하여 음성 합성 회로를 구성하면 그림6과 같다.

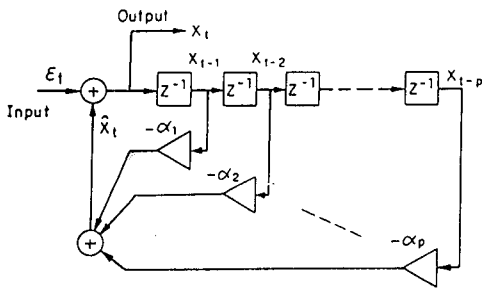


그림 6. LPC 합성 회로

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a(k) z^{-k}} \qquad (2)$$

선형 예측 계수를 구하는 방법으로는 autocorrelation법, covariance법, lattice법 등이 있으며 각각의 방법은 연산량, 저장 데이타량, 안정성 등에서 장단점을 갖는다. 실제 선형 예측 계수를 저장 또는 전송할 때에는 일정한 bit으로 양자화 하는데 양자화 오차에 의해 안정성이 보장되지 않으며, 이 경우 안정성을 보장할 수 있는 선형 예측 계수의 범위가 명확하지 않다. 이 문제를 해결하기 위하여 실제 구현시에는 PARCOR 계수를 많이 이용한다.

3) PARCOR(Partial Correlation) 합성법

PARCOR 계수는 전방향 예측 오차와 후방향 예측 오차의 상관계수로 정의된다. 전방향 예측 오차를 ef라하고 후방향 예측오차를 eb라하면 PARCOR계수 ki는 식(3)과 같이 정의된다.

$$ki = \frac{\sum_{m=0}^{n-1} ef^{(i-1)}(m)eb^{(i-1)}(m-1)}{\left\{\sum_{m=0}^{n-1}\left(ef^{(i-1)}(m)\right)^2 \sum_{m=0}^{n-1}\left(eb^{(i-1)}(m-1)\right)^2\right\}^{1/2}} \quad (3)$$

PARCOR 계수와 선형 예측 계수는 반복 연산을 통하여 상호간 변환이 가능하다. PARCOR 계수는 절대값이 항상 1보다 작으며 이경우 안정성이 보장되므로 실제 구현시 선형 예측 계수보다 많이 사용되며 대략 2.4kbps의 데이타 양을 갖는다. PARCOR 합성 회로는 그림7과 같이 격자형 필터로 구현된다.

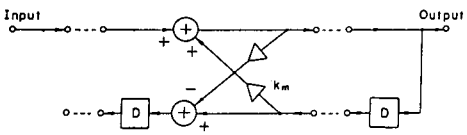


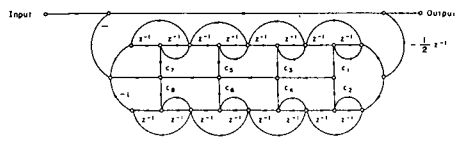
그림 7. PARCOR 합성회로

4) LSP(Line Spectrum Pair) 합성법

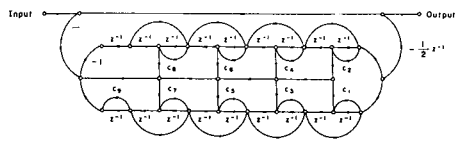
앞에서 설명한 PARCOR 합성법과 마찬가지로 성도를 all-pole 모델로 모델링한다. 성도의 음향 특성을 에너지 손실이 없는 무손실(lossless) 시스템으로 가정하면 공진 주파수에서의 Q값은 무한대가 되며 델타 함수쌍과 같은 공진 특성을 갖게 되는데 이를 LSP라 한다. LSP 계수는 PARCOR 계수에 비하여 낮은 bit rate에서도 좋은 음질을 얻을 수 있는 것으로 알려져있다 (Itakura, 1975). 이는 PARCOR 계수는 시간영역에서 작용하는 계수인데 반해 LSP 계수는 주파수 영역에서 작용하므로 양자화 또는 선형 보간법에 의한 왜곡이 비교적 적기 때문이다. 또한 LSP 계수의 값이 포만트와 유사하므로 파라메터 변환에 따른 음성 특성 제어가 용이하다. 반면 계산량이 약간 증가하며 분석 차수에 따라 LSP 계수값이 달라지는 단점이 있다. LSP 합성 회로는 그림8과 같이 구현된다.

3. 혼합 코딩법에 의한 합성

파형 코딩법과 음원 코딩법의 장점을 혼합하여 낮은 정보량으로 고음질의 음성을 합성해 내는 방법들이 고안되었다. 주로 선형 예측 방법의 변형으로써 음성 코딩 및 제한 어휘 합성에 많이 쓰이며, 피치 변화등 운율 제어가 어려우므로 무제한 어휘 합성에



(a) p=even:



(b) p=odd

그림 8. LSP합성회로

는 적용하기에 한계가 있다. 그밖의 음성합성 방법으로는 고음질 무제한 어휘 합성에 많이 쓰이는 PSOLA(Pitch Synchronous Overlap-and-Add) 방법과 대칭형 음편 합성법등이 있다.

1) RELP(Residual Excited LPC) 합성법

선형예측 분석에 의한 잔차신호의 저주파 성분을 부호화 하였다가 이를 선형예측 합성시 여기신호로 사용하여 음성을 합성한다.

2) MPC(Multipulse-Excited LPC) 합성법

선형예측 합성시 여기신호로써 다중 펄스를 이용한다. 이방법은 음성신호의 무성, 유성음 판별이 필요없으며 피치 검출과정도 필요 없기 때문에 음원신호의 모델링 오차에 의해 생기는 음성 품질 저하가 적다는 강점이 있다.

3) CELP(Code-Excited LPC) 합성법

선형 예측 합성시 여기신호로써 벡터 양자화된 여기신호를 이용한다. 여기신호는 음성신호의 주기성을 고려한 장기 예측과 표본간의 상관성을 고려한 단기 예측에 의하여 생성된다. 이 방법은 MPC 방법의 변형으로써 다중 펄스에 의한 여기를 벡터 양자화된 펄스열에 의한 여기로 대체한 것이다.

4) PSOLA(Pitch Synchronous Overlap-and-Add) 합성법

음성신호를 파형 부호화 방식으로 합성할 경우 무제한 어휘 합성시 가장 큰 문제중의 하나는 운율 제어의 어려움이다. 즉 피치의 변화 및 음 길이의 변화 등을 제어하기 어렵다. 음성 신호의 대부분을 차지하는 유성음은 주기적 성질을 가지며 피치 주기를 제어

함으로써 운율 제어가 가능하다. 그림9는 PSOLA 방법에 의해 유성음의 피치 주기를 변화시키는 설명도이다. 음성신호의 각 피치 단위의 신호를 피크값을 중심으로 해닝창을 이용해 분류하여 음편을 만든 후 합성시 각각의 음편을 중첩시켜 연속된 음성신호를 만드는데, 중첩 길이를 변화시킴으로써 피치 주기를 조절할 수 있다. 음성 신호를 시간 영역에서 부호화하므로 음원 코딩법에 의한 합성음에 비하여 음질이 좋고 합성시 연산량이 많지 않아 실시간 처리가 용이하다. 반면 합성시 필요한 데이타량이 많으며 음편을 자동적으로 분류해내기 어려우므로 음편 사전 구성시 수작업이 많이 필요하다. 주로 고음질의 무제한 어휘 합성기에 응용된다.

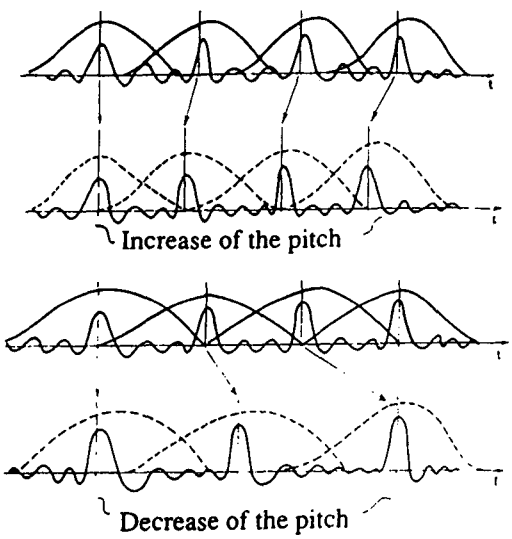


그림 9. PSOLA 합성법에서의 피치 주기 조절 방법

5) 대칭형 음편 합성법

음성 신호를 일정 길이의 프레임으로 나누어 이를 주파수 영역으로 변환하면 주파수 포락선 정보와 위상 정보가 얻어진다. 이중 위상 정보를 모두 영으로 치환하여 시간영역으로 역변환하면 좌우 대칭인 음편이 얻어지는데 이를 중첩하여 연결함으로써 음성을 합성한다. 음성 합성 과정은 2-3-4절에서 설명한 PSOLA 합성법과 같으나 PSOLA 합성법과는 달리 자동적인 음편 추출이 용이하며 음편이 좌우 대칭형이므로 음편의 반만 저장하면 전체를 복원할 수 있어서 합성시 필요한 데이타 저장량이 줄어든다는 장점이 있다. 반면 영위상 복원에 의해 단조로운 소리가 발생하며 중심부에서 비정상적 피크가 생성되어 음질이 저하된다. 이를 없애기 위해 음편을 비선형 변환을 한다. 그림10은 음편 추출 과정 및 음편 연결 과정을 나타낸다. PSOLA 합성법과 마찬가지로 고음질의 무제한 어휘 합성기에 응용된다.

Ⅲ. 문자 음성 변환 장치(Text-to-Speech System)

TTS 시스템은 임의의 문자열을 음성으로 변환하는 장치이다. 이는 무제한 어휘 합성 기능외에 문자-음

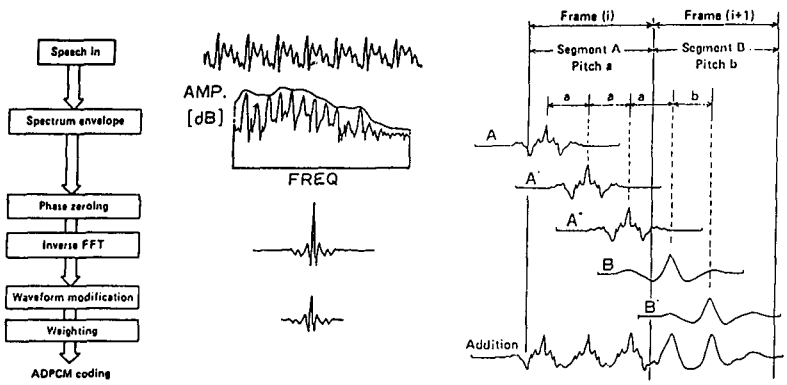


그림 10. 음편 추출 과정 및 음편 연결 과정

소 변환 기능, 구문 파싱 기능등 여러가지 기술의 복합에 의해 가능하며 음성 합성 분야의 궁극적 목표이다. 따라서 본 절에서는 음성합성 분야중 TTS 시스템에 대하여 자세히 설명하기로 한다.

1. 전체 구성

그림11은 TTS 시스템의 전체 구성도이다. 문자열이 입력되면 전처리 및 구문 파싱부에서 전체 문장의 구성을 해석한다. 이는 후의 운율 처리부에 정보를 제공한다. 문자-음소 변환부에서는 문자열을 음소열로 변환하는데 이 과정에서 음운 규칙 및 불규칙 처리용 사전이 필요하다. 운율 제어부에서는 음성 발성시 필요한 운율 파라메터를 추출하는데 억양, 강세, 리듬 등의 제어에 필요한 피치 포락선, 음소길이, 음량, 휴지기 길이 등의 파라메터를 추출한다. 합성부에서는 전단에서 추출한 운율 파라메터 및 음성 파라메터를 이용해 2절에서 설명한 여러가지 합성 방법에 의해 음성을 합성해 낸다.

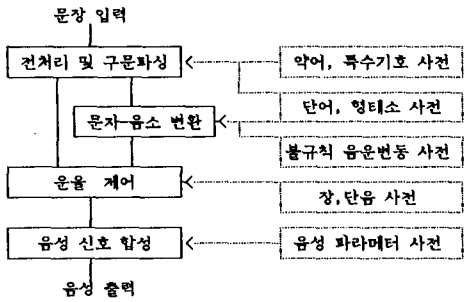


그림 11. TTS 시스템의 구성도

2. 합성 단위

합성 단위는 연속음을 합성해 내기 위한 기본 단위이다. 합성 단위로는 구절(phrase), 단어(word), 음절(syllable), 음소(phoneme)등의 단위와 이들을 음성 합성에 알맞는 형태로 변형시킨 반음절(demisyllable), diphone, triphone, CV (Consonant-Vowel), VC, VCV, CVC등 여러가지가 사용된다. 합성 단위는 클수록 자연도는 증가하나 조합하여 만들수 있는 대상은 제한되며 합성단위가 작은 경우는 그 반대이다. 주로 제한 어휘 합성시에는 구절 및 단어를 합성 단위로 많이 이용하며 무제한 어휘 합성시에는 음절 이하의 단위가 많이 이용된다. 합성 단위가 작으면 합성에 필요한 전체 데이타량이 작아지며 각각의 단위를 제어하기는 쉬우나, 합성 단위의 연결시 발생하는 상호 조음현상을 정확히 구현하기 어려우며 접합부분의 불연속성 및 이에 따른 음질 저하 현상이 생긴다. 따라서 이와같은 문제점을 최소화 하기 위하여 모음의 안정된 구간에서 음절을 절단한 반음절 및, CV, VC, VCV등이 합성 단위로 많이 쓰인다. 또한 최근에는 합성 단위를 한가지로 제한하지 않고 음소 환경에 따라 합성 단위가 자동적으로 생성되는 COC (Context-Oriented-Clustering) 방법이 제안되었다(Nakajima and Hamada, 1988). COC는 음소열을 음소환경이 같은 집단으로 분할하는 방법이다. 즉 음소 레이블링이 되어있는 음성 DB로부터 음소환경에 따른 분할 과정을 거쳐 합성 단위를 자동 생성하는데 이 과정이 그림12에 나타나있다. COC 방법에 의한 음성 합성 시스템의 블록도는 그림13과 같다.

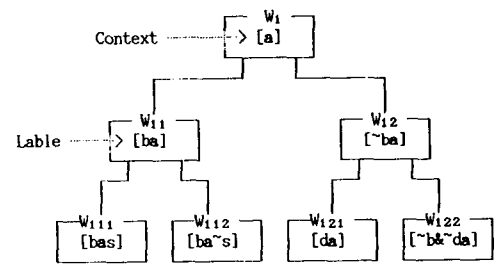


그림 12. Cluster분할 과정의 예

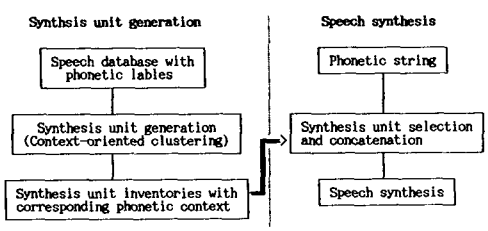


그림 13. COC합성 시스템의 블록도

3. 문자-음소 변환

표기되는 문자와 발음되는 음소는 서로 정확하게 일치하지는 않는다. 따라서 음성합성을 위하여 문자열을 음소열로 변환하는 과정이 필요하게 된다. 이 과정은 음운 변동 규칙에 의하여 대부분 자동적으로

변환이 가능한데 대표적인 규칙으로는 대표음화, 연음, 경음화, 격음화, 자음접변, 니은첨가, 구개음화 등이 있다. 그러나 경음화, 니은첨가등은 규칙으로 처리하지 못하는 부분이 존재하므로 이경우는 불규칙 변환 사전을 이용하여 처리하여야 한다.

4. 운율 제어

운율이란 발성시 나타나는 억양, 강세, 리듬등의 특성을 말하는데 이는 기본 주파수, 음소길이, 음량, 휴지기 길이 등에 의해 결정된다. 사람이 한번 숨을 쉬어 발성하는 말의 단위를 발화단위라 하는데 발화단위 내에서는 기본 주파수가 점차 낮아지는 경향을 갖는다. 이를 억양의 기본 기울기라 한다. 억양의 기본 기울기에 단어, 음절의 강세 및 문 구조에 따른 억양 패턴이 더해져서 전체 억양 패턴이 구성된다. 일본의 Fujisaki는 1960년대 후반에 선형 필터 모델을 이용하여 일본어의 억양 구조를 모델링하였다. 전체 구조는 그림14에 나타나 있다. 구 및 절의 위치에서 임펄스 입력을 발생시키는 구명령기와 단어의 고유 강세 패턴을 구형파로 생성하는 강세 명령기로 구성되며 각각은 선형 필터 시스템을 거쳐 더해짐으로써 전체의 억양 패턴이 생성된다. 음소길이 및 휴지기길이는 억양과 함께 합성음의 자연도를 결정하는 중요한 요소이다. 음소의 길이는 음소 자체의 성질뿐만아니라 주변의 음소환경, 한 단어내의 음소 갯수, 단어내에서 음소의 위치, 강세여부등 다양한 요소에 의해 영향을 받는다. 휴지기 길이도 음소길이와 마찬가지로 전후의 음소환경에 의해 영향을 받는데, 그 이외에 앞에서 언급한 발화단위사이에서 긴 휴지기가 존재한다. 그런데 발화단위는 하나의 발화단위 내의 어절갯수, 음절갯수 뿐만아니라 문장의 구조 및 의미에 의해 결정된다. 예를 들어 발화단위의 경계는 구, 절등의 경계에서 발생하기 쉬우나 수식어와 피수식어 사이에서는 발생하기 어렵다. 따라서 문장 구조의 분석을 위하여 형태소 분석, 구문 분석, 의미 분석등이 필요하며 운율 제어부에서는 이러한 정보를 이용하여 운율 파라메터를 추출한다.

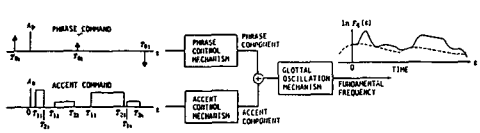


그림 14. Fujisaki 모델

5. 합성음의 성능 평가

합성음의 음질은 이해도와 자연도로 평가된다. 이해도는 다시 음소이해도, 단어이해도, 문장이해도 등으로 구분되는데 여러가지의 합성음 평가 방법을 표3에 보였다.

표 3. 합성음의 성능 평가 방법

이해도	o Diagnostic rhyme test (Voiers, 1983) o Modified rhyme test (House, 1965) o Harvard sentences (Egan, 1948) o Haskins anomalous sentences (Nye/Gaitenby, 1974) o Reading/listening comprehension (Pinsoni/Hunnicutt, 1980)
자연도	o Paired comparisons (IEEE, 1969 : Logan/Pisoni, 1986) o Subjective ratings (Nusbaum, 1984)

표 4. 국내외 연구 동향 및 관련 제품

—국내

연구기관	제품명/Type	합성방법	기타특징
삼성정보통신	"한국어 문서-음성 변환장치"/독립형 또는 PC 내장형	LSP	o 남성음성 o 상품화 단계 o μPD7720 사용
금성사	"Home PC"의 음성 합성보드/PC내장형	Formant	o 남성음성 o 상품화 단계 o Adlib호환 sound card 이용
Digicom	"가라사대"/ PC 내장형	LPC	o 남성음성 o 상품화 단계
ETRI	"금소리"/ PC 내장형	LSP	o 남성음성 o Prototype o TMS320C25 사용 o PSOLA 합성 연구
서울대	/독립형 또는 내장형	LPC	o 남성음성 o Prototype o TMS320C25

—국외

국가	연구기관	Model명	대상언어	합성단위	합성방법
미국	MIT	MITalk	영어	음소	Formant
	Bell Lab.			반음절	LPC
	Vortrax Inc.	TYPEN TALK	영어	음소	Formant
	Cornell 대학		일	음절	Formant
	Texas Instruments	Magic Wand	영어	음소	LPC
	Street Electronics Co.	ECHO-GP	영어	음절	LPC
	Ackeman Digital System	Synthetalker	영어	음소	Formant
	Intex Micro System	Intex Talker	영어	음소	Formant
일본	東京 大學		일어	가나음절	Formant
	朝倉 大學		일어	가나음절	Formant
	와세다 大學		일어	가나음절	PARCOR
	九州[illegible]術 工科 大學		일어	CV	LSP
	日本 電氣	PC6601	일어	CV, VC	Formant
	電氣 通信 研究所		일어		PARCOR
	東京 三洋 電氣	VSS100	일어	가나음절	PARCOR
	Fujitsu	FMVS-101	일어	CV, V	PARCOR
	OKI		일어	CV, VCV	Sym. Wave
서독	Ruhr	SYNTEX	독어	음소	Formant
프랑스	X-Com	DICTON	불어	음소	Formant

6. 국내외 연구동향

현재 미국, 일본 및 유럽 여러나라에서는 TTS 시스템이 실용화되어 있으며 국내에서도 학교, 연구소 및 몇몇 기업체에서 연구가 활발히 진행되고 있다. 국내외 연구 동향 및 관련 제품이 표4에 나타나있다.

Ⅳ. 결론 및 향후 전망

음성합성 분야중 파형 코딩방법에 의한 제한 어휘 합성은 현재 장난감에서부터 자동응답 시스템에 이르기까지 많은 부분에 적용되고 있다. 또한 집적 회로 기술 발전에 의해 여러종류의 음성합성 chip이 개발되었다. 무제한 어휘 합성 분야에 있어서는 선진국의 경우 실용화 단계에 와 있으며 국내에서도 최근들어 몇몇 업체에서는 TTS 응용 제품을 상품화 하기도 했다. 그러나 현재까지의 국내 기술 수준으로는 아직 합성음의 자연도면에서 미흡하므로 본격적인 상품화가 어려운 형편이다. 앞으로 무제한 어휘 합성분야는 언어학 및 자연어 처리 분야와의 연계를 통해 자연도를 향상시킬 수 있으리라 생각되며 수년내에 오디오 텍스 및 각종 정보 안내 시스템에의 적용이 예상된다.

參 考 文 獻

[1] L.R.Rabiner and R.W.Schafer, Digital processing of speech signals, Prentice-Hall,Inc.

[2] Jonathan Allen, M.Sharon Hunnicutt and Dennis Klatt, From text to speech: The MITalk system, Cambridge University Press.

[3] Sadaoki Furui, Digital speech processing, synthesis, and recognition, Marcel Dekker, Inc.

[4] Sadaoki Furui and M.Mohan Sondhi, Advances in speech signal processing, Marcel Dekker, Inc.

[5] J.D.Markel and A.H.Gray, Linear prediction of speech, Springer-Verlag, New York.

[6] Takashi Yazu and Kozo Yamada, "The speech synthesis system for an unlimited Japanese vocabulary", ICASSP 86, Tokyo, pp2019-2022, 1986.

[7] M.R.Schroeder and B.S.Atal, "Code-excited linear prediction(CELP): High-quality speech at very low bit rates", ICASSP-85, pp.937-940, 1985

[8] F.Soong and B.-H.Juang, "Line spectrum pair and speech data compression", *Proc.* ICASSP-84, pp.1.10.1-4, Mar.1984.

[9] K.Hakoda, S.Nakajima, T.Hirokawa and T.Mizuno, "A new Japanese text-to-speech synthesizer based on COC synthesis method", ICSLP 90,pp.809-812, 1990.

[10] S.Nakajima and H.Hamada, "Automatic generation of synthesis units based on context oriented clustering", ICASSP-88, pp.133-136, 1988.

[11] K.Ozawa and T.Araseki, "High quality multi-pulse speech coder with pitch prediction", ICASSP 86, pp.1689-1692, 1986

[12] C.Hamon, E.Moulines and F. Charpentier, "A diphone synthesis system based on time-domain modifications of speech", ICASSP-89, pp. 238-141, 1989.

[13] K.Hakoda, K.Kabeya and T.Hirahara, "Japanese text-to-speech synthesis based on residual excited speech synthesis", ICASSP-86, pp.2431-2434, 1986.

[14] D.H.Klatt, "Review of text-to-speech conversion for English", J.A.S.A., vol. 82, no.3, pp.737-797, Sept.1987.

[15] 이현복, 한국어의 표준발음, 교육과학사, 1989.

[16] 허웅, 국어 음운학, 샘문화사, 1985
[17] K.Hirose, H.Fujisaki and H.Kawai, "Generation of prosodic symbols for rule-synthesis of connected speech of Japanese", ICASSP-86, pp.2415-2418, 1986.
[18] H.Fujisaki and H.Kawai, "Realization of linguistic information in the voice fundamental frequency contour of the spoken Japanese", ICASSP-88, pp.663-666, 1988.
[19] H.Fujisaki and H.Sudo, "A model for the generation of fundamental frequency contours of Japanese word accent", *J.Acoust.Soc.Jpn.*, vol.27, no.9, pp. 445-453, 1971.

筆者紹介

李 潤 根
1964年 2月 18日生
1986年 2月 서울공대 제어계측공학과 졸업(공학학사)
1988年 2月 KAIST 전기 및 전자공학과 졸업(공학석사)

1988年 2月 ~ 현재 금성중앙연구소 기초 4연구실(선임연구원)

주관심분야: DSP, 음성신호처리

安 承 權
1957年 10月 20日生
1980年 2月 서울대학교 공과대학 전자공학과 (공학학사)
1982年 2月 서울대학교 공과대학 전자공학과 (공학석사)
1992年 7月 서울대학교 공과대학 전자공학과 (공학박사)

1988年 2月 ~ 현재 금성중앙연구소 기초 4연구실 책임연구원

주관심분야: 디지탈 시호처리, 신경회로망, 패턴인식