음소단위를 이용한 소규모 문자-음성 변환 시스템의 설계 및 구현

Design and Implementation of Simple Text-to-Speech System using Phoneme Units

박 애 희*, 양 진 우*, 김 순 협*
(Ae-Hee Park*, Jin-Woo Yang*, Soon-Hyob Kim*)

※본 연구는 광운대학교 신기술연구소의 연구비 지원에의해 이루어진 것임

요 약

본 논문은 소규모 시스템에 적용 가능한 한국어 문자-음성 변환 시스템의 설계 및 구현에 대한 연구를 목적으로 한다. 본 논문에서 채택한 음성합성 방법은 파라메터 합성법으로서 LPC(Linear Predictive Coding)계열의 PARCOR(PARtial autoCORrelation)계수를 음향 파라메터로 사용하였으며, 음성합성 단위로는 가장 기본적인 단위인 음소를 채택하였다. 합성 파라메터로는 유성음의 경우 PARCOR계수, 피치, 진폭을 무성음의 경우 잔차신호와 PARCOR계수를 사용하였다. 특히 무성음의 경우 LPC합성시 음질이 떨어진다는 단점이 있었으나, 본 논문에서는 LPC분석시 얻어지는 잔차신호를 무성음의 여기신호로 사용하여 단어 단위의 합성에서 60%의 이해도를 얻을 수 있었다. 합성결과 단어 단위의 합성에 적용 가능하였고, 문장단위의 합성을 위해서는 음소 지속시간 조절에 대한 연구가 진행되어야 할 것이다.

본 논문의 구현환경으로는 486 PC상에서 음성의 입,출력을 위해 70[Hz]-4.5[KHz] 대역통과 필터와 증폭기, 그리고 TMS320C30 디지탈 신호처리 프로세서를 장착한 DSP보드를 사용하였다.

ABSTRACT

This paper is a study on the design and implementation of the Korean Text-to-Speech system which is used for a small and simple system.

In this paper, a parameter synthesis method is chosen for speech synthesis method, we use PARCOR(PARtial autoCORrelation) coefficient which is one of the LPC analysis. And we use phoneme for synthesis unit which is the basic unit for speech synthesis. We use PARCOR, pitch, amplitude as synthesis parameter of voice, we use residual signal, PARCOR cofficients as synthesis parameter of unvoice. In this paper, we could obtain the 60% intelligibility by using the residual signal as excitation signal of unvoiced sound. The result of synthesis experiment, synthesis of a word unit is available. The controlling of phoneme duration is necessary for synthesizing of a sentence unit.

For setting up the synthesis system, PC 486, a 70[Hz]-4.5[KHz] band pass filter for speech input/output, amplifier, and TMS320C30 DSP board was used.

*광운대학교 전자계산기공학과
Department of Computer Engineering, KwangWoon University
접수일자 : 1995년 1월 18일

Ⅰ. 서 론

최초의 음성 합성기는 1779년에 만들어졌다고 전해지며, 1791년에 Von Kempelen에 의해 기계식 음성 합성기가 제작되었다. 최초의 전기적 음성 합성기는 1922년에 J. Q. Stewart 에 의해 만들어졌는데 전기적 공진 회로에 의해서 모음을 생성해 냈다. 최초의 연속음성을 합성할 수 있는 음성 합성기는 Voder라고 불리는 합성기로서 1939년에 H. Dudley에 의해 개발되었는데 발판 및 10개의 건반을 이용해 기본 주파수 및 공진 필터의 특성을 제어하여 음성을 합성하도록 설계되어 있다. 1960년 Fant에 의해 음성 생성의 음향학적 원리 및 음성 생성의 디지틀 모델이 발표되고 1980년대 후반 이후로 디지틀 신호처리 기술 및 컴퓨터의 발달로 인하여 본격적인 음성 합성의 연구가 진행되고 있다. 국내에서는 ETRI를 비롯한 연구 기관과 금성사등의 기업체에서 LSP(Line Spectrum Pair) 및 LPC(Linear Predictive Coding) 합성 방법을 이용한 연구가 진행되고 있다. 국내 TTS의 시제품으로는 Digicom의 "가라사대"와 LSP를 이용한 삼성의 "한국어 문서-음성 변환장치"등이 있으며, 국외에서는 미국, 일본 및 유럽 여러나라 등에서 음소 및 음절을 합성단위로 포만트(formant), PARCOR 계수를 이용한 TTS 시스템이 실용화 단계에 있다.

본 논문에서는 한국어 문자-음성 변환 시스템에서 소규모 시스템에 적용 가능하도록 하기 위하여 가장 기본적인 합성단위인 음소를 합성 단위로 사용하였다. 분석 파라메터로 10차의 PARCOR계수와 피치, 영교차율, 에너지 등을 사용하였고, 합성 파라메터로는 유성음의 경우 PARCOR 계수, 피치, 진폭을 사용하였고, 무성음의 경우 명료도 향상을 위해 잔차신호와 PARCOR 계수를 사용하였다.

Ⅱ. 기본 개념

2.1 음성 합성

2.1.1 대상어휘에 따른 분류

제한 어휘 합성은 합성하고자 하는 어휘들을 미리 분석하였다가 이들의 조합에 의해 말을 합성하는 방법으로 합성 대상 어휘가 제한된다. 주로 단어 또는 소문장 단위의 음편들을 연결하여 말을 합성하는데 현재 ARS(Audio Response System), 지하철 안내 방송 등에 이용되고 있다. 구현이 용이하며 무제한 어휘 합성에 비하여 높은 음질을 얻을 수 있으나 음편들의 연결부분이 부자연스러우며 합성 대상 어휘가 바뀔 때마다 다시 녹음, 분석하여야 하는 단점이 있다. 그림 2.1는 음성 합성의 분류를 나타내고 있다.

무제한 어휘 합성은 언어의 기본 단위인 음소, 음절등의 조합에 의해 말을 합성해 내므로 합성 대상 어휘에 제한이 없으며 주로 TTS(Text-to-Speech) 장치등에 적용된다. 그러나 음소, 음절등의 연결시 상호 조음현상의 처리 및 자연스러운 운율처리등이 아직 미흡하여 현재까지는 제한 어휘 합성에 비하여 음질이 매우 떨어지는 실정이다.

2.1.2 합성방법에 의한 분류

파형 합성법은 구현이 용이하고 합성음의 음질이 좋은 반면 저장해야 할 데이타 양이 많아 제한 어휘 합성기에 주로 쓰인다. 최근 자동응답 시스템에 많이 이용되는 방식으로 유럽등지에서는 PSOLA(Pitch Synchronous Over Lap and Add) 방식이 규칙 합성에 적용되고 있다.

파라메터 합성법은 인간의 성도 특성을 모델링하여 특징 파라메터의 시간적 변화 정보에 의해 음성을 합성한다. 파형 합성법에 비해 연산량이 많고 음질도

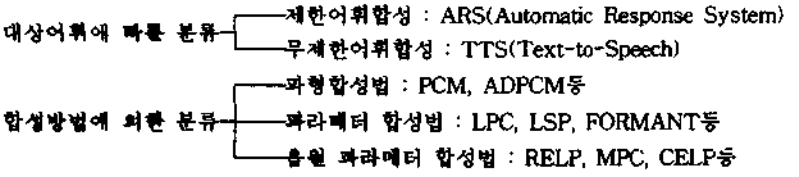


그림 2.1 음성 합성의 분류
Fig 2.1 Classification of Speech Synthesis

떨어지나 데이타의 압축률이 높고 특징 파라메터의 변환에 따라 말의 속도, 음높이, 스펙트럼 변환 등이 용이하여 주로 무제한 어휘 합성에 응용된다. 대표적인 합성 방법으로 포만트 합성법, LPC 합성법 등이 있다[5].

가) 포만트 합성법

인간의 성도는 여러가지 단면적을 갖는 관의 연결로 모델링 될 수 있는데 각각의 관의 단면적에 따라 고유의 공진 주파수를 가지며 이는 입모양, 혀의 위치등에 따라 변하므로 이에 따라 발음의 변별적 특성이 생긴다. 성도의 공진 주파수를 포만트(formant)라 하며, 3-5개의 포만트로써 성도 특성을 나타낼 수 있다. 이와 같이 청취(지각)의 관점에서 음성신호의 주파수 스펙트럼에만 주목하여 전기적으로 흉내 낸 것이 포만트 방법이다. 이 방법은 인간의 청취가 음성의 스펙트럼에 근거하고 있다는 개념을 이용한 것이다. 음성 스펙트럼을 포만트라고 부르는 성도의 3-5개의 공진 주파수와 그 대역폭으로 표현하고, 포만트에 의한 공진회로를 여러 개 접속해서 성도와 등가인 전달특성을 실현한다. 공진 특성을 직접 표현하므로 음성 파형과의 대응도 쉽고 합성음 제어의 규칙화도 쉽지만 포만트의 완전한 자동 추출이 어렵다는 단점이 있다. 주로 미국에서 많이 사용된다.

나) 선형 예측(LPC) 합성법

음성신호는 표본간의 상관성이 높으므로 과거의 N개의 표본으로 현재의 표본값을 예측할 수 있다. 이 예측 계수를 이용하여 all-pole 성도 모델 필터를 구성하고 음성신호를 필터링하면 음성신호를 합성해 낼 수 있다. 전통적인 방식에서 음원 여기신호로 유성음의 경우 펄스열을 무성음의 경우 백색 잡음신호를 이용한다. 음성신호를 all-pole모델로 합성하므로 유성음의 경우 좋은 합성음을 만들어 낼 수 있으나, zero 특성이 나타나는 비음의 처리가 미흡하다. 선형 예측 계수를 a(k), 이득 계수를 G라 하면 p차 선형 예측 계수에 의한 성도 전달 함수 H(z)는 아래 식(2.1)과 같다.

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a(k) z^{-k}} \qquad (2.1)$$

선형 예측 계수를 구하는 방법으로는 autocorrelation 법, covariance법, lattice법 등이 있으며 각각의 방법은 연산량, 저장 데이타의 양, 안정성 등에서 장단점을 갖는다. 실제 선형 예측 계수를 저장 또는 전송할 때에는 일정한 bit으로 양자화하는데 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[6].

PARCOR 분석합성 시스템은 LPC분석에 의해 얻어지는 스펙트럼의 포락이 불안정하기 때문에 편자기상관함수(PARtial autoCORrelation)를 이용하여 파라메터를 추정하는 방법이다. 또 최근에는 LSP 합성법를 많이 사용하는데 앞에서 설명한 PARCOR 합성법과 마찬가지로 성도를 all-pole 모델로 모델링한다. PARCOR 계수는 시간영역에서 작용하는 계수인데 반해 LSP계수는 주파수 영역에서 작용하므로 양자화 또는 선형 보간법에 따른 음성 특성 제어가 용이하기 때문이다. 그러나 계산량이 증가하며, 분석 차수에 따라 LSP 계수값이 달라지는 단점이 있다. 본 논문에서는 PARCOR 분석 합성을 이용하였다. 음원 파라메터 합성은 파형 합성법과 파라메터 합성법의 장점을 혼합하여 낮은 정보량으로 고음질의 음성을 합성해 내는 방법이다. 주로 선형 예측 방법의 변형으로써 음성코딩 및 제한 어휘 합성에 많이 쓰이며, 피치변화등 운율제어가 어려우므로 무제한 어휘 합성에는 적용하기에 한계가 있다.

2.2 문자-음성 변환 시스템

2.2.1 규칙합성의 원리

정서법으로 표기된 텍스트를 입력으로 하는 규칙합성 시스템의 경우에는 어휘에 제한이 없어야 하고, 텍스트 정보상에는 운율 정보가 직접 표현되어 있지 않으므로 원음성의 분석에 의해 그대로 합성해 내는 분석합성만으로는 합성음성을 만들어 낼 수 없다. 이러한 처리에 앞서서, 텍스트가 어떤 단어로 구성되어 있고 주부와 술부의 경계는 어디인가와 같은 언어정보의 추출이 필요하며, 추출된 언어정보와 음성의 음향적 특징을 관련지우는 조작이 필요하다. 따라서 텍스트로부터 음성을 합성하려면 그림 2.2과 같은 4단계의 처리가 필요하다. 각 단계에서의 처리내용은 각각의 합성 시스템에 따라 다르겠지만 개략적으로 다음과 같이 정리될 수 있다[7].

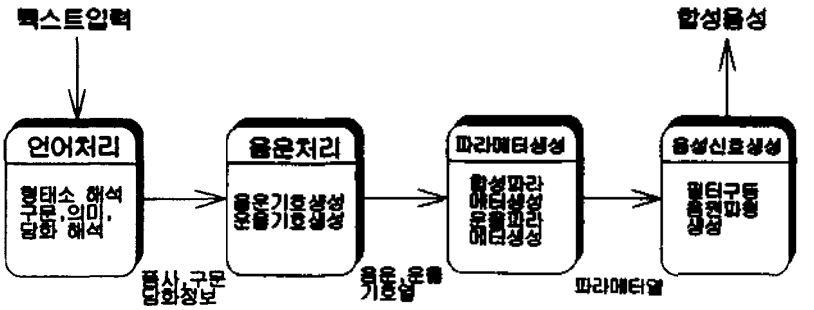


그림 2.2 규칙합성 시스템의 구성
Fig 2.2 Configuration of synthesis by rule system

표 2.1 규칙합성에 이용하는 음성단위와 그 특징[12]
Table 2.1 Speech units and their characteristics for synthesis by rule

음 성 단 위	단위의 총수	특 징
음운(음소)	30~50	가장 기본적인 단위. 단위의 갯수는 적으나 각종 음운 변형 규칙 필요.
음 절	100~2000	음운과 같은 음성학적 기본 단위. 언어에 따라 갯수가 많아지는 것이 결점.
dyad(diphone)	150~2000	음운 과도부를 포함하므로 명료성을 쉽게 확보가능. 갯수가 적당하여 사용이 용이.
demi-syllable	100~1000	조음결합에 의한 음운변화의 완전 흡수는 불가능.
VCV CVC 등 음운복합단위	수백~수천	조음결합에 의한 음운변화를 어느 정도 흡수할 수 있어 음운변형 처리를 줄일 수 있음. 언어에 따라 갯수가 너무 많아지는 결점이 있음. 단위간 접속에 유의할 필요.

2.2.2 합성 단위

규칙에 의한 음성합성에서는 임의 어휘의 출력이 가능해야 하므로 단어보다 작은 단일 음운(음소), 또는 2-3개의 음운 연쇄를 합성의 기본 단위로 한다. 이들 단위 중 가장 원리적인 것이 단일 음소를 이용한 것이다. 그밖의 합성단위로는 이중음소(diphone), 반음절(demisyllable), 단어(word) 및 이중음소에서 실현되지 않은 조음 결합에 의한 음운변동을 흡수하기 위해 VCV나 CVC등과 같은 음운 복합단위도 이용할 수 있다. 표 2.1은 규칙 합성에 이용하는 음성단위와 그 특징에 대하여 나타내고 있다.

Ⅲ. 합성 파라메터의 추출

3.1 분석 실험

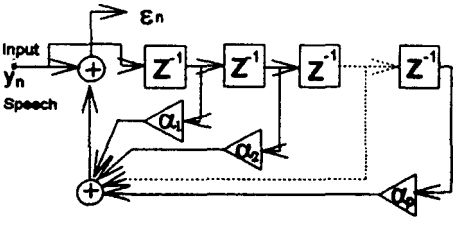
3.1.1 선형 예측 분석

시스템 이론에 의하면 입력 여기 신호{e_n}와 출력 음성 신호{x_n}사이의 관계식인

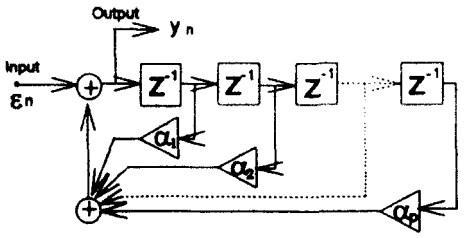
$$x_n + \sum_{i=1}^{p} a_i x_{n-1} = e_n \tag{3.1}$$

는 AR(Auto-Regressive) process라고 불린다. 양측에 z-변환을 행하면 시스템 함수 H(z)를 얻을 수 있는데 이 함수는 pole만을 포함하고 있다.

$$H(z) = \frac{1}{1 + a_1 z^{-1} + \ldots + a_p z^{-p}} \tag{3.2}$$



(a)



(b)

그림 3.1 LPC 분석 및 합성회로 (a) 분석회로 (b) 합성회로
Fig 3.1 LPC analysis and synthesis circuit
(a) analysis circuit (b) synthesis circuit

이 시스템을 all-pole 시스템 또는 모델이라 한다.

입력 음성 파형의 표본에 대한 선형 예측 해석은 음성 파형을 AR process로 가정하고 이에 대한 all-pole 시스템 모델을 구하는 과정이라고 할 수 있다. 아래의 그림 3.1은 선형 예측 계수(LPC)를 이용한 음성 분석 및 합성 회로이다.

3.1.2 PARCOR 해석방식

선형 예측 파라메터는 시간 영역에서 정의된 것으로 프레임간 또는 단위의 접속시 인접 구간과의 선형 보간(linear interpolation)에 취약하다는 단점이 있다. 이는 주파수 영역에서 정의된 파라메터를 이용하면 극복될 수 있는데 LSP(line spectrum pair)방식이 그 예이다.

본 연구에서는 분석 및 합성 파라메터 선정 시 PARCOR와 LSP를 고려하였지만 LSP 역시 무성음의 합성에는 그렇게 좋은 특성을 나타내지 못하였고, 두 방식에서 합성음의 큰 차이를 발견할 수 없었기 때문에 상대적으로 계산량이 적고 구현하기 쉬운 PARCOR를 합성 파라메터로 사용하였다. PARCOR 합성 과정은 무손실 음향관을 통과하는 음파의 진행으로 이해되고 있다.

PARCOR 계수의 추출 방법은 LPC 분석 중 Levinson-Durbin's 일고리즘을 사용하는 방법과 lattice 방법 등이 있는데 전자의 방법이 LPC 계수와 함께 여러가지 파라메터를 같이 구할 수 있기 때문에 본 논문에서는 Levinson-Durbin's 일고리즘을 이용하였다. 피치 검출에는 3-level center clipping 방식을 이용하였으며 진폭은 잔차 신호의 에너지를 이용하였다.

PARCOR 음성 분석 방식은 입력 파형 표본에서 순차적으로 correlation을 제거하여 나가는 방식으로 all-pole 스펙트럼의 포만트 구조를 차례차례로 역 필터링 하여 스펙트럼을 평탄하게 만들어 가는 과정이다. 따라서 그 해석 결과인 잔차 신호는 유성음의 경우 임펄스열과 같은 모양을 갖고, 무성음의 경우에는 백색 잡음(white noise)과 같게 된다. PARCOR 음성 합성은 분석 방법과 반대로 잔차 신호에 순차적으로 correlation을 더해 주는 과정으로 평탄한 주파수 스펙트럼을 갖는 잔차 신호에 포만트 구조를 만들어 가는 것이라 할 수 있다. 만약 이 때 음성 분석에서 얻은 잔차 신호 파형을 입력으로 사용하면 PARCOR 음성 합성 후 원래의 입력 음성과 동일한 파형을 얻을 수 있다. 그림 3.2는 PARCOR 합성 필터 회로이다[16].

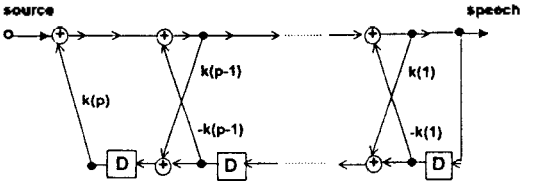


그림 3.2 PARCOR 합성 필터
Fig 3.2 PARCOR synthesis filter

3.1.3 무성음 구간의 잔차 신호

앞에서 설명하였듯이 합성 필터에 LPC 분석 중 구해지는 잔차 신호를 입력하였을 때 원신호와 같은 신호가 합성된다. 그림 3.3은 원신호와 잔차신호로 부터 합성된 합성신호를 보여준다.

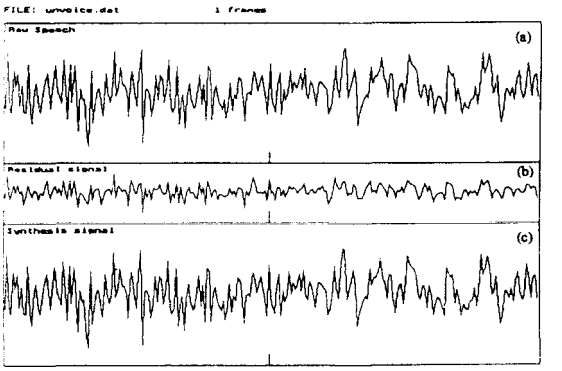


그림 3.3 무성음구간의 원신호와 합성신호 (a) 원신호 (b) 원신호로부터 추출된 잔차신호 (c) 합성신호
Fig 3.3 The original and synthesized signal in unvoice sound (a) raw signal (b) residual signal extracted from raw signal (c) synthesized signal

3.2 피치 및 진폭

3.2.1 3 level center clipping방식을 이용한 피치의 검출

피치 검출의 목적은 피치 주기에 연관된 두가지의 모델 계수를 얻는 데 있다. 이것은 유/무성음의 여부에 대한 것과 피치 주기의 값이다. 피치 주기의 여부 검출로 유/무성음에 대한 판별을 하고 피치 주기 추출로 이 계수들을 구하게 된다. 피치 정보는 음성의 자연성에 큰 영향을 미치기 때문에 인간의 청각은 피치 변화에 매우 민감하게 반응한다. 그러므로 정확한 피치의 추출은 합성음성의 음질을 좌우하는 중요한 요소이다.

자기 상관 함수를 이용한 피치 주기 검출에 있어서 입력 신호의 작은 값들을 0로 center clipping한 후 구한 자기상관함수를 사용하면 피치 검출기의 성능을 향상시킬 수 있다. 또한 입력신호의 level을 +1, 0, −1로 양자화하여 자기상관을 구하면 계산량을 대폭 줄일 수 있다. 이것을 3-level center clipping이라 한다[15].

3-level center clipping의 출력을 Y(n)으로 하면, 자기상관함수에서 생산조건 y(n+m)y(n+m+k)은

$$R_n(k)=\sum_{m=0}^{N-k-1} y(n+m)\,y(n+m+k)$$

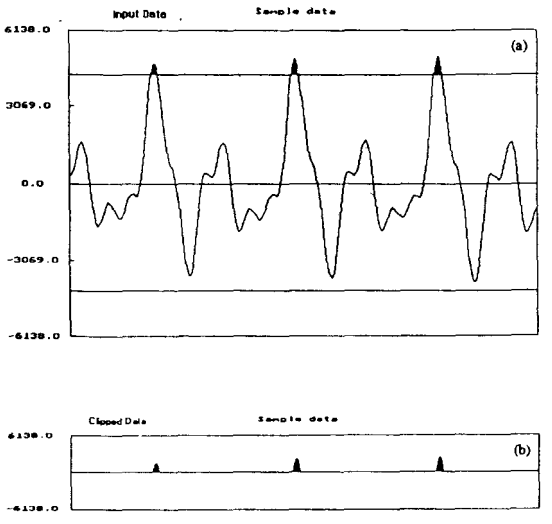


그림 3.4 3-level center clipping 피치검출
(a) 입력 data (b) center clipped 데이타

Fig 3.4 Pitch detection using 3-level center clipping
(a) input date (b) center clipped data

세가지 다른 값을 가질수 있다.

$$y(n+m)y(n+m+k)=\begin{cases}0 & \text{if } y(n+m)=0 \text{ or } y(n+m+k)=0\\ +1 & \text{if } y(n+m)=y(n+m+k)\\ -1 & \text{if } y(n+m)\neq y(n+m+k)\end{cases}$$

그림 3.4는 3-level center clipping 피치검출기의 피치 검출 예이다.

그림 3.5는 유성음 구간의 원신호와 합성신호의 피치를 비교해 주고 있다.

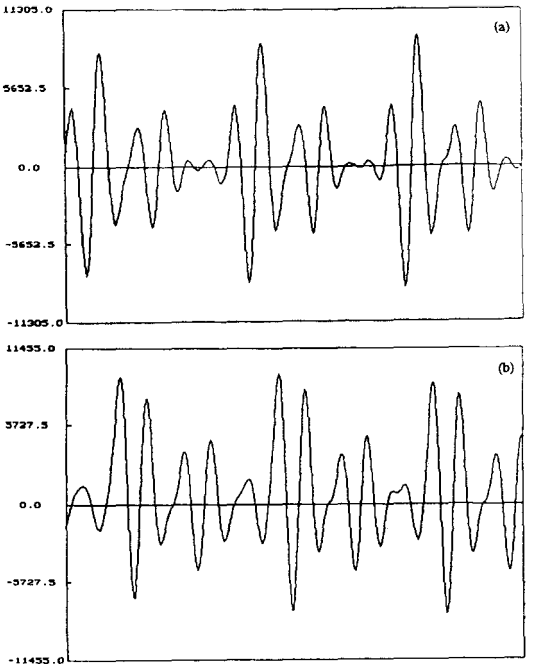


그림 3.5 원파형과 합성파형에서 Pitch비교
(a) 원파형 (b) 합성파형

Fig 3.5 The comparision of pitch in original and synthesized signal
(a) raw date (b) synthesized data

3.2.2 진폭의 검출

PARCOR 합성 회로에 잔차신호가 입력되었을 때 원래의 신호와 같은 파형이 합성된다. 합성 필터의 여기신호 모델링은 유성음의 경우 임펄스로, 무성음

의 경우 백색 잡음(white noise)으로 한다. 임펄스와 백색 잡음(white noise)모두 평평한 스펙트럼 구조를 갖는다.

음성의 한 구간을 진폭 A와 주기 T로 표현할 때 T=0이면 무성음의 잡음원, T>0이면 임펄스원이다. T=A=0이면 묵음 구간이다. A와 T에 관한 이러한 모든 정보는 잔차 신호로 부터 추출될 수 있다. 또한 잔차 신호의 에너지 값으로 원 진폭을 결정한다. 그리고 잔차 신호의 자기상관이 주기적인가 아닌가에 따라 유/무성음 판별을 할 수 있다[14][17].

Ⅳ. 문자-음성 변환 시스템의 구현

4.1 파라메터 데이타베이스의 구축

4.1.1 합성단위의 선정

규칙 합성음의 이해성 평가를 수행하는 경우 평가 단위는 매우 중요하다. 평가 단위에는 "음절", "단어", "문장" 등이 있으며 각각 평가 대상 및 평가 목적에 따라 구분하여 사용할 필요가 있다. 즉, 합성음의 음운에 대한 명료성을 진단적으로 평가하고 싶은 경우에 평가 단위로서 음절이 적당하고, 단어와 문장은 합성 방식의 이해성에 관한 전체적인 성능 평가에 적합하다.

특히, 음운의 접속에 따른 품질 열화의 가능성이 있는 규칙 합성 방식에서는 단어나 문장이 평가 단위로서 매우 중요하다. 한편, 문장을 이용하여 규칙 합성음의 이해성 평가를 수행할 경우 평가 결과가 문장의 내용에 따라 영향을 받기 쉬우며 또, 조건의 수가 많아지면 방대한 문장이 필요해지고 평가 시험을 실시할 때 시간상, 처리상 어려움이 발생하는 문제점이 있다. 그러나, 단어는 문장과 비교하여 앞에서 언급한 문제점이 적을 뿐 아니라 다른 나라에서도 "단어 이해도"를 이해성 평가를 위한 기본 단위로 적합하여 수행하고 있는 점을 고려할 때 평가 단위로서 단어가 매우 적당하다. 그러나 규칙 합성 방식에 의한 시스템은 일반적인 말을 합성하여 음성으로 출력하는 시스템이기 때문에 이해성 평가를 위한 시험용 단어의 범위를 가능한 넓혀야 한다. 자연 음성이라면 보통 처음 듣는 단어일지라도 정확하게만 발성되면 이해할 수 있으며, 합성 음성에 있어서도 충분히 이해성을 얻을 수 있다면 복잡한 단어일지라도 비슷한 단어로 이청하지 않고 알아들을 수 있다. 그러나, 단어의 범위를 넓힘에 따라 합성음의 이해성에 대한 품질면에서 이해도의 차가 생길 것이다. 즉, 시험용 단어에서 단어의 수가 적게 제한되고, 평이한 단어로 구성되어 있으면 그 단어를 반복하여 시험에 사용할 경우 피험자가 한번 청취한 단어의 대부분을 기억해 버리는 결과를 초래하기 때문에 이해성 평가에 부적당하게 된다.

따라서, 합성 방식의 개선된 정도를 조사하는 경우와 다른 방식과의 우월을 비교하는 경우 등에 있어서 가능한 많은 단어를 이용하여 평가 시험을 수행하여서 합성음의 이해성을 정량화시킬 필요가 있다고 판단되어 본 연구에서는 이와같은 입장을 고려하여 음소 단위를 이용한 지속적인 연구가 필요하다고 생각된다.

본 논문에서의 음성합성 단위는 음소로 결정하였다. 음소단위 합성은 음소간 연결부분의 음질저하가 문제이기는 하지만 소규모의 시스템에 이용할 수 있다는 장점이 있다. 먼저 한국어의 초성, 중성, 종성의 구성에 대해 조사하였다. 표 4.1은 음가가 다른 각 음소들을 초, 중, 종성 별로 구분한 표이다.

초성과 중성은 모든 음소를 그리고 종성의 경우 대표음 7개에 대하여 각 음절 단위로 녹음을 하였다. 녹음 환경은 표본화 주파수는 10kHz, 양자화 레벨은 16bit로 A/D 변환하여 컴퓨터의 하드디스크에 PCM 데이타로 저장하였다. 입력된 2byte 조합형 한글을 음소단위로 분리한다. 이때 각 음소의 자음, 단모음, 이중모음 여부와 초성, 중성, 모음여부가 결정된다. 이 정보에 따라 음소열로 부터 연접된 음소간의 결합규칙을 찾아 각 규칙에 따른 해당 코드(조합형 한글) 열을 생성해 낸다. 생성된 코드들에 해당하는 번호

표 4.1 합성단위로 선정된 음소

Table 4.1 The selected phonemes for synthesis units

구 분	음 소	갯 수
초 성	ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ, ㄲ, ㄸ, ㅃ, ㅆ, ㅉ	18 개
중 성	ㅏ, ㅓ, ㅗ, ㅜ, ㅡ, ㅣ, ㅐ, ㅒ, ㅟ, ㅚ, ㅔ, ㅖ, ㅑ, ㅛ, ㅠ, ㅙ, ㅞ, ㅝ, ㅢ, ㅘ	21 개
종 성	ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅇ	7 개

열과 세그멘테이션 정보 열이 출력된다.

```
합성할 문장을 입력하세요 : 안녕하세요
  아스키 ' ' _
      중성 'ㅏ' V   종성 'ㄴ' C
      초성 'ㄴ' C   중성 'ㅕ' gV'  종성 'ㅇ' C
      초성 'ㅎ' C   중성 'ㅏ' V
      초성 'ㅅ' C   중성 'ㅔ' V
      중성 'ㅛ' gV
  아스키 ' ' _
```

또한 운율처리에 필요한 음소의 길이를 제어하기 위하여 각 음소의 길이를 Klatt의 분절음 지속 시간 규칙을 이용하여 구한다. 자연성 향상을 위한 합성음의 운율조절은 문자-변환 시스템에 있어서 필수적이다. 합성음의 운율 조절을 위해 사용되는 파라미터로는 지속시간(duration), 휴지, 에너지, 기본주파수의 변환 패턴 등이 있다. 발성의 지속시간과 휴지의 삽입은 합성음의 명료도 및 자연성에 아주 큰 영향을 미친다. 본 논문에서는 Klatt의 분절음 지속시간 규칙을 이용하여 음절 내 음운환경에 따른 각 음소에 대한 시간을 산출하였다. 그 규칙은 다음의 식으로 정리된다.

$$Pdur = MINDUR + ((INHDUR - MINDUR) \times PRCNT)/100$$

MINDUR : 각 음소의 최소 지속시간
INHDUR : 각 음소의 고유 지속 시간
PRCNT : 음운환경에 따른 음소의 신축비

MINDUR, INHDUR의 값은 ETRI에서 조사한 연속 음성에서 각 음소의 평균 지속시간 분포를 참조로 하였으며, PRCNT의 값과 휴지 구간의 삽입은 Klatt의 지속시간 규칙을 참조로 하였다.

4.1.2 음소 단위 구분화(segmentation)

그림 4.1은 음절단위로 녹음된 데이타에 대해 음소단위로 구분화하는 과정을 보여주고 있다. 구분화는 음성의 스펙트로그램을 보면서 목측을 통해 행하였고, 각 구분화 된 음소들은 화일 단위로 저장된다. 본 논문에선 무성음 합성시 잔차신호를 합성 필터의 여기신호로 사용하여 합성음의 명료도를 향상시킬 수 있었다. 이런 방법은 저장 용량이 커집다는 단점이 있지만 현재 문자-음성 변환 시스템의 연구에 있어서 가장 중요한 요소인 음질 향상을 위해서는 불가피한 방법이다.

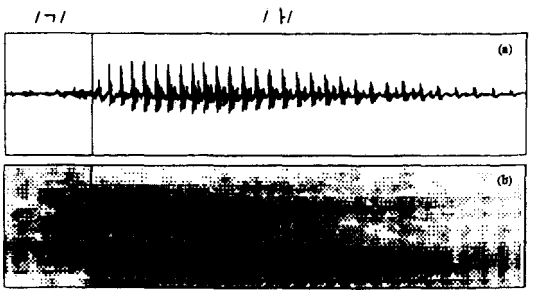


그림 4.1 음절로 부터 음소단위 구분화
(a) 음절 /가/ (b) 음절 /가/의 스펙트로그램
Fig. 4.1 The phoneme unit segmentation from syllable
(a) syllable /ga/ (b) spectrogram of /ga/

또한 명료도와 자연성의 향상을 위해서 몇가지 에너지 조절 규칙을 제안했다. 합성 단위간의 에너지 평활화 규칙을 적용하여 합성 단위간의 녹음 레벨 불일치로 인한 음질 저하를 최소로 할 수 있었다. 또한 유성음 구간의 에너지 평활화를 통해 합성음의 자연성을 향상시킬 수 있었다 이러한 평활화 규칙을 이용한 합성음은 평활화 이전의 합성음 보다 약간 울리는 소리가 합성되었지만 전체적인 합성음의 자연성 면에서는 좋은 결과를 나타냈다.

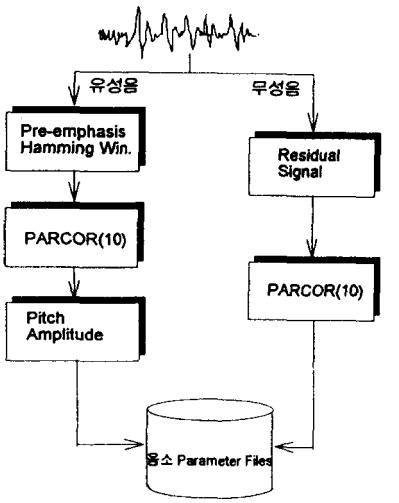


그림 4.2 합성 파라메터의 추출
Fig. 4.2 The extraction of synthesis parameter

4.1.3 파라메터 추출 및 저장

음소 단위로 구분화된 PCM 음소 데이타에 대해 합성파라메터를 추출한다. 3장에서 설명한 바와같이 유/무성음 여부에 따라 프레임별로 각각 다른 파라메터가 사용된다. 그림 4.2는 파라메터추출 과정을 유/무성음에 따라 보여주고 있다. 그리고 유성음 구간 한 프레임의 저장에는 48바이트가 사용되고 무성음 구간의 경우 552바이트가 사용된다.

추출된 파라메터들은 화일 단위로 저장되는데 그 구조는 그림 4.3과 같다.

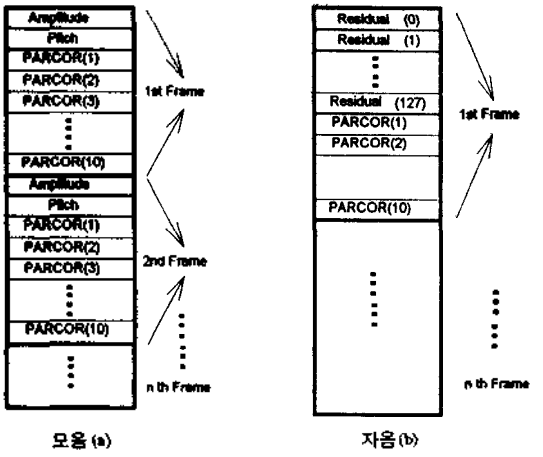


그림 4.3 파라메터 화일의 구조
(a) 모음 파라메터화일 (b) 자음 파라메터 화일

Fig. 4.3 The structure of parameter File
(a) The parameter file of vowels
(b) The parameter file of consonants

4.2 합성 필터의 구현

PARCOR 합성에 사용되는 디지탈 필터는 lattice 필터이다. 그 구조는 앞의 그림 3.2에서 보여 주고 있다. 그림 4.4는 합성 필터를 포함한 음성 합성부의 구조이다. 각 합성 파라메터가 합성부에 어떻게 적용되는가를 나타내고 있다.

4.3 구현된 문자-음성 변환 시스템의 구조

본 절에서는 본 논문에서 구현된 음소 단위 문자-음성 변환 시스템에 대해 설명한다. 그림 4.5는 본 논문에서 구현된 문자-음성 변환 시스템의 구조이다.

문장 해석부에서는 입력된 2 바이트 조합형 한글을

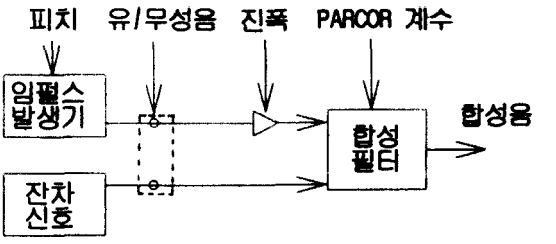


그림 4.4 음성 합성부의 구조

Fig. 4.4 The structure of speech synthesis part

음소 단위로 분리한다. 이때 각 음소들의 초, 중, 종성 여부가 결정된다. 각 음소에 해당하는 파라메터 화일의 정보가 출력되어 파라메터 생성부로 입력된다. 파라메터 생성부는 각 음소에 해당하는 합성 파라메터들을 파라메터 사전으로 부터 메모리로 적재한다. 마지막으로 음성합성부에서는 입력된 파라메터 열에 대해 합성 필터링을 한다. 이때 사용되는 필터가 lattice 필터이다. lattice 필터에 대해서는 3.1.2 절에서 설명하였다. 합성 필터의 출력으로 음성파형이 생성된다. 이 합성 데이타는 하드디스크에 저장되었다가 D/A 변환되어 스피커를 통해 출력된다.

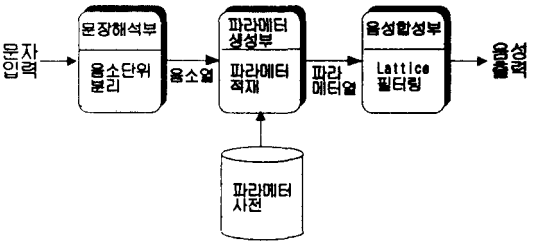


그림 4.5 구현된 문자-음성 변환 시스템의 구조

Fig. 4.5 The structure of text to speech system

V. 실험 및 고찰

5.1 실험 환경

본 논문은 컴퓨터상에서 소프트웨어로 구현하였다. 문자는 키이 보오드를 통해 한글 2바이트 조합형으로 486 PC에 입력되며 저장된 파라메터 중 필요한 파라메터를 메모리로 가경가 운율 조절을 한다. 합성 파라메터 화일들은 하드디스크에 저장되어 있다가 그때 그때 필요한 파라메터들이 메모리로 적재된다.

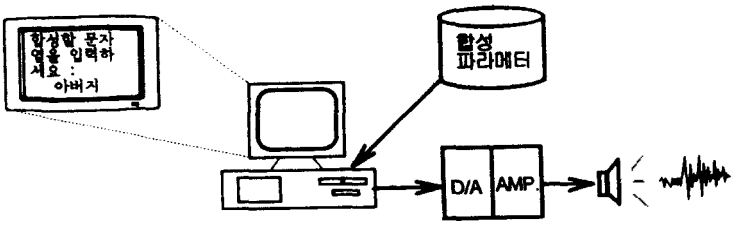


그림 5.1 실험 환경
Fig. 5.1 The experimenting environment

합성필터를 통해 출력된 합성파형은 PC에 장착된 TMS320C30 DSP 보드를 통한 접속 방법으로 D/A 변환되어 증폭기를 통해 스피커로 출력된다. 그림 5.1은 실험 환경의 개략적 그림이다.

5.2 단음절 합성 실험

먼저 합성기의 출력상태를 알아보기 위하여 무의미 단음절에 대해 합성실험을 행하였다. 그림 5.2.1과 그림 5.2.2는 각각 단음절 /서/와 /전/의 합성파형을 보여준다.

5.3 단어 합성 실험

다음은 단어 단위의 합성 실험을 하였다. 그림 5.3.1과 그림 5.3.2는 각각 단어 /아버지/와/ 컴퓨터/의 합성 파형이다. 일반적으로 LPC 계열의 합성법은 비음의 합성에 매우 취약하다. 그러나 본 논문에서 잔차신호를 무성음의 여기 신호로 이용했기 때문에 /아버

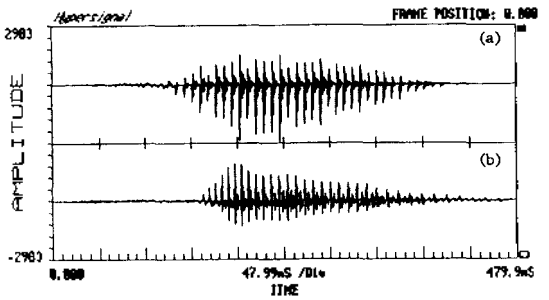


그림 5.2.1 단음절 합성실험 /서/ (a) 자연음 (b) 합성음
Fig. 5.2.1 The synthesis experiment of monosyllable /seo/
(a) natural speech (b) synthesized speech

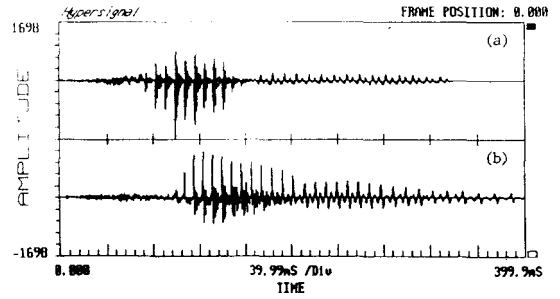


그림 5.2.2 단음절 합성실험 /전/ (a) 자연음 (b) 합성음
Fig. 5.2.2 The synthesis experiment of monosyllable /jeon/
(a) natural speech (b) synthesized speech

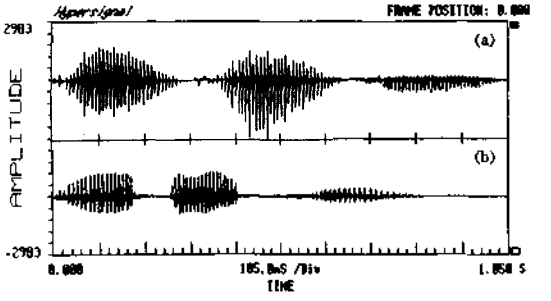


그림 5.3.1 단어 합성 실험 /아버지/ (a) 자연음 (b) 합성음
Fig. 5.3.1 The synthesis experiment of word /abeogi/
(a) natural speech (b) synthesized speech

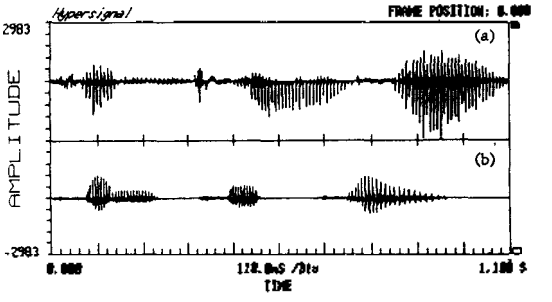


그림 5.3.2 단어 합성실험 /컴퓨터/ (a) 자연음 (b) 합성음
Fig 5.3.2 The synthesis experiment of word /computer/
(a) natural speech (b) synthesized speech

지/에서 /ㅂ/의 비음 구간이 명확하게 합성 되었음을 확인할 수 있다.

5.4 청취 실험

합성음의 객관적인 평가를 위하여 청취 실험을 행하였다. 청취 대상으로는 음성 합성에 대한 사전 지식이 없는 20대 남녀 5인으로 구성하였고, 본 실험에서 선택한 단어들은 본 연구실 논문 "규칙 합성음의 이해성 평가를 위한 단어표 구성 및 실험법"에서 제시된 108개의 1-4음절 단어중 각 음절어에서 5개씩을 선정 발췌한 것이다. 이상의 단어들을 1번씩 들려 주어 청취한 결과를 받아쓰게 한 결과 다음 표 5.1과 같은 청취율을 얻었다.

위의 청취 실험 결과 음절 수가 많을수록 청취율이 저하됨을 알 수 있었다.

이는 음소 지속 시간에 대한 연구가 뒷받침 되지 않았기 때문인 것으로 앞으로 이에 대한 지속적인 연구가 필요하다[18]. 반면 기본 주파수와 음소 지속시간 조절을 통한 합성음의 운율 조절 실험에는 만족할 만한 결과를 얻을 수가 없었다. 이는 우리말에 대한 통계적 특징의 규칙화에 대한 선행 연구가 부족했기 때문이다. 특히 합성음의 품질에 가장 큰 영향을 미치는 요소 중의 하나인 음소 지속시간 및 휴지의 조절은 본 연구에서 매우 어려운 부분이었다.

Ⅵ. 결 론

본 논문에서는 소규모 시스템에 적용 가능한 한국어 문자-음성 변환 시스템을 구현하였다. 합성 파라메터로는 유성음의 경우 PARCOR 계수, 진폭, 피치를 사용하였고, 무성음의 경우 PARCOR 계수, 잔차 신호를 사용하였다. 합성 단위로는 소규모 시스템의 구현에 적합한 46개의 음소를 사용하였다. 음소를 이용한 합성은 충분한 품질의 합성음을 얻기 위해서는 고도의 합성 규칙이 요구되나, 현재까지는 한국어 음성에 대한 분석 자료의 부족으로 뛰어난 합성 음질의 구현은 어려운 실정이다. 일반적으로 LPC 합성시 비음과 무성음의 경우 음질 저하가 문제시 되었으나, 본 논문에서는 PARCOR 분석시 얻어지는 잔차 신호를 무성음 합성시 필터 여기 신호로 사용하여, 명확한 합성음을 낼 수 있었다. 합성음의 객관적인 평가를 위해 청취 실험을 한 결과 60%의 비교적 양호한 이해도를 얻을 수 있었다. 따라서 본 논문에서 구현된 시스템은 단어 단위의 합성에 적용 가능하리라 사료된다. 그러나 음소 길이가 길어질수록 이해도가 떨어지는 현상이 발생하였다. 이와같이 음소 길이에 따라 청취율이 낮은 이유는 음소 지속 시간에 대한 연구가 뒷받침되지 않았기 때문으로 향후 이에 대한 지속적인 연구가 필요하다.

참 고 문 헌

1. 이재홍, 하재규 외, "한국어 무제한 단어음성 합성기 제작", pp. 144, 1988년 음성통신 및 신호 처리 workshop.
2. 이용주, "반음절 단위, LSP방식에 의한 한국어 음성의 규칙합성에 관한 연구", 고려대학교 박 사학위 논문, 1992.
3. Hiroya Fujisaki, "State of art and future prospects

표 5.1 청취 실험을 위한 단어 및 청취율

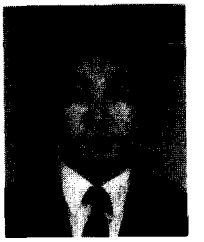
Table 5.1 Word list and hearing rate for hearing experiments

구 분	1 음절	청취율	2 음절	청취율	3 음절	청취율	4 음절	청취율
대상단어	나	5/5 100%	우리	4/5 80%	자동차	3/5 60%	장난치다	1/5 20%
	너	5/5 100%	안녕	4/5 80%	사랑방	3/5 60%	가깝하다	1/5 20%
	남	4/5 80%	사랑	4/5 80%	나그네	3/5 60%	살아나다	2/5 40%
	자	5/5 100%	안내	3/5 60%	가로등	2/5 40%	가공식품	1/5 20%
	수	5/5 100%	가족	3/5 60%	갈대밭	1/5 20%	야단법석	1/5 20%
	평 균	96%		72%		48%		24%
	총 청취율 60%							

of speech synthesis", 일본 전자 정보통 신학회 음성인식, 합성기술의 현상과 장래과제 강습회 교제, 1991. 7.
4. 최영하, 성굉모외, "반음절 데이타베이스를 이용한 MPLPC 한국어 무제한 단어 음성 합성기의 제작", pp. 298-303, 1991.
5. 이윤근, 안승권, "음성 합성 기술 분야", 1993년도 5월 전자공학회지, 한국어 음성 인식 및 합 성기술 특집, pp. 524-531.
6. 이황수, "Vocoder 기술의 최근 동향", 1988년 음성신호처리 workshop.
7. K. Hirose, "Current stends and future prospects of speech synthesis", JASI, vol.48-1, 1992.
8. 이태원외, "양질의 음성합성을 위한 최적의 합성단위 추출에 관한 연구", 한국전자통신연구소, 1993.
9. 이규호 "한국어 문자-음성 변환을 위한 음소분석에 관한 연구", 고려대학교 박사학위 논문, 1993.
10. 정인종, 경연정, 이양희, "고품질의 한국어 합성에 관한 연구-자연음성으로 부터의 다이폰 자 동 추출", 제1회 ETRI 음성, 언어 및 음향 정보처리 workshop 논문집, pp. 128-142, 1993.
11. Lovins, Macchi, Fujimura, "A Demisyllable Inventory For Speec Synthesis", 97th meeting of ASA, 1979.
12. Y. Sagisaka, "음성합성을 위한 운율제어의 연구", 일본 와세다대학 박사학위논문, 1985.
13. L. Rabiner, Biing-Hwang Juang, "Fundamentals of Speech Recognition", Prentice Hall, 1993.
14. S. Saito, K. Nagata, "Fundamentals of Speech Signal Process", Academic Press, 1985.
15. L. R. Rabiner, R. W. Schafer, "Digital Processing of Speech Signal", Presntice-Hall, 1978.
16. 김순협, "한국어 음성의 분석과 자동인식에 관한 연구", 연세대학교 대학원 박사학위 논문, 1982. 11.
17. 김명환, 김순협, "유, 무성음 및 묵음 식별에 관한 연구", 광운대학교 전자계산기공학과 석사학 위 논문, 1985.
18. 김성한, "규칙 합성음의 이해성 평가를 위한 단어표 구성 및 실험법", 광운대학교 전자계산기 공학과 석사학위 논문, 1991. 6.

▲박 애 희(Ae-Hee Park : 학생회원)
1991년 2월 : 서울 산업대학교 전자계산학과 졸업(공학사)
1995년 2월 : 광운대학교 전산대학원 전자계산기공학과 졸업(공학석사)
현재 : 쌍용 정유
※주관심분야 : Speech recognition & synthesis, Speech signal processing 등

▲양 진 우(Jin-Woo Yang : 정·종신회원) 1959년 9월 30일생



1982년 2월 : 원광대학교 전자공학과 졸업(공학사)
1985년 2월 : 광운대학교 대학원 전자공학과 졸업(공학석사)
1994년 8월~현재 : 광운대학교 대학원 전자계산기공학과(박사수료)
1993년 7월~1994년 3월 : 광운대학교 전자계산기공학과 전임조교
※주관심분야 : Voice dialing, Real-time processing, Neural network, Speech recognition & synthesis, Speech signal processing 등

▲김 순 협(Soon-Hyob Kim)
현재 : 광운대학교 컴퓨터공학과 교수
한국음향학회지 제14권 1호