음성합성기술 개발의 현황과 과제

이 양 희 (동덕여자대학교 전자계산학과)

I. 서 론

최근 음성언어에 의한 Man-Machine 인터페이스는 그 편리함 때문에 많은 관심의 대상이 되고 있다. 이를 실현하기 위하여 기계가 음성을 이해하고, 기계로 부터 음성을 생성하는 기술이 필요하다. 이러한 기술 중에서 기계로부터 음성을 생성하는 음성합성은 현재 실용화 단계에 있으며, 우리 주변에서 접할 수 있는 실정이다.

음성합성 방식은 크게 파형부호화, 분석합성형과 규칙합성형으로 나눌 수 있다[1]. 그 중에서 파형부호화 방식과 분석합성 방식은 어휘의 수에 제한을 받으며, ARS(Audio Response System)와 전철의 안내방송등에 이용되고 있다. 한편 규칙합성방식은 언어 정보와 음운 규칙으로부터 무제한 어휘의 음성을 생성하는 방식으로서 활용범위가 다양하며, 근년 반도체 집적 회로의 진보에 따라 선진 각국을 중심으로 활발히 연구되어 일부 실용화 되었고, 음질의 개선을 위한 연구도 계속 진행되고 있다[2]-[5].

근래에 국내에서도 한국어 음성의 규칙 합성에 대한 연구가 활발히 진행되고 있다[6]-[10]. 그러나 아직 실용화하기에는 음질면에서 미흡한 실정이므로 좋은 규칙 합성 시스템을 만들기 위해서는 우리말에 대한 많은 분석을 토대로 하여 규칙 합성 시스템을 이루고 있는 부분 시스템들을 확립하여야 한다. 본 고에서는 양질의 문-음성변환 시스템을 실현을 위해, 개요 및 규칙 합성의 언어 처리 기능과 음성 합성 기능의 기술개발 현황과 과제를 소개한다.

II. 음성합성 방식

인간이 어떠한 정보를 음성으로 표현하는 과정을 몇가지로 나누어 모델화 하면 그림1과 같이 나타낼 수 있다.

사고의 내용, 의도를 말로서 표현하는 인간과 같은 정도의 표현능력을 갖기 위한 「개념으로부터의 음성합성」을 생각하면, 그림1의 첫째 과정인 개념의 표현과 개념의 언어화가 필요하다. 현 시점에서는 실현하기 어려운 문제이나, 「개념으로부터의 음성합성」은 인

간과 기계, 또는 인간과 인간의 원활한 대화를 실현하는 데 하나의 목적이 있고, 단순히 문장의 기저구조를 음성으로 변환한다는 시도에 그치지 않고, 회화나 통신의 입장으로부터 폭 넓게 음성합성의 문제를 생각해 가는 것이 중요하다.

그림1.에서 언어화 과정을 생략한 나머지 부분이 통상 문자열로 나타내어진 문장을 음성으로 변환하는 문-음성변환(Text-To-Speech, TTS)이며, 규칙합성(Synthesis -by-rule)으로 이를 실현할 수 있다.

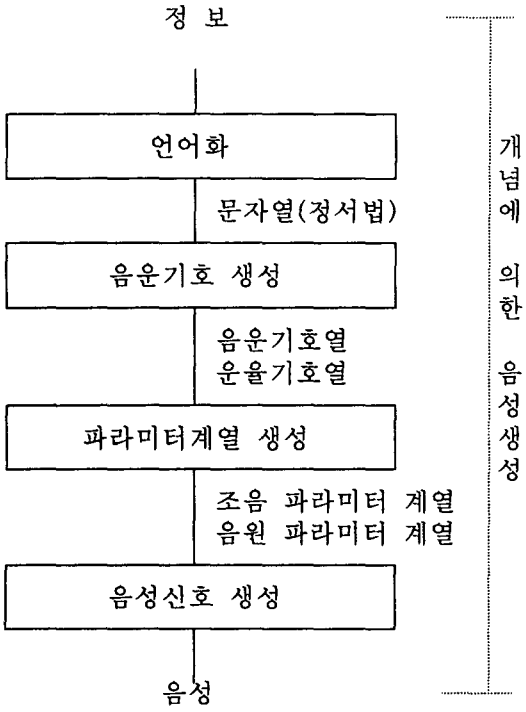

그림1. 음성생성과정의 모델

음성을 합성하는 방식들은, 크게 파형코딩, 분석합성, 규칙합성으로 분류할 수 있으며, 각 합성방식의 특징과 비교가 표1.에 나타나 있다.

문-음성변환 시스템에 적용할 합성방식은 규칙합성이므로, 본고에서는 주로 문-음성변환 시스템에 대하여 기술하고자 한다.

음성을 규칙합성하기 위한 시스템을 구성하기 위해서는, 먼저 언어처리 기능에 의해 입력문장으로 부터 음운기호 열과 운율기호 열이 생성되고, 또 이들로 부터 조음, 음원파라미터열이 생성된다. 이 파라미터열들은 각각의 규칙들을 사용하여 생성할 수 있지만, 이 규칙들은 파라미터-음성변환 방식에 따른다. 따라서, 규칙합성시스템에 있어서 파라

미터-음성변환 방식의 선택은 시스템 전체의 각 부분 시스템과 밀접한 관계가 있다.

표 1. 3가지 음성합성법의 특징과 비교

특징		파형 코딩	분석합성	규칙합성
음질	명료도	높음	높음	중간
	자연성	높음	중간	낮음
어휘의 수		작다(<500)	크다(1000)	무제한
Bit Rate		24-64 kbps	2.4-9.6 kbps	50-75 bps
1Mbit 메모리에 저장되는 음성의 길이		15-40 s	100 s-7 min	거의 무제한
저장 단위		단어, 음절, 문장	단어, 음절, 문장	음소, 음절 (CV, VCV, CVC, 형태소등)
복잡도		낮음	중간	높음
주요 하드웨어		메모리	메모리와 프로셋서	프로셋서
예		아날로그 또는 PCM녹음, ADPCM	채널보코더 LPC(PARCOR, LSP)	MITalk, DECtalk (합성기는 analog, LPC, Cepstrum등 이다)
Terminal				

문자열로 부터 음성으로 변환하는 시스템을 크게 2레벨 구조 즉 언어처리 기능과 음성합성기능으로 나누어 기술한다.

III. 규칙합성계

문-음성변환 시스템은 언어에 따라 다소 다르나, 일반적인 문-음성변환 시스템을 구성하는 주요 요소는 그림 3과 같다. 이 시스템은 크게 언어처리 기능과 음성합성 기능으로 나누어 생각할 수 있다. 언어 처리 기능부에서는 정서법으로 쓰여진 입력 문장을 음운기호와 운율 기호로 변환하여 출력하며, 음성 합성 기능부에서는 음운 기호와 운율 기호를 조음 파라미터 열과 음원 파라미터 열로 변환하고 또한 음성 생성 필터를 구동하기 위한 여기 신호를 생성한다. 이러한 문-음성변환 시스템의 구현에는 음원과 조음파라미터 열을 자연음성의 데이터를 사용하지 않고 완전히 규칙에 의해 음성을 합성하는 순수

규칙합성 방식(포만트 합성)과 조음 파라미터 계열의 일부를 직접 자연음성으로 부터 추출하여 사용하는 보코더(Vocoder)를 이용한 분석합성 방식이 있다[11]. 순수합성 방식은 원리적으로는 이상적인 방식이지만, 조음파라미터열을 생성하기 위한 규칙을 얻기가 매우 어렵다. 이에 반해, 분석합성계를 이용하는 규칙합성 방식은 적어도 단위음성의 데이터로서는 자연음성으로부터 직접 구한 음성특징파라미터를 사용할 수 있다.

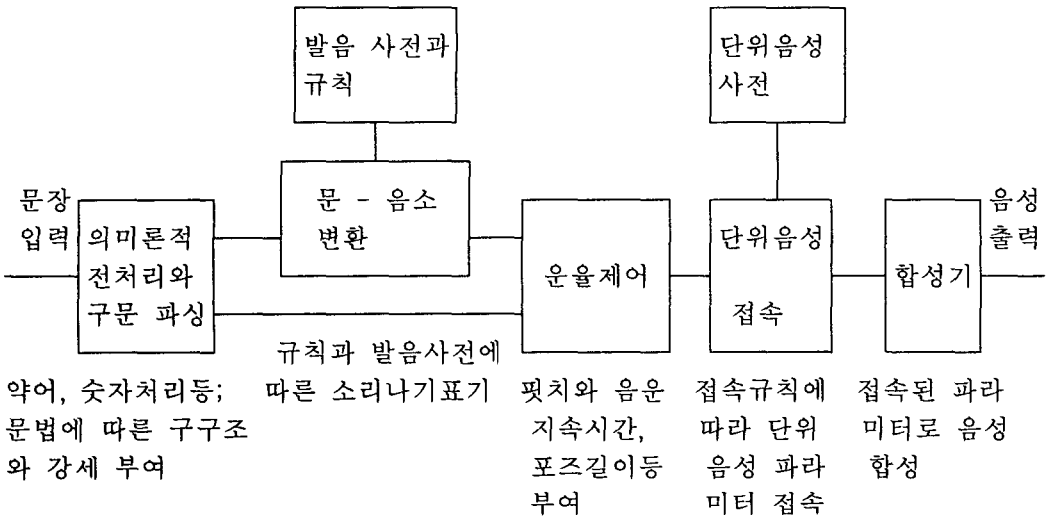

그림2. 음성의 문-음성변환 시스템을 이루는 주요요소

완전한 규칙 합성 시스템의 구축을 위해서는 먼저 음성 합성 기능부의 음성 생성을 위한 시변 필터(Time-varying digital filter)와 단위 음성 파라미터의 접속부, 단위 음성 데이타베이스 구축이 필요하며, 실험적인 방법에 의한 운율 규칙등의 부분 시스템이 확립되고 적용되어야 한다.

이러한 시스템을 구축하기 위해 해결하여야 할 과제는 다음과 같다.

1. 언어처리기능

언어처리 기능에서는 정서법으로 쓰여진 입력 문장을 형태소 및 구문해석, 의미해석을 거쳐 음운기호 및 운율 기호로 변환 할 해석기(Parser)가 필요하다. 입력된 문장은 이 해석기를 거쳐 운율 기호와 음운 기호로 변환되며 기호, 약어, 알파벳등도 여기서 발음 형태의 텍스트로 변환된다. 구문해석에 의하여 문절 접속도에 따른 발화의 운율 세그멘트(운율 그룹, Pause 할당)가 생성되며 피치(Pitch)의 변화와 강세(Stress)기호가 생성되어,

이 들 세그멘트에 적절한 운율 패턴, 즉 운율 기호열이 생성된다. 이러한 운율 기호열은 규칙에 의해 나타내어 진다. 그러나 아직 한국어에 있어서 음운 기호, 운율 기호를 생성하기 위한 언어처리 기능이 미흡한 상태이므로 이 분야의 많은 연구가 요구된다.

1.1. 음운 기호열

입력된 문장은 사전 정보에 의한 형태소 및 구문, 의미해석을 기반으로 하여, 음운 변동 처리되어 소리나기 표기, 즉 음운 기호열로 변환된다. 이 음운 기호 계열은 변이음 처리와 장/단음 처리에 의해 음성 기호로 변환된다. 음운과 음운이 만날 때 양쪽 모두 또는 어느 한쪽의 소리가 변할 때 이를 음운 변동이라 하며 우리말의 음운변동은 규칙적인 음운변동과 불규칙적인 음운변동으로 나눌 수 있다. 우리말의 경우, 문장을 소리나기 표기로 변환하는 알고리즘이 어느정도 확립되어 있으나[8],[9], 불규칙변동과 장단음에 처리는 아직 확립되어 있지 않은 실정이다. 따라서, 이러한 문제를 해결할 수 있는 발음사전에 대한 연구가 필요하다.

1.2. 운율 기호열

음성언어의 운율은 단어의 액센트, 문장의 구조 또는 화자의 감정 상태등과 같은 언어학적 요소에 의해 변화하게 된다. 이러한 운율의 특징을 표현할 수 있는음향학적인 요소로서는 기본 주파수(Fundamental frequency), 강세(Intensity) 또는 어절의 길이(Segmental duration)등이 있는데 이 가운데 기본 주파수는 음성 언어의 억양과 직접적인 관계가 있다[11]. 타 언어에서는 음운학적, 실험음성학적인 연구결과를 토대로 구문해석기법을 이용하여 입력 문장을 해석하므로서 음성합성을 위한 운율기호를 생성하는 처리 시스템이 구현되어 있으나[2][5], 우리말에는 이들에 대한 연구가 미흡한 실정이므로 자연스러운 한국어 합성음 생성을 위해서는 한국어에 대한 운율 기호 생성 처리 시스템의 구현을 위한 기초 연구가 시급하다.

운율을 표현하기 위한 음향학적 요소로서 일반적으로 다음과 같은 것들이 있다.

① 기본 주파수

합성음에 자연성을 주는 운율요소중 기본주파수의 패턴은 가장 중요한 요소이다. 영어[2]와 일본어(후지사키 모델)[5]의 경우, 기본주파수 모델이 거의 확립되어 있는 상태이다. 문장에 대한 피치 패턴은 각각의 호흡 단락 그룹(Breath group) 단위로 만들어 지고, 미리 정의된 길이 만큼의 휴지가 호흡 단락들 사이에 삽입되어 피치 패턴이 만들어 진다. 일본어의 후지사키 모델은 피치 패턴의 생성 모델로서 시간 t의 함수인 다음의 식으로 정의하고 있다[5].

$$\ln F_0(t) = \ln F_{min} + A_{u1}G_u(t - T_0) + A_{u2}G_u(t - T_3) + A_aG_a(t - T_1) - G_a(t - T_3)$$

$G_u(t)$는 구 제어지시(Phase control mechanism)의 임펄스 응답(Impulse response함수이고 $G_a(t)$는 액센트 제어 지시(Accent control mechanism)의 스텝 응답(Step response)함수이다.

여기서

Fmin: 액센트성분이 없을 때 F0의 점근치

A_{u1}: positive utterance command의 크기

A_{u2}: negative utternace command의 크기

A_a : 액센트 명령의 진폭

T_0: positive utterance command의 timing

T_3: negative utternace command의 timing

T_1:액센트 명령의 시작점

T_2: 액센트 명령의 끝

이다.

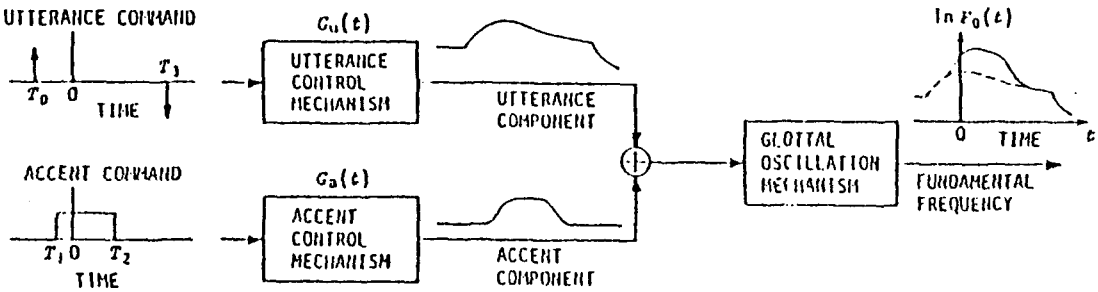

그림3. 일본어의 피치패턴을 생성하는 후지사키 모델

그림 3은 기본 주파수 윤곽(F0 contour)을 만들어내기 위한 후지사키 모델의 개념도이다. 위에서 살펴본 후지사키 모델은 영어에 적용된 바 있는데, 피치-액센트 언어인 영어나 일어의 경우 후지사키 모델은 잘 적용되는 것으로 알려져 있다[5]. 그러나 피치-액센트 언어라 하기 어려운 한국어의 경우에는 후지사키 모델이 잘 적용된다고 하기는 어렵다. 한국어의 억양에 관한 연구에서는 발음된 문장의 피치가 시간에 따라 최소값으로 나타나는 기저선(Baseline)이 한국어에서는 어떻게 나타나는가를 조사하였다[12].

한국어의 인토네이션을 규칙화하기 위하여 기초주파수의 하강을 조사한 연구에서는 기초주파수 패턴이 문장의 길이에 따라 크게 달라지지 않으며, 중문에서는 기초주파수의 리셑(Reset)현상이 나타나고, 주어나 동사가 생략된 경우에 그렇지 않은 경우와 다소 다른 피치 패턴을 보이며, 삽입절은 주절의 피치 패턴에 거의 영향을 미치지 않는다고 보고되었다[12].

② 타이밍 제어를 위한 패턴

자연스러운 리듬과 템포의 부여를 위해 연속 음성에 있어서 각 음운의 지속 시간을 조사하여 음성 합성에 적용할 필요가 있다. 이는 많은 기초 연구를 요하고 있으며 타언어에서는 이에 대한 연구가 행해져 모델화되어 음성합성에 적용되고 있는 실정이나 [13], 우리말은 아직 이에 대한 연구가 부족하다. 우리말에 대한 연구의 한 예로, 합성음의 타이밍을 제어하기 위해, 타이밍 제어의 기준이 되는 박동기(Beat duration)와 박동기점을 정의하고 발성하는 음절수와 박동기와의 관계를 자연음성을 분석 조사하여 합성음의 음절 지속시간을 제어하는 방법이 보고되었다[6].

이 연구에서는, 합성음에 자연스러운 템포와 리듬을 부여하기 위한 시도로서, 음절수가 같은 단어에 있어서 음절의 종류가 같다면 음절의 박동기가 거의 같음을 확인하였고, 같은 종류의 음절에 있어서는 음절의 수가 증가함에 따라 박동기가 감소하는 관계를 1차식으로 모델화하였으며, 타이밍의 기준점이되는 CV, V의 박동기의 위치를 설정하였다.

③ 휴지(Pause) 길이

단순한 묵음 구간 뿐만 아니라 말의 흐름이 끊긴다고 느껴지는 부분 또한 휴지라고 정의할 수 있다.문절의 접속도에 따른 휴지 패턴은 자연음성의 분석을 통한 모델링에 의하여 합성음에 적용할 수 있다.

한국어 산문 낭독을 통한 휴지 분석의 연구 결과, 묵음 구간 뿐만 아니라 길이 증가와 억양등도 휴지를 지각하게 하는 주요 요소이고, 하강조의 억양을 갖는 모든 문장과 단락 경계에서는 성질 변화가 관찰되며, 상위 경계로 올라갈수록 긴 묵음 구간, 하강 억양, 성질 변화, 길이 증가등으로 강하게 나타난다고 보고되었다[14]. 그러나 구구조에 따른 밀접도와 포우즈의 길이와의 관계를 정량적으로 표현할 수 있는 연구가 요구된다.

④ 에너지 패턴

에너지는 음성의 인지도에 영향을 미치는 intensity를 나타내는 요소로서 스펙트럼 에너지나 파형의 크기를 합산하는 방법등으로 추정할 수 있다[14]. 문장내에서 에너지의 흐름을 에너지 윤곽선이라 하며 문장의 운율 성분과 관련있다. 음절의 에너지는 앞부분이 크며, 자음보다 모음이 크다는 경향이 있다[14]. 운율에 영향을 미치는 하나의 요소로서

에너지 패턴은 우리말의 액센트와 강세에 대한 분석을 행하여 모델화되어야 한다.

이상에 소개한 운율패턴에 대하여 영어, 일본어등에서는 많은 연구가 되어, 그 결과가 음성합성에 적용되어 양질의 합성음을 생성하고 있스나, 우리말의 경우 아직은 미흡한 단계이므로, 더욱 섬세하게 제어할 수 있는 음운지속시간, 인토네이션, 휴지, 액센트등에 대한 물리적인 특성의 모델화에 대한 연구와, 구문과 운율기호와의 관계를 나타내기 위한 연구가 행해져야 한다.

2. 음성 합성 기능

음성 합성 기능부에서는 변이음과 장단음과 같은 음운 변동의 처리와 길이, 에너지, 억양, 휴지등의 운율 패턴에 따라 단위 음성 파라미터를 접속하여 조음 파라미터와 음원 파라미터를 생성하고 이로서 시변 필터를 구동하여 합성음을 생성한다.

2.1. 음운 생성

성도의 변동을 나타내기 위한 규칙을 줄이기 위하여 음성의 단편을 미리 저장하며, 저장된 단편들을 접속하여 음성을 생성한다. 이를 위하여 최적의 단위 음성 선택 문제가 해결되어야 하고, 또 운율 파라미터의 수정이 음질에 영향을 미치지 않도록 하며, 접속점에서 생기는 조음 파라미터의 불연속을 줄일 수 있고, 저장될 데이터 양을 효율적으로 감소 시킬 수 있는 기술이 요구된다.

표2.는 단위 음성 즉 접속단위의 종류와 단위음성의 길이에 따른 장단점을 나타낸다. 표2.로부터 음소의 경우 데이타 양은 적으나 인접하는 음 사이의 매끄러운 천이의 실현이 어렵다. 음절의 경우 데이타 양이 많고 변이음 처리와 타이밍 제어가 어렵다. 다이폰(Diphone) 또는 반음절의 경우는 데이타 양이 음소보다 많으나 인접하는 음 사이의 포만트 천이를 유지할 수 있으므로 접속점에서 불연속을 최소화할수 있어 가장 많이 사용되고 있다[7]. 한국어에 있어서 유성화하는 변이음의경우, 전후음의 영향이 강하게 나타난다. 이러한 면에서 볼때 변이음 처리를 위하여는 불연속성이 적고 포만트의 천이를 유지하고 있는 다이폰이 좋은 합성 단위라 생각된다. 최근 연구에서는 연결문맥의 전후를 고려하여 접속단위를 정의하는 COC(Context -Oriented Clustering)[3][4]와 CDU(Context-Dependent Unit) [9]를 합성단위로사용하여 합성음의 음질을 향상시킨 예도 있다. 이 합성단위는 양질의 합성음을 생성하나 단위음성 데이타베이스의 크기가 커진다는 단점이 있다. 그러나 하드웨어의 발달로 이는 쉽게 극복할 수 있다.

표2. 단위음성(접속단위)

단 위 길 이	음성 합성 단위	장 점
짧다	단일 핏치 파형 음 소 CV DYAD 반음절	크기가 작은 단위음성 집합 수정이 용이함
길다	VCV CVC CVCV 형태소 단어 구	조음결합 규칙 없이 고품질 음 성 생성 접속이 용이함

2.2 음성신호생성부

음성 변환방식 선택은 시스템 전체의 각 부분과 밀접한 관계가 있으므로 합성음의 음질이 좌우된다고 할 수있다. 현재 일반적으로 사용되고 있는 변환방식은 크게 3가지 1) 포만트(Formant)합성방식[2] 2)분석합성방식으로서 PARCOR방식, LSP방식과 켑스트럼 방식, 3) PSOLA (Pitch-synchronous overlap add)방식도 이용되고 있다[1][11].

· 포만트(Formant)합성방식

이 방식은 순수규칙합성 방식으로서, 영어에서는 실용화된 방식이다[2].성도의 변화/필터특성은 각 음소 그리고 음소간의 포만트변화를 나타내는 규칙을 사용하여 기술한다. 순수합성방식은 원리적으로는 이상적인 방식이지만, 조음파라미터계열을 생성하기 위한 규칙과 음성기본단위의 데이터를 얻기가 매우 어렵다. 한국어음성합성에 포만트 합성방식을 적용한 연구가 다소 있으나[10], 아직 우리말에 대해 충분히 분석하여 얻어진 데이터가 적기 때문에 한정된 단어를 합성하는 정도이며 음질 또한 좋지 못한 형편이다. 그러나 우리말의 음운특성에 대한 많은 연구가 행하여 질 때, 양질의 합성음의 생성이 가능하다.

· 음성분석계(Vocoder)를 이용한 방식

분석계(Vocoder)를 이용하는 규칙합성방식은 적어도 단위음성 데이터로서는 자연음성으로 부터 직접 구한 것을 사용할 수 있고, 파라미터 음성합성에서는 이미 실용화되고 있는 분석합성계를 이용할 수 있기 때문에 비교적 간단한 시스템으로 양질의 음성을 합성할 수 있다. 분석합성 방식에는 PARCOR과 LSP의 LPC계열과 켑스트럼등[1][7]이 대표적이다.

IV. 결론

본 고에서는 문음성 변환 시스템의 개요 및 규칙 합성 시스템의 부분 시스템인 언어 처리 기능과 음성 합성 기능에 대한 기술 및 과제를 소개하였다.

자연스럽고 명료한 음성을 합성하기 위해서는 먼저 음성 신호 생성기술 뿐만 아니라 운율 처리에 대한 연구가 자연 음성의 분석을 바탕으로 하여 이루어져야 한다. 그리고 운율기호로 부터 운율패턴 즉 피치 패턴과 음운지속시간, 에너지 패턴등을 생성하는 모델에 관한 기초적인 연구가 많이 수행되어야 할 것이다. 또한 텍스트로 부터 운율, 언어정보를 구하는 문제로서 기본주파수, 구경계, 액센트 위치등과 말의 부분, 어형변화의 타입과 형태, 구 의존 구조등에 대한 연구 역시 많이 수행되어야 할 것이다. 이는 신호 처리, 실험 음성학과 언어 처리등 제분야의 협조 체제가 절실하게 요구됨을 의미하는 것이다.

[참 고 문 헌]

[1] Edited by S. Furui & M.M. Sondhi,"Advanced in speech signal Processing", Dekker,(1992.).

[2] D. H. Klatt, "Structure of phonological rule component for synthesis -by-rule program",IEEE Vol.ASSP-24 No.5,pp.391-398, 1976.

[3] K.Takeda外, "選拓的に合成單位を用いる規則音聲合成", 電子情報通信學會論文誌(D- II), Vol. J73-D-II No.12, pp. 1945-1951, 1990年 12月.

[4] Y.Sagisaka外, "Concatenative Speech Synthesis By Minimum Distortion Criteria", IEEE. Trans. ASSP, pp.65-68, 1992.

[5] H. Fujisaki and K.Hirose,"Comparison of acoustic features of word accent in English and Japanese", J.Acoust.Soc.Jpn(E)7.1 pp.57-63. 1986.

[6] 이 양희," 한국어 음성의 규칙합성에 있어서 음절계속시간제어",음향학회 음성 통신 및 신호처리 workshop 논문집, pp118-123, 1989.

[7] 이 양희, "음성합성 기술 개발 현황 및 전망", 제9회 음성 통신 및 신호처리 워크샵 논문집, pp.88-93, 1992.

[8] N. K. Ha and S. R. Kim, "Development of Korean text-to-speech system", Korea-Japan joint Symposium on Acoustics, pp.261-267, 1991.

[9] 지민제외, "글소리II에서의 신호처리:PSOLA합성방식", 제1회 ETRI 음성,언어 및 음향정보처리 워크샵 논문집, pp.91-96, 1993.

[10] 조 철우외,"한국어 파열음의 규칙 합성을 위한 파라미터 추정에 관하여", 제 1 회 신호처리 합동 workshop논문집,pp.51-54, 1988.

[11] D. O'Shaughnessy, "Automatic Speech Sythesis", IEEE Communication magazine pp26-34, 1983.12.

[12] 김 진영외, "한국어 억양에 관한 연구", Korea-Japan joint Symposium on Acoustics, pp.292-297, 1991.

[13] 勾坂芳典, "音韻持續時間の制御と知覺", 日本音響學會, 1981年 8月.

[14] 지 민제외,"한국어 Pause Pattern의 음향음성학적 분석",음향학회 1990년도 음성통신 및 신호 처리 논문집, pp.169-171. 1990

[15] 星野雅孝, "單語中における單語の持續時間の決定のための定量的モデル",日本音響學會,1982.