

Text-to-Speech Conversion의 기술 동향

정 유 현*

I. 서론

인간의 가장 자연스러운 정보 수단인 음성을 인위적으로 만들려는 음성합성의 시도는 1779년 Krautzenstein의 기계적 음성합성이래 디지털 기술, VLSI 기술, 컴퓨터 및 주변 기술의 발전에 의해 급속하게 진전되어, 현재 전화를 이용하여 각종 정보의 안내, 예약등의 서비스를 제공하는 전기통신 분야와 가전, 완구 등의 각종 민생품 분야에서 실용화되고 있다. 현재 사용되고 있는 대부분의 음성합성 방식은 합성할 단어나 회화음을 전구간에 대한 frame마다의 파형(녹음편집방식) 혹은 음성 파형에서 분석 추출한 파라메터(분석합성방식) 등으로 미리 저장하고 있다가 필요에 따라 선택·편집하여 응답문을 출력하는 것이며, 출력 가능한 어휘는 미리 저장된 단어나 회화음으로 한정되어 있다.

최근 기계와 인간간의 정보전달 수단으로서 음성이 지난 장점때문에 음성합성 응용분야가 계속 확대되고 있으며, 그에 따라 인명, 지명, 회사명 등의 고유명사에 대한 음성출력 요구가 증가하고 있다.

목 차

- I. 서 론
- II. Text. 음성변환 기술
- III. Text. 음성변환의 응용분야 및 시스템의 예
- IV. 결 론

* 음향연구실 연구원

그러나 고유명사는 어휘수가 무한하여 녹음편집방식 또는 분석합성방식으로는 음성파일의 작성 및 변경에 따른 문제가 많고, 대용량의 기억장치가 필요하므로 비경제적이다. 따라서 이러한 문제를 해결하기 위한 방안으로 임의어 음성합성 기술에 대한 연구가 활발히 진행되고 있다.

임의어 음성합성 기술은 단어보다도 작은 음성 단위(음소, 단음절, VCV 음절 등)를 음성 파라메터로써 저장하고 있다가, 문자열이나 발음기호열이 입력되면 필요한 음성단위의 음성 파라메터를 접속규칙으로 결합하여 음성합성기에 의해 음성을 생성하는 “법칙합성 (Synthesis by Rule)” 연구분야에서 각종의 방식이 검토되어, 미국, 일본 등에서는 문자(Text)로 작성한 임의의 문장을 음성으로 변환하는 “Text. 음성변환기술(Text – to – Speech Conversion Technology)”에 대한 연구가 활발히 진행되고 있다.

Text. 음성변환은 문장을 이해해서 책을 낭독하는 인간의 복잡한 과정을 기계적으로 실현하는 것으로, 입력한 문자계열에 대응한 임의의 단어나 문 레벨의 음성합성이 가능하여 합성 어휘수에 대한 제한이 없다.

Text. 음성변환의 연구는 미국을 중심으로 시작되었으며, 1979년에 MIT 연구진에 의해 영어 Text. 음성변환 시스템 “MITalk” 가 발표된 것을 기점으로 연구가 활성화되었다. 현재 각국에서 자국어의 문자언어를 입력대상으로 하는 시스템이 상품화되고 있으나, 아직 음과 음의 연결이나 억양이 어색하여 합성음의 질이 좋지 못하다. 이를 해결하기 위한 방법으로 발음의 단위와 연결방법, 억양 부여 방법 등이 대한 연구되고 있다.

Text. 음성변환은 원리적으로 합성 어휘수가 무

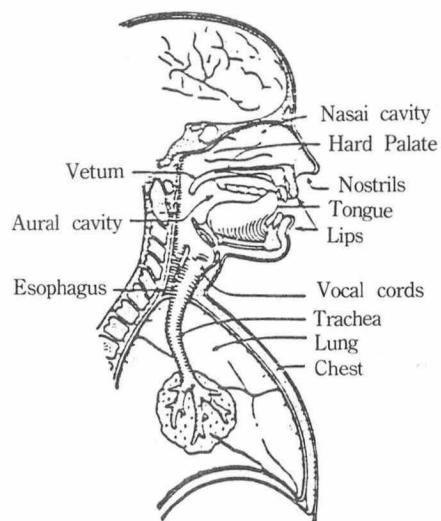
한하기 때문에 맹인용 독서기, 자동통역 전화기, Audiotex, 컴퓨터 음성출력 등 기계의 처리결과를 음성으로 출력하는데 필요한 기반기술로 앞으로의 기술발전이 크게 기대되고 있다.

본고에는 Text. 음성변환기술의 개요, 관련 연구 과제 및 최근에 상품화된 시스템에 대하여 정리하였다.

II. Text. 음성변환기술(Text – to – Speech Conversion Technology)

1. 인간의 음성생성 과정

음성생성은 말을 전달하기 위한 수단으로서 음성을 만들어 내는 것을 말하며, 그 음성을 구체화한 음성파형은 인간의 음성기관에 의해 생성된다. 음성기관은 음성 발생에 관여하는 기관을 총칭하며, (그림 1)과 같다



(그림 1) 음성기관

물리음향의 생성이론에서는 음성파형의 생성과정을 음원의 생성(generation of voice source), 조음(articulation), 방사(radiation)의 3단계로 생각하고 있다.

가. 음원의 생성

음원은 음성의 발성에 사용되는 주요한 energy 원으로 성대음원(glottal source), 난류잡음원(turbulent noise source), 파열음원(plosive source) 등이 있다.

성대음원은 성대 진동에 의한 준주기적인 조밀파이며, 모음 등의 유성음의 음원이다. 성대 진동주기에 의해 고저(pitch)가, 또 그 주기의 시간적인 변화에 따라 액센트나 억양감각이 부여된다. 난류잡음원은 성도(vocal tract)의 좁은 부분을 날숨(expiration)이 급속하게 빠져 나갈 때의 난류에 의한 잡음원으로 마찰음의 음원이다. 파열음원은 성도의 폐쇄가 급격히 개방될 때에 생기는 임펄스 음원으로 파열음의 음원이다.

나. 조음

음원에 의해 만들어진 음향energy(음파)가 성대에서 입술까지의 입안을 지나는 동안에 조음기관(articulatory organ : 혀, 입술, 턱 등)의 운동에 따라 변화하는 입안의 모양에 의해 주파수 별로 energy의 선택적인 강화(공명) 또는 억압(반공명)이 일어나, 고유의 음색을 지닌 음으로 형성된다. 이러한 과정을 조음이라 말한다.

조음기관중에서 자유롭게 움직일 수 있는 혀, 입술 등은 조음기(articulatory)라 하고, 고정된 부위

인 경구법, 이 등은 조음위치(place of articulation)라 한다.

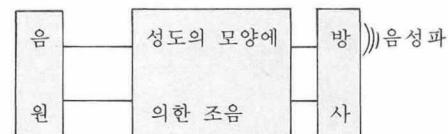
다. 방사

방사는 고유의 음색을 지닌 음으로 형성된 음파가 입술(콧구멍)에서 공중으로 전달되는 것을 말한다.

라. 음성

이상의 과정(그림 2)으로 생성된 음성은 모음(a, i, u, e, o), 유성자음(b, d, g 등), 반모음(j, w)등의 유성음과 무성자음(p, t, k)으로 분류하며, 음성에 의해 전달되는 정보로는 의도적 정보와 비의도적 정보가 있다.

의도적 정보는 주로 언어정보를 전달하는 것을 목적으로 하고 있으며, 음성의 음운적 내용을 담고 있는 음운성 정보와 문형의 강조(액센트, 억양), 희노애락 등을 표현하는 표현성 정보가 있다. 비의도적 정보는 말하는 사람의 사회적 환경(성별, 경력, 방언 등)과 생리적 제약(연령, 병리 등)에 대한 것을 포함하고 있다.

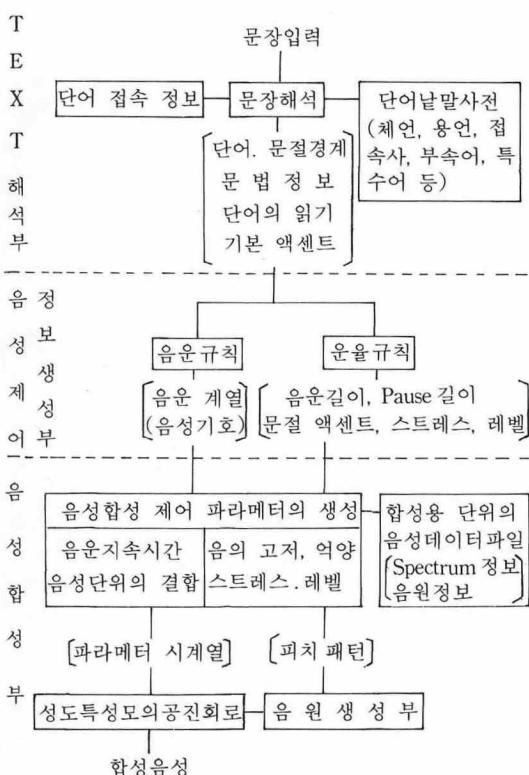


- 음의 고저
 - 음의 형성
 - 공중에의 전달
 - 유성/무성별
- (그림 2) 음성생성의 기본형

2. Text. 음성변환 개요

Text. 음성변환은 Text로 표현된 임의의 문장을

음성으로 합성하는 것이며, Text. 음성변환에 필요한 기술은 언어에 의한 처리상의 어려움이나 언어 특유의 문제점은 있지만 기본적으로는 같다. 따라서 본고에는 한자를 사용하고, 문법체계가 우리말과 비슷한 일본어를 입력대상으로 하는 일본어 Text. 음성변환 시스템(그림 3)을 예로 Text. 음성변환에 대해서 개략적으로 설명하고자 한다.



(그림 3) 일본어 Text. 음성변환 시스템의 구성

한자/가나로 작성된 일본어 문장이 입력되면은 문장해석부에서 일본어의 단어 낱말 사전과 일본어 문법을 반영한 단어간의 접속정보를 기초로 문장해석이 실시된다. 이 해석에 의해 어, 문절의 경계, 한자의 읽기, 액센트정보 및 문법정보(품사, 활용형)

가 결정된다.

문장해석부에서 얻어진 이러한 결과는 음성제어 정보생성부에서 음운규칙, 운율규칙이 적용되어, 음성합성에 필요한 각종의 제어정보가 작성된다. 음운규칙은 단어의 읽기를 기초로 계조사[は] 및 격조사[へ] [を]의 변환, 장음화변환, 모음의 무성화변환 등을 실시하여 음운 계열을 생성한다. 운율규칙은 단어. 문절경계 등의 문법정보를 기초로 각운의 지속 시간 길이, pause 길이, 한숨의 위치를 결정하고, 단어의 기본 액센트 등을 기초로 문절액센트 및 그것의 상대적인 크기를 표시하는 스트레스. 레벨 등의 음조제어 정보도 생성한다.

음성합성부에서는 생성된 음운계열에 따라 음성데이터파일에 일정한 단위로 저장되어 있는 음성신호의 특징량을 읽어내며, 음운길이 데이터를 근거로 합성단위의 신축, 결합을 행하여 성도특성을 표시하는 스펙트럼 파라메터의 시계열을 생성한다. 또, 음조제어 정보를 기초로 기본주파수의 시간변화패턴(pitch)이 생성되어, 음성 데이터 파일에서 얻은 음원정보와 함께 음원생성부에 전송된다. 음원생성부에서는 이러한 정보를 기초로 구동음원신호가 생성되어, 성도특성을 모의하는 공진회로에 의해 입력 Text에 대응하는 합성음성이 얻어진다.

3. 연구과제

a. 음성합성방식의 연구

합성의 기본단위를 연결하여 연속음성파형을 얻기 위한 기본방식의 연구를 행하며, 주요 연구과제는 다음과 같다.

- 고품질의 합성 음성을 얻기 위한 합성방식의 확립
- 연속음성을 합성할 때 기본이 되는 음성합성 단위의 연구
- 합성음성의 품질평가 방법의 확립

나. 합성규칙의 연구

임의어 연속음성을 합성하기 위해 필요한 합성 규칙의 연구를 행한다. 합성규칙에는 합성단위의 연결·변형에 관한 음운규칙(분절규칙)과 억양 등에 관한 운율규칙(초분절규칙)이 있으며, 주요 연구 과제는 다음과 같다.

- 연속음성에 있어서의 음운변형규칙과 조음결 합규칙의 확립
- 음소천이의 동적특성에 대한 분석과 합성단위 연결에 적용
- 언어정보를 충분히 활용한 운율정보의 자동 추출법
- 비의도적 요인에 의한 운율의 실현

다. 언어처리기술의 연구

문자나 문법·의미정보에서 음성을 합성하기 위한 과정에서 필요한 언어처리 방법을 연구하며, 주요 연구과제는 다음과 같다.

- 언어정보에서 운율레벨, 음운레벨의 정보 추출기술 확립
- 음성합성을 위한 언어처리계 구성법

III. Text. 음성변환의 응용분야 및 시스템의 예

1. Text. 음성변환의 응용분야

Text. 음성변환은 원리상 무한한 음성을 합성할 수 있기 때문에 많은 응용분야가 예상되며 〈표 1〉, 그에 따른 정보 산업의 과급효과도 크다.

〈표 1〉 Text. 음성변환의 응용분야

분야	실현 가능한 제품 및 기술 응용분야
통신 서비스	<ul style="list-style-type: none"> • 예약 시스템, 자동 통신 판매 • Audiotex
OA	<ul style="list-style-type: none"> • 음성출력 Word Processor • 음성출력 전자메일
FA	<ul style="list-style-type: none"> • 음성 로버트 • 음성 CAD
PA	<ul style="list-style-type: none"> • 음성백과사전 • 농아자용 독서기 • 음성사전
교육	<ul style="list-style-type: none"> • 음성입출력 CAI • 외국어 회화연습기 • 발성연습기
민생	<ul style="list-style-type: none"> • 회화형 자동판매기 • 음성에 의한 제품취급 설명
의료	<ul style="list-style-type: none"> • 음성건강 진단 • 음성구급처리 설명

2. Text. 음성변환 시스템의 예

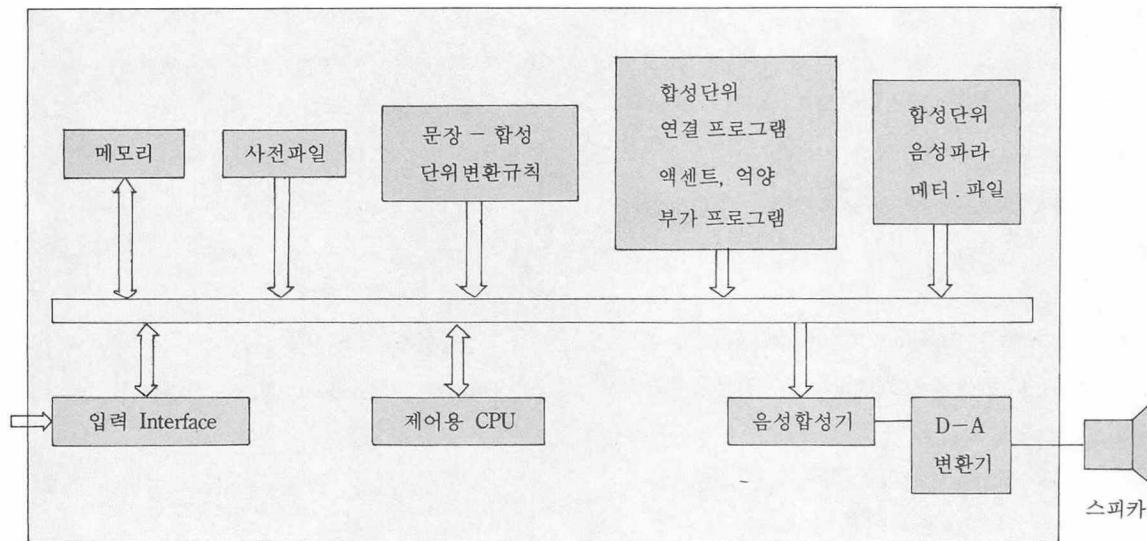
Text. 음성변환 시스템은 미국의 MIT에서 1979년에 개발한 MITalk 이래로 현재에는 다국어, 복수의 음색이 합성 가능한 장치가 제품화되고 있다. 〈표 2〉는 연구개발중인 시스템을 포함한 대표적인 Text. 음성변환 시스템을 나타낸 것이며, 기본 구성은 (그림 4)와 같다.

〈표 2〉 Text-to-Speech Conversion System 의 예

국 가	연구·개발기관	Model 명	대상언어	합성단위	합성방식
미	MIT	MITalk, Klattalk	영어	음소	Formant
	Bell Lab.	.	영어	dyad, demi-syllable	LPC
	Digital Equipment Corp.	DECtalk	영어	음소	Formant
	Speech Plus Inc.	Prose 2000	영어	음소	Formant
	Votrax Inc.	TYPE'N TALK	영어	음소	Formant
	Cornell 대학	.	영·일어	음소	Formant
	Texas Instruments	Magic Wand	영어	음소	LPC
	Street Electronics Corp.	ECHO-GP	영어	음소	LPC
	Ackeman Digital Systems Inc.	Synthetalker	영어	음소	Formant
국	Don't Ask Computer Software	Software Automatic Mouth	영어	음소	---
	First Byte Inc.	Smooth Talker	영어	음소	---
	Intex Micro Systems	Intex Talker	영어	음소	Formant
스웨덴	Infovox AB	SA201/PC SC2000	다국어	음소	Formant
서독	Ruhr 대학	SYNTEX	독어	음소	Formant
프랑스	X-Com	Dicton III	불어	음소	LPC
일본	NTT	.	일어	CVC	LSP
	東京 大學	.	일어	가나음절	Formant
	明治 大學	.	일어	가나음절	Formant
	와세다 대학	.	일어	가나음절	PARCOR
	九州藝術工科大學	.	일어	C V	LSP
	富士通	MB22441, MB22437	일어	가나음절	PARCOR
	日本電氣	PC6601	일어	CV, VC	FORMANT
	電氣總合研究所	.	일어	---	PARCOR
	東京三洋電機	VSS-100	일어	가나음절	PARCOR

현재 시판중인 시스템의 합성단위는 단일의 음소나 2~3개의 음운연쇄로 된 단위를 사용하고 있으며, 합성단위의 선택은 대상으로 하는 언어의 음운적 구조와 밀접한 관계가 있다. 영어에는 총

음절수가 12,000개 정도가 되므로 음소단위 혹은 음운과도부를 포함한 Dyad, Demi-syllable라는 단위가 사용되고 있다. 이러한 단위에 의해 합성을 행하는 경우 조음결합현상을 고려하는 규칙이 필



(그림 4) Text. 음성변환장치의 기본구성도

요하다.

이 규칙화를 피하고, 명료성이 좋은 자연적인 음성을 얻기 위하여 일본어에 있어서는 VCV(모음-자음-모음), CVC(자음-모음-자음) 등의 복합 음성단위를 사용하는 방법이 제안되고 있다. VCV는 전 모음의 정상부 후반에서 후 모음의 정상부 전반까지를 (CVC는 같은 방법으로 자음부분을) 접속하는 방법이다. 예를 들어 [sakura]는 VCV의 경우 /sa/ 와 /ku/ , /ra/ 로 CVC의 경우 /sak/ 와 /kur/ , /ra/ 로 분해된다. 일본어의 경우 VCV와 CVC의 음절의 개수는 각각 800개와 1,300개이고, CVC가 합성음질이 VCV보다 좋다.

합성단위의 음향적 표현과 합성방식으로는 임의 어의 합성을 행하는 과정에는 음색을 부여하는 스펙트럼. 파라메터와 피치 감각을 부여하는 음원신호를 분리하여 독립적으로 제어하는 방법이 요구 되기 때문에 포르만트 혹은 PARCOR, LSP로 대

표되는 선형예측분석에 의해 얻어지는 파라메터가 많이 사용되고 있다. 이러한 파라메터를 합성에 사용하는 경우 종래에는 구동음원으로 유성음부는 pulse 열로, 무성음부는 백색자음으로 모델화한 신호가 많이 사용되었지만, 최근에는 음성품질 향상을 목적으로 예측잔차신호를 구동음원으로 이용하는 방법 등 각종의 방식이 제안되고 있다.

IV. 결론

지금까지 임의의 문장을 음성으로 변환하는 Text. 음성변환에 대한 개요, 응용분야, 연구과제, 상품화된 시스템 등에 대하여 계략적으로 살펴보았다. Text. 음성변환은 음성합성의 최종적인 목표로 현재 실용화 단계에 들어서고 있으며, 합성음의 품질개선을 위한 방안으로 조음결합 현상의 분석과 규칙화, 운율제어 규칙의 고도화, 합성방식의 개선 등이 연구되고 있다. 이러한 Text. 음성변환의 연구성과와

병행하여 불특정 연속음성인식 연구의 진전에 의해
공상과학이나 영화 등에서나 가능한 “인간과 기계
간의 음성에 의한 자유로운 정보교환(Man-Ma-
chine Communication by Voice)”의 꿈이 조만간
실현될 것이다.

참 고 문 헌

- (1) Y.Sagisaka ; H.Sato, “Review of Text-to-Speech Conversion Technology”, 일본음향학회지, 41권12호, 1985
- (2) G. Kaplan, E.J.Lerner ; “Realism in Synthetic Speech”, IEEE Spectrum, 32 – 37, 1985.4
- (3) K.Nakata ; “On the Recent Trend of Speech Information Processing”, 일본음향학회지, 42권 12호, 1986
- (4) Naoki Ishii ; “Trends of Speech Synthesis Research”, 일본음향학회지 42권12호, 1986
- (5) Ichikawa ; “Speech Synthesis and Audio Response Equipment”, 정보처리, July, 1978
- (6) Hirokazu SATO 외 ; “日本語 Text からの音聲合成”, 연구실용화보고 제32권 제11호, 1983
- (7) 住永洋子 ; “注意の 文章を 音聲に 變換する 規則音號合成が 實用化”, NIKKEI ELECTRONICS, 1984.7
- (8) DECtalk A Guide to Voice, Digital Equipment Corporation
- (9) Flanagan, J.L ; Speech Analysis, Synthesis and Perception, Springer-Verlag, N.Y., 1972
- (10) 齊藤, 中田 ; 음성정보처리의 기초, OHM사, 1981
- (11) 中田和男 ; 음성, Corona사, 1977
- (12) Markel, J.D. and Gray, Jr. A.H. ; Linear Prediction of Speech, Springer Verlag, N.Y. 1976