다이폰단위의 합성방법을 이용한 오디오텍스 시스템의 구현에 관한 연구

이승훈, 강동규, 최준혁, 한민수
한국전자통신연구소 음성응용연구실

Development of a Diphone-Based Audiotex System

Siong-Hun Yi, Dong-Gyu Kang, Joon-Hyuk Choi, Minsoo Hahn
Speech Application Section
Electronics & Telecommunications Research Institute

요 약

당 연구실에서 개발했던 초기의 오디오텍스 시스템은 LSP 파라미터를 이용한 무제한 한국어 음성합성 장치로서 합성데이타베이스는 640개의 반음절로 구성되어 있었다. 그러나 이 시스템은 일반 사용자들에게 음성합성 서비스를 제공하기에는 음질이 너무 미흡하였으므로 음원모델의 수정, 에너지 contour의 조절등을 사용하여 어느 정도의 음질개선을 꾀하였으나 만족할 만한 수준에는 도달하지 못했다. 그래서 합성단위를 다이폰단위로 수정한 새로운 오디오텍스 시스템을 구현하였다.

다이폰단위의 오디오텍스 시스템은 한국어의 여러가지 음운환경을 고려하여 1228개의 합성단위로 구성되어 있으며 LSP 파라미터를 이용한 합성방식을 채택하고 있다. 또한 음원생성시 수정된 LF모델에 자음의 명료도 및 자연성을 높이기 위해 residual 신호를 추가한 혼합방식의 음원모델을 사용하였다. 그리고 실시간으로 음성을 합성하기 위해 TMS320C30 DSP chip, MC68020 CPU, 고속 메모리소자, 및 VRTX OS를 사용하여 시스템을 구현하였으며, 청취실험 결과 기존의 합성방법보다 자연성 및 명료도에서 개선된 음질을 얻을 수 있었다.

1. 서 론

컴퓨터의 보급이 늘어나면서 하이텔과 같은 정보통신 서비스망을 이용하여 서로 정보를 교환하고 필요한 자료를 검색하는 사용자들이 계속적으로 증가하고 있으며, 이에 따라 정보의 미디어변환 서비스인 문자/음성 변환기술에 대한 욕구도 계속적으로 확대되고 있는 추세에 있다. 이 기술은 주로 신문이나 소설 등을 읽어주는 서비스, 주식시세 서비스, 은행잔고조회 서비스 등의 분야에서 활용될 수 있으나 현재 국내의 합성기술이 상용서비스에 사용될 정도로 만족스러운 양질의 합성기를 만들기에는 많은 어려움이 있으므로 아직도 학계 및 여러 연구소에서 계속적으로 연구를 해오고 있다.

현재 개발중인 오디오텍스 시스템은 통신처리장치 내에서 미디어변환 서비스 기능을 수행하는 시스템으로서, 하이텔 서비스 망에서 일반 전화가입자에게 DTMF key를 이용하여 원하는 정보를 선택 및 검색하도록 하며 그 내용을 음성으로 변환하여 들려주는 서비스를 제공한다. 이 시스템은 초기의 640개의 반음절을 이용한 LSP 파라미터 합성방식을 탈피하고 새로운 합성단위인 다이폰을 이용한 합성 데이타베이스를 구성하여 보다 자연스럽게 음성의 천이구간을 표현할 수 있도록 하였다. [1] [2] 명료도 면에서는 residual 신호를 사용하여 자음을 합성함으로써 합성 데이타베이스의 크기는 늘어났지만 보다 정확하게 자음을 표현하는 잇점을 얻을 수 있었다.

2. 한국어 문자/음성 변환

2.1 합성 데이타베이스

규칙을 이용한 음성합성방식을 이용하여 인간에 가까운 음성을 만들어 내기 위해서 다양한 합성단위들이 연구되어 왔으며, 이러한 단위들은 주로 음소 또는 여러 개의 연속적인 음소로 구성이 되었다. 일반적으로 합성 단위의 크기가 커지면 커질수록 합성단위 간의 연결 규칙이 간단해지고 합성음의 음질이 양호하지만 무제한의 음소열을 만들어 내는데 필요한 합성단위의 갯수가 많아진다. 또한 이러한 합성단위는 앞이나 뒤의 음소, 액센트, 문장의 형태, 또는 문맥 등에 영향을 받아서 그 특성이 변하게 되므로 가장 효율적이고 적절한 합성단위 뿐만 아니라 필요한 합성단위의 갯수를 결정하는 것은 매우 어려운 일이다.

지금까지 연구 중인 합성단위 중 다이폰단위는 음

소 사이의 천이 구간을 합성단위 내에 포함시키고 있으므로 결합규칙은 복잡해지더라도 음소간의 연결이 자연스러우므로 오디오텍스 시스템에서는 이 단위를 채택하였다. 확보된 음성 데이타 베이스는 한국어에서의 대부분의 allophonic variation을 포함시키고 또한 길이 조정이 용이하도록 고안된 1228개의 다이폰으로 구성되었다. 이 음성 시료는 여자 아나운서로 하여금 발음하게 하여 DAT에 녹음한 후 16bit, 10KHz로 표본화한 후 12차 LSP파라미터, 피치, 에너지 조절을 위한 가중치, 및 양질의 자음의 합성을 위한 residual신호를 추출하여 합성 데이타 베이스를 구성하였다.

사용된 다이폰 종류는 표1과 같으며 각각의 단위는 한 음소의 안정구간의 중간에서 시작해서 다음 음소와의 천이구간을 포함하고 이 음소의 안정구간의 중간까지 포함하고 있다. 표2는 이 때 사용된 기호에 대한 약어표이다.

표3은 결합규칙을 나타낸 것으로서 '+' 기호는 기호 양측의 합성단위를 연결한다는 뜻이며 '( )' 표시는 합성단위 내에서 사용되는 부분을 의미한다. 이러한 다이폰의 결합은 안정 구간이므로 접속시 스펙트럼의 왜곡이 적을 뿐만 아니라 연음효과가 합성단위 내에 포함되므로 자연스러운 합성음을 만들어 낼 수 있다.

표1. 다이폰 종류

kind	number	kind	number
_V_	8개	_gV_	12개
_CV_	144개	_CgV_	162개
_VC_	140개	_VV_	64개
Vge	16개	_VCe_	152개
egV	12개	_eCV_	133개
eCgV	208개	_Vnt`a_	20개
amC!	8개	_Vnt`a_	20개
anC!	8개	_Vngt`a_	20개
angC!	8개	_V!t`a_	20개
alC!	7개	_V!le_	20개
_el!V_	20개	RPV	20개
RPC	4개		

표2. 사용된 약어표

C : Consonant	_ : silence	m : /ㅁ/
V : Vowel	a : /아/	n : /ㄴ/
g : Glide	e : /에/	l : /ㄹ/
j : Glide /이/	I : /으/	ng : /ㅇ/
w : Glide /오/	t` : /ㄸ/	
RPC:rising pitch	RPV:rising pitch	

표3. 결합 규칙

unit	concatenation rule
_V_	(_V)_+_(V_)
_gV_	(_gV)_+_(V_)
_VC_	(_V)_+_(VC_)
_CV_	(_CV)_+_(V_)
CgV	(_CgV)_+_(V_)
gVC	(_gV)_+_(VC_)
CVC	(_CV)_+_(VC_)
CgVC	(_CgV)_+_(VC_)
VCV	_(VC)e_+_e(CV)_
VCgV	_(VC)e_+_e(CgV)_
VV	_(VV)_
VgV	_(Vg)e_+_e(gV)_
_V	(_V)_
_gV	(_gV)_
_CV	(_CV)_
_CgV	(_CgV)_
V_	_(V_)
VC_	_(VC_)
VCiCjV	_(VC_)+(_CV)_ _(VC)t`a_+(_CV)_ _(VC)t`a_+_a(CC)1_+_e(CV)_
VCCgV	_(VC_)+(CgV)_ _(VC)t`a_+(_CgV)_ _(VC)t`a+_a(CC)1_+_e(CgV)_

2.2 음성합성

문자를 음성으로 변환하는 과정은 그림1과 같으며 크게 언어처리과정과 실제적인 음성을 합성하는 음성합성과정으로 구성되어 있다.

입력되는 완성형 문자는 3바이트의 내부 코드로 변환된 후 숫자, 약어처리, 발음규칙처리, 및 운율정보처리를 거쳐서 구분정보가 포함된 소리나는 형태의 발음기호열로 바뀌어 진다. 음성합성과정에서는 이 기호열을 이용하여 KLATT의 분절음 지속규칙을 사용한 음소의 길이조절을 수행한다.[3] 그 다음으로 표3의 결합규칙을 사용하여 합성데이타베이스로 부터 파라미터를 가지고 온 후 인접단위 사이에서 발생하는 불연속 현상을 제거하기 위하여 LSP파라미터의 평활화 처리 및 에너지 가중치를 이용한 에너지 조절을 수행함으로써 자연스러운 합성음이 되도록 한다.

기본주파수는 문장의 구조, 의미, 감정등에 대한 정보를 얼마나 잘 표현하는가에 있어서 중요한 요소인데 오디오텍스에서는 다음과 같은 이차함수 P(t)를 사용하여 구현하였다.[4] [5]

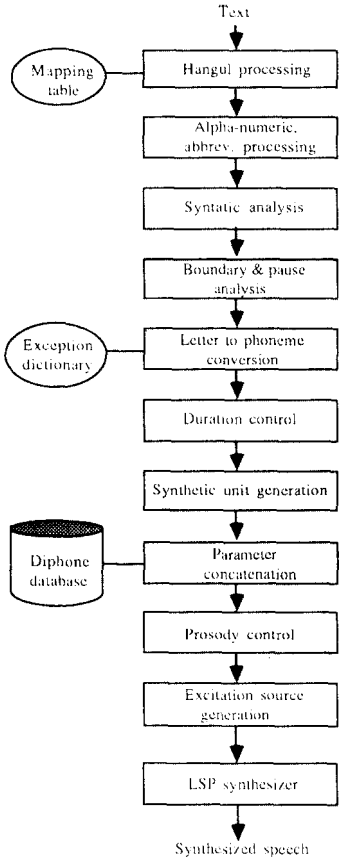

그림1. 문자/음성 변환 과정

$$P(t) = Pb - (Pb-Pa) * ((Tb-t)/(Tb-Ta))^{**2}, \quad Ta <= t < Tb$$
$$P(t) = Pc, \quad Tb <= t < Tc$$
$$P(t) = Pb - (Pb-Pd) * ((t-Tc)/(Td-Tc))^{**2}, \quad Tc <= t < Td$$

이 때 Pa, Pb, Pc, Pd는 상수이다.

그림2는 합성기에 사용된 음원모델로서 보다 자연스러운 음성을 합성하기 위해서 유성음의 경우 임펄스 음원 대신 수정된 LF모델을, 무성음의 경우에는 합성자음의 명료도를 확보하기 위해서 random noise 대신 residual신호를 사용하였으며 적용결과 만족할 만한 성능을 보였다. [6]

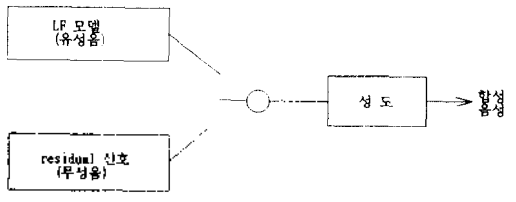

그림2. 음원생성모델

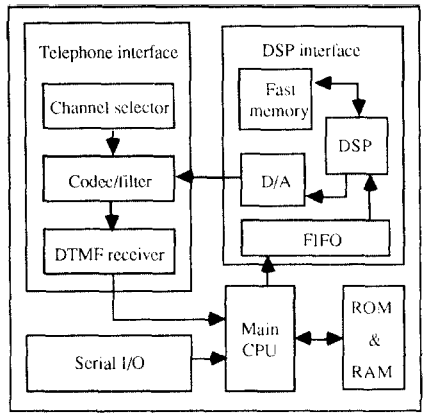

그림3. 오디오텍스 H/W 구조

3. 구현

3.1 시스템 H/W

오디오텍스 시스템은 실시간으로 문자를 음성으로 변환하기 위해 그림3과 같은 형태로 구성하였다.

시스템은 크게 나누어 보면 주 제어부, 전화기 인터페이스부, Serial I/O부, DSP 부로 구성이 되어있다. 주 제어부는 MC68020 CPU 및 각각 4Mbyte의 RAM과 ROM으로 이루어져 있으며 시스템 전체를 관리한다. 전화기 인터페이스부는 가입자 전화기와 시스템을 연결시켜주는 부분으로서 Hook on/off와 DTMF key 수신등을 담당한다. Serial I/O부는 MC68901 MFP와 Z8530 SCC로 구성이 되어 있으며 서비스망과의 데이타 송수신작업을 수행한다. 4Kbyte의 FIFO, 16Kbyte ROM, 128Kbyte RAM, 및 TMS320C30 DSP chip으로 구성된 DSP부에서는 주 제어부에서 생성된 합성파라미터를 FIFO를 통해 넘겨 받아 LSP 합성기를 구동시켜 합성음을 생성하는 기능을 담당한다.

3.2 시스템 S/W

S/W구조 역시 문자/음성 변환을 실시간으로 처리하기 위해 VRTX OS를 사용하여 구현하였으며 그 구조는 그림4와 같다.

각각의 제어 task는 task priority scheduling에 의해서 동작하며 mailbox, queue, 및 interrupt flag등을 사용하여 서로 정보를 주고 받는다. Hook task는 가입자로 부터 전화가 걸려오면 이를 감지하여 관련된 다른 task들을 생성시켜 합성을 수행하고, 서비스가 완료되면 생성했던 task들을 소멸시키고 초기 상태로 돌아간다. DTMF task는 서비스메뉴를 선택하고 정보를 검색하기 위해 누르는 DTMF key를 전화기 인터페이스부로 부터 수신하여 서비스 망

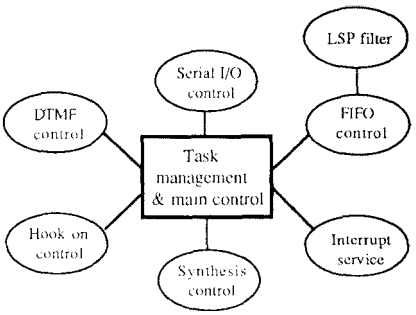

그림4. 오디오텍스 S/W 구조

에 전달하는 기능을 수행한다. Synthesis task는 서비스망으로 부터 입력받은 문자를 합성기에 사용할 수 있는 파라미터열로 변환하여 FIFO task를 통하여 DSP부에 넘겨주어 합성을 수행하도록 한다. DSP부에 내장된 LSP필터는 12차 all-pole모델로서 residual bit가 '1'이면 residual신호를 음원으로 사용하고, 그렇지 않으면 수정된 LF모델을 사용하여 합성을 수행한 후 매 10KHz마다 16bit D/A변환기를 통하여 합성음을 내보낸다.

4. 검토

초기의 반음절단위를 사용하였던 시스템에서 생성되었던 합성음은 파형의 처우침과 미음화등의 문제점을 안고 있었으며 지연성과 명료성이 매우 미흡하였다. 이러한 합성음질의 문제점을 해결하기 위해서 음원의 경우 임펄스모델을 수정된 LF모델로, deemphasis 계수의 변경, 유/무성음의 천이구간에서 mixed source의 사용, 및 에너지 contour의 적용등의 기법을 이용하였으나 커다란 성과는 얻지 못하였다. [7] 따라서 다이폰단위의 합성데이타베이스를 새로이 구성하여 오디오텍스 시스템에 적용함으로써 반음절단위에는 없었던 음소의 천이구간의 자연스러운 합성, 다양한 음운환경의 표현이 가능하였으며 residual신호를 사용함으로써 자음의 합성시 명료도를 높일 수 있었다. 합성음의 전체적인 지연성 및 명료성의 개선은 실험실에서 청취실험을 통하여 확인 할 수 있었으나 지금보다 더욱 다양한 allophonic variation의 표현, 다양한 화자의 데이타베이스 구성, 훨씬 더 강력한 언어처리기를 이용한 구문정보 생성 및 정확한 분석을 통한 최적의 합성파라미터 추출등의 분야에서 아직도 많은 점을 개선해야 할 것으로 본다. 만약 이러한 부분들을 개선한 다면, 오디오텍스 시스템을 실 상용화하는데 한걸음 더 나아 갈 수 있을 것으로 본다.

5. 결론

본 논문에서는 다이폰단위를 이용한 한국어 문자/음성변환 시스템인 오디오텍스의 구현에 대해서 설명하였다. 이 시스템은 일반 전화가입자들에게 하이텔서비스를 제공하기 위한 것으로서 합성 데이타베이스는 1228개의 다이폰으로 구성되어 있으며 유성음을 합성하는 경우에는 수정된 LF모델로, 무성음인 경우에는 residual신호를 사용하여 LSP합성방식으로 합성하도록 구현하였다. 또한 고속연산 및 실시간 처리를 위해 TMS320C30 DSP chip을 사용하여 LSP합성필터를 구현하였으며 VRTX OS를 사용하여 multi tasking이 가능하도록 하였다. 이 시스템은 기존의 반음절방식에 비해서 자연성 및 명료도 면에서 개선된 것을 청취실험을 통하여 확인 할 수 있었으나 통신처리장치 내에서 상용 서비스를 제공하기에는 아직도 많은 점이 미흡하다고 판단되며 운율조절규칙의 세분화, 합성 데이타베이스의 보강 및 최적의 합성파라미터 추출 등에 더욱 노력을 기울인다면 앞으로 좋은 결과를 얻을 수 있으리라 생각된다.

참 고 문 헌

1. E.Kim et.al, "Implementation of Audiotex System: Korean Text-to-Speech Synthesis," Proc. of ICSPAT'92, pp.34-43, Nov. 1992.

2. J.Lee, "Study of Korean Text-to-Speech Synthesis Based on LSP Representations," Proc. of Workshop on Speech Comm. and Signal Proc., Aug. 1990, in Korean.

3. J.Allen, M.S.Hunnicutt, and D.Klatt, *From text to speech : The MITalk system*, Cambridge University Press, 1987.

4. J.C.R. Licklider, "Effects of differentiation, integration and infinite peak clipping upon the intelligibility of speech," J. Acoust. Soc. Am, pp.42-51, Jan. 1948.

5. A.Waibel, *Prosody and Speech Recognition*, PITMAN, 1988.

6. A. Lalwani, "Experiments on LF Model," Ph.D. Dissertation, Univ. of Florida, 1990.

7. D.G.Childers, M.Hahn, and J.N.Larar, "Silence and voiced/unvoiced/mixed excitation(four-way) classification of speech," IEEE Trans. Acoust., Speech, Signal Processing, Vol.ASSP-37, pp.1771-1774, Nov.1989.