1999년도 한국음향학회 학술발표대회 논문집 제18권 제2(s)호

PC 용 Text-to-Speech 시스템 개발

최무열, 황철규, 김순태, 김정곤, 이서배*, 장석복*, 표경란*, 안혜선*, 김형순
부산대학교 전자공학과
*부산대학교 인지과학협동과정

Development of Text-to-Speech System for PC

Muyeol Choi, Cholgyu Hwang, Soontae Kim, Junggon Kim,
Sopae Yi*, Seokbok Jang*, Kyungnan Pyo*, Hyesun Ahn*, Hyung Soon Kim
Dept. of Electronics Eng., Pusan National University
*Dept. of Interdisciplinary research program of cognitive science, Pusan National University
E_mail : {shallom,kimhs}@hyowon.pusan.ac.kr

요 약

본 논문에서는 PC 응용을 위한 고음질의 한국어 text-to-speech(TTS) 합성 시스템을 개발하였다. 개발된 시스템의 합성방식으로는 음의 고저 조절, 인접음 사이의 연결 처리 및 음색제어 등에서 기존의 PSOLA 방식에 비해 장점을 가지는 정현파 모델 기반의 방식을 채택하였고, 자연스러운 운율 모델링을 위하여 통계적 기법중의 하나인 Classification and regression tree(CART) 방법을 사용하였다. 또한 음소 경계의 불연속성 문제를 줄이기 위한 합성단위로 초성-중성 및 종성 단위를 사용하였고, 다양한 음색표현이 가능하도록 음색제어 기능을 갖추었다. 그리고, 표준 Speech Application Program Interface(SAPI)를 준용한 TTS engine 형태로 구현함으로써 PC 상에서의 응용 프로그램 개발 편의성을 높였다. 합성음의 청취평가 결과 음질의 우수성 및 음색제어 기능의 유효성을 확인할 수 있었다.

1. 서 론

최근 국내에서도 다수의 TTS 시스템이 상용화되어 많은 응용 분야에 활용되기 시작하고 있다. 그러나 TTS 기술이 더 널리 보급되기 위해서는 합성음의 명료성과 자연스러움의 개선과 더불어 다양한 음색표현 등의 부가적인 기능들이 요구되고 있다. 본 논문에서는 차세대 TTS 시스템의 요구조건을 고려하여, 기존의 합성방식에 비해 고음질 생성과 음색제어 등에 장점을 가지는 정현파 모델에 기반을 둔 TTS 시스템을 구현하였다. 개발된 시스템은 통계적 방법에 의한 운율 모델링 방식을 채택하였으며, 다양한 음색을 표현하는 음색 제어 기능도 가지도록 하였다.

본 논문의 구성은 다음과 같다. 서론에 이어 2 장에서는 음성 합성 시스템의 구성에 대해 기술하며 3 장에서 실험 및 결과를 다룬 후 4 장에서 결론을 맺는다.

2. 구현된 TTS 시스템의 구성

본 논문에 구현된 TTS 시스템의 구성은 그림 1 과 같이 크게 언어학적 처리부와 운율 제어부, 그리고 음성신호 처리부의 세 부분으로 나눌 수 있다.

2.1 언어학적 처리부

언어학적 처리부는 입력된 문장을 분석하여 여러 가지 언어학적 정보를 생성하는 단계로서 일반적으로 text 전처리, 발음표기변환, 문장 분석의 세 과정을 거치게 된다. text 전처리 단계에서는 입력된 문장에서 한글이 아닌 한자, 숫자, 영어, 특수문자 및 기호등을 적절한 한글로 바꾸고, 발음표기변환 단계에서는 문법에 맞게 입력된 문장을 실제 발음되는 형태의 음소열로 변환하게 된다.

문장 분석 과정에서는 운율 생성에 필요한 정보를 만들기 위해 입력 문장의 품사 정보 및 문장 분석 정보를 생성한다. 언어학적 처리부의 결과는 발음표기 변환

된 음소열에 품사 및 문장 분석 정보가 부가된 형태로 나타난다.

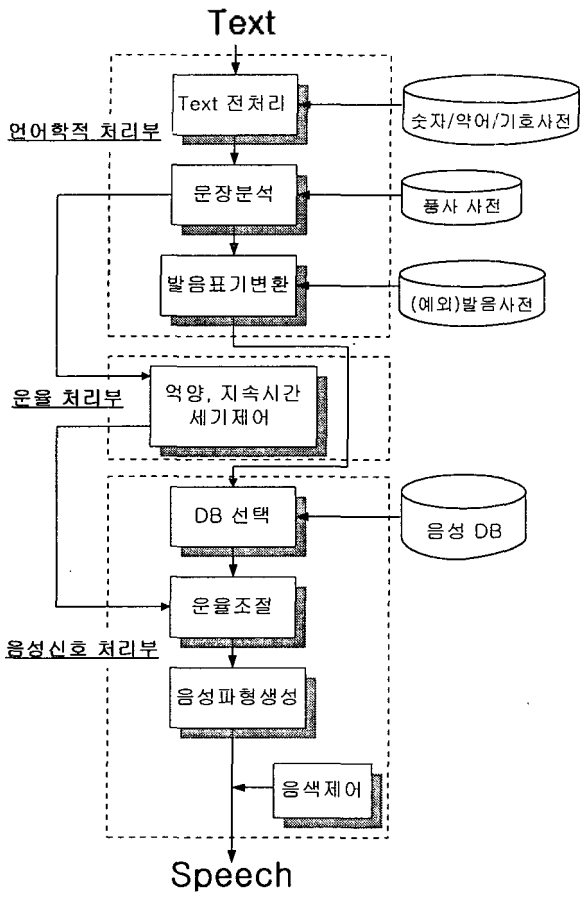


그림 1. TTS 시스템 구성도

2.2 운율 제어부

2.2.1 지속시간(Duration) 모델

음소의 지속 시간은 여러 가지 문맥 정보에 의해서 변화하므로 단순한 제어 규칙에 의존하기 보다, 방대한 데이터베이스를 이용하여 통계적인 기법으로 음소의 지속 시간에 변화를 주는 요인을 찾아내려고 하는 것이 지금의 추세이다. 본 연구에서도 트리 기반 모델링 방법중의 하나인 CART 방법[1]을 사용하여 회귀 트리를 생성하고, 생성된 트리에 기반하여 음소의 지속 시간 예측 모델과, 자연스러운 끊어 읽기를 위한 휴지 기간 예측 모델을 구현하였다. 음소의 지속 시간 예측 모델에서는 음성코퍼스와 문서코퍼스에서 분석한 22 개의 파라미터를 사용하여 음소 문맥을 고려한 회귀 트리를 생성한 후, 각 트리의 단말노드의 평균값으로 모델링하였다. 휴지기간 예측 모델은 음성코퍼스에서 추출한 데이터의 분산도가 커서 회귀 트리만으로는 정확하게 모델링 되지 않았기 때문에, 품사 bigram 을 사용하여 후처리를 하였다.

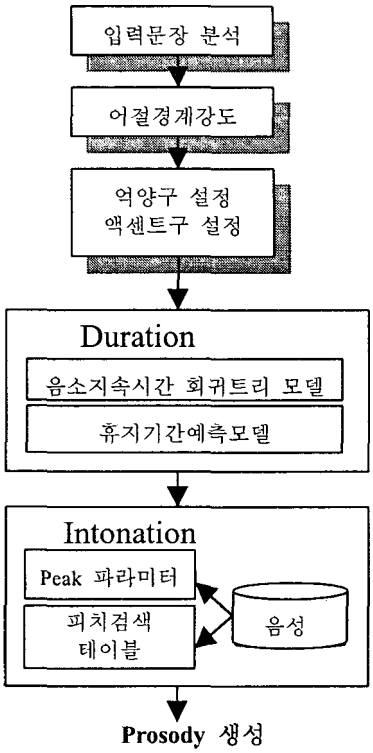


그림 2. 운율제어부 구성도

2.2.2 억양(Intonation) 모델

TTS 시스템에서 사용할 억양 모델을 위해 음성 코퍼스(corpus)에서 모델 파라미터와 피치 검색테이블을 추출하여 미리 구성하고, 합성시에는 이를 추정하여 최종 F0 값을 생성하는 자료기반 접근방식(data-driven approach)을 사용한다.

어절 경계강도(break-index)는 경계강도의 특성에 따라 고정적 경계강도와 가변적 경계강도로 세분화하여 사용하였고, 예측된 경계강도를 기준으로 억양구(Intonation Phrase)와 액센트구(Accentual Phrase)를 설정하였다. 특히, 액센트구 모델은 인지적, 음향적으로 중요한 정점(peak)을 정확하게 모델링하는 것에 주안점을 두어 정점의 시간축, 주파수축 값과 이를 기준으로 앞뒤 기울기를 추정하여 4 개의 파라미터로 설정[2]하였고, 이 파라미터들은 CART 를 이용하여 예측규칙을 만들었다. 경계음조가 나타나는 조사, 어미는 정규화된(normalized) 피치값과 key-index 로 구성되는 검색테이블을 만들어 보다 정교하게 피치값을 예측하였다.

2.3 음성신호 처리부

2.3.1 음성 합성 단위(DB)

본 시스템에 적용한 합성 단위로는 초성과 중성(C^iV) 형 또는 중성(V)형 및 종성(C^f)형의 단위를 사용하였다. 이와 같이 변형시킨 음절을 합성 단위로 하여 DB를 만들 때 C^iV 형, V 형, 그리고 C^f형에 미치는 앞뒤 음소의 조음현상을 고려하여 그 현상을 분석하면 유사음소를 묶을 수 있게 된다. 이렇게 묶은 유사음소 그룹은 C^iV 형 뒤에 13 개, V 형 뒤에 6 개, 그리고 C^f형 뒤에 3 개씩 두었다. 이렇게 되면 분류 범위의 확대에 따른 합성 DB 의 증가 문제를 어느 정도 해결할 수 있다. 반면에 C^f형 중에 불파음 /ㄱ/,/ㄷ/,/ㅂ/은 분리하지 않고 C^iV 형에 붙임으로써 하나의 음절 형태를 취하여 명료성을 높였다. 이와 같은 분류 방법은 합성 단위간의 반복적인 실험과 청취과정을 통해 마련한 규칙에 근거하였다.

이와 같은 분류 방법에 의해서 구해진 합성단위의 개수는 모두 6,080 개로 표 1 과 같다.

표 1. 합성단위 개수

C^iV 형	V 형	C^f형
5,173	816	91

또한 음성 데이터베이스를 효율적으로 압축함으로 DB size 의 최적화를 이루었다. 여성 합성 DB 의 경우 사용된 PCM file size 의 50~60% 정도인 15Mbyte 의 파일 크기를 가지며 압축에 대한 계속적인 연구를 진행 중이다.

2.3.2 정현파 모델 기반의 합성방식[3][4]

정현파 모델은 각기 다른 주파수, 진폭 그리고 위상을 가지는 정현파들의 합으로 음성신호를 모델링하는 방법으로서, 음성신호를 다음 식(1)과 같이 시간에 따라 크기가 변하는 L 개의 정현파들의 합으로 추정한다.

$$s(n) = \sum_{l=1}^{L} A_l(n)\cos(\omega_l n + \theta_l) \qquad (1)$$

여기서 $A_l(n)$은 시간에 따라 변하는 각 정현파의 진폭이고 ω_l와 θ_l는 주파수와 위상이다. 이 방법은 각 정현파들의 주파수, 진폭 및 위상을 변경시킴으로써 파라미터 조작을 용이하게 할 수 있고, 고속 Fourier 변환(FFT) 등의 빠른 알고리즘을 사용하여 계산량 문제도 어느 정도 극복할 수 있다.

음성신호로부터 각 정현파들의 파라미터들을 추출하기 위한 방법으로는 이산 Fourier 변환(DFT)을 이용한 스펙트럼 영역에서의 peak 추출 방법과 Analysis-by-Synthesis Overlap Add(ABS/OLA)방법이 있는데, 후자의 경우가 성능면에서 더 우수한 것으로 알려져 있다. 본 논문에서는 ABS/OLA 방법을 사용하였다. ABS/OLA 방법에서는 원음 $x[n]$과 추정된 음 $\overline{x[n]}$ 사이의 최소 자승 오차(MSE)를 최소화하도록 하는 식(1)의 파라메터를 구한다. 그림 3 은 ABS 방법의 블록도 이다.

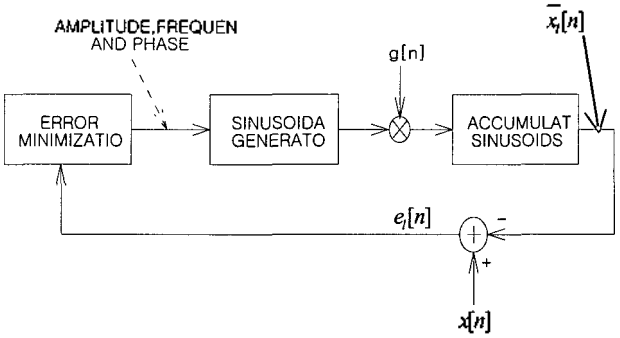


그림 3. Analysis-by-synthesis 방법

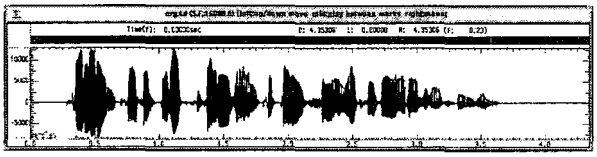
이렇게 추출한 파라미터를 가지고 연결합성 방법을 통하여 합성음을 생성하게 된다. OLA sinusodal model 을 기반으로 하여 서로 다른 음성으로부터 발화된 세그멘트들을 연결할 때 그 둘 사이에는 시간영역 혹은 주파수영역에서의 신호 특성에 커다란 차이가 존재하게 된다. 그러므로 두 세그멘트의 경계 근처에서 정현파들을 적절히 조절하여 연결부분에서의 pitch, phase, energy 그리고 spectral envelop 불일치를 최소한으로 줄여야 한다[5].

여기서는 합성 프레임의 중앙에서 기본 주파수와 phase 가 일치하도록 한 후 시간 영역에서 두 프레임 사이의 cross correlation 이 최대가 될 때 중첩 가산 함으로서 서로 다른 음성의 발화에서 얻은 두 프레임 사이의 phase mismatch 를 제거 하는 방법을 사용하였다. DB 전체에 대한 에너지 정규화와 더불어 세그먼트 경계에서 생길 수 있는 잡음 제거를 위해 세그먼트 경계에 있는 프레임들에 대해서 energy 및 spectral envelope 를 smoothing 하였다.

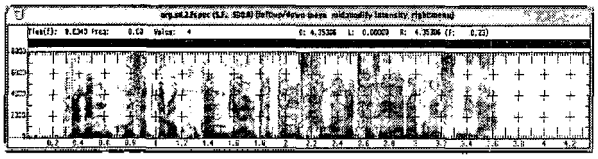
이러한 정현파 모델 기반의 합성방식은 개별적인 harmonic 들을 제어함에 의해 원래의 합성음과는 다른 음색을 지닌 합성음을 만들 수 있는 추가적인 장점이 있다.

3. 결과 및 고찰

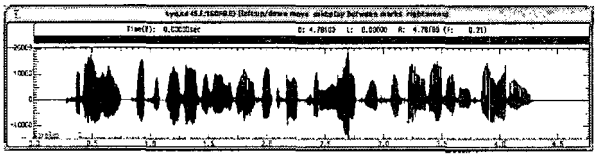
본 시스템으로 생성시킨 합성음을 동일한 문장의 자연음과 비교한 결과 그림 4 에서 보는 바와 같이 자연음과 유사한 결과를 얻을 수 있었다. 비공식적인 청취 실험 결과, 명료성 면에 있어서는 현재 최고 수준의 상용 시스템들에 비해 아직 뒤떨어지며, 이를 보완하기 위한 추가적인 DB tuning 이 진행중이다. 합성음의 자연성의 측면에서는 일기예보, 교통안내 등의 평가 문장에 대해 자연스럽다는 평가가 우세했다.



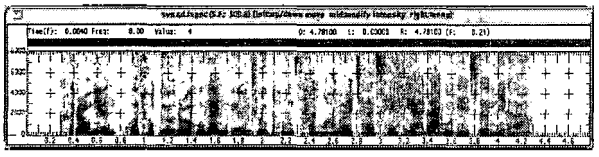
(a) 자연음 파형



(b) 자연음 스펙트로그램



(c) 합성음 파형



(d) 합성음 스펙트로그램

그림 4. 문장 "그러면 고속도로 구간별 교통상황을 자세히 알아보겠습니다" 에 대한 결과

본 시스템의 특징인 음색변환 부분에서는 정현파 모델 기반의 다양한 파라메터 제어를 통해 여러 가지 음색표현이 가능함을 확인했으며, 이로써 단일 합성 DB 로 여러 화자의 음성 DB 를 가지고 있는 효과를 얻을 수 있었다.

4. 결 론

본 논문에서는 고음질의 음성합성 및 다양한 음성변환의 유연성을 위해 정현파 모델에 기반한 합성방식을 이용하여 TTS 시스템을 개발하였다. 음소 지속시간 및 휴지기간은 음성 corpus 를 기반으로 CART 방법을 이용하여 모델링 하였고, 억양모델은 인지적으로 중요한 정점(peak) 파라미터와 경계음조를 나타내는 피치검색 테이블을 작성하여 생성함으로써 자연성의 향상을 이루어낼 수 있었다.

이 시스템은 15 Mbytes 의 DB 크기로 자연스러운 합성음을 생성해 낼 수 있으며, 특히 단일 화자의 합성 DB 만으로도 여러 화자의 음색을 표현할 수 있도록 음색제어 기능을 갖추었다. 그리고, PC 상에서 최소한의 계산량으로 음성합성을 수행하도록 알고리즘을 최적화 시켰으며, PC 상에서의 다양한 응용 프로그램에 쉽게 적용시킬 수 있도록 Microsoft 사의 SAPI 규격에 맞는 인터페이스를 구현하였다.

현재 합성음의 음질 개선을 위한 합성단위 및 운율의 tuning 작업이 진행되고 있으며, PC 이외의 응용분야를 고려하여 메모리 용량 및 계산량 면에서 scalable 한 합성시스템의 구축도 검토중에 있다.

이 논문은 산업자원부가 지원하는
산업기반기술개발사업 과제의 결과입니다

참 고 문 헌

[1] M.D.Riley, "Tree-based modelling of segmental duratuion," *Talking Machines : Theories, Models, and Designs*, G. Bailly, C. Benoit, and T.R. Sawallis, editors, pp.265-273, Elsevier Science, 1992.

[2] B. Heuft, T. Portele, "Synthesiziong prosody: a prominence-based approach," in proc. ICSLP'96,pp.1361-1364,1996.

[3] M. W. Macon and M. A. Clements, "Speech concatenation and synthesis using an overlap-add sinusoidal model," *in Proceedings of ICASSP*, Vol. 1, pp. 361-364, 1996.

[4] E. B. George and M. J. T. Smith, "Speech Analysis/Synthesis and Modification Using an Analysis-by-Synthesis/Overlap-Add Sinusoidal Model, " *IEEE Trans. on SAAP*, Vol. 5, No. 5, pp. 389-406, 1997.

[5] 구자형, 최무열, 김형순, "Analysis-By-Synthesis/Overlap-Add(ABS/OLA) Sinusoidal Model 을 이용한 음성변환과 연결음성합성," 제 15 회 음성통신 및 신호처리 워크샵(KSCSP '98 15 권 1 호), pp.339-342.