한소리 : 무제한 음성합성시스템

○
김응인, 김재인
한국통신 소프트웨어연구소 자동통역연구팀

HanSoRi : an Unlimited Synthesis System

○
Eung-In Kim, Jae-In Kim
Automatic Interpreting Telephony Team, S/W Research Laboratories, Korea Telecom

요 약

본 논문에서는 무제한단어 음성합성 시스템인 한소리에 대해서 간략히 기술하고 청취실험을 통한 성능평가에 대해서 논한다. 음성합성시스템의 음질을 결정하는 주요 요소들은 합성의 기본단위, 합성방법, 음운학적 전처리방법 및 운율조절방법이다. 한소리 합성시스템은 반음소를 음성합성의 기본단위로 하고, 형식형태소를 이용 음성학적 전처리를 실행하며, 개선된 한국어 운율조절방법이 적용되고, 음성단편 조합 방식을 합성방식을 사용한다. 청취실험결과 매우 한소리 합성시스템의 합성음이 자연스러움을 알 수 있었다.

1. 서 론

인간과 기계와의 대화에 있어서 기계로 하여금, 인간의 가장 자연스러운 통신수단인, 말을 하게 하는 음성합성에 관한 연구가 국내외에서 꾸준히 연구되어 왔다. 국내에서는 현재 상용화되어 있는 합성시스템으로 디지콤의 가라사대가 있고, 연구소에서도 몇몇 합성시스템을 개발하였으나, 그 음질에 있어 아직까지는 명료성(intelligibility)은 있으나, 자연스럽지 못한 점이 많다.

이 자연성(naturalness)의 미확보로 음성 정보 검색 시스템이나 700번 서비스에서는 안내음으로 합성음을 이용하지 않고 양질의 음질을 가진 발성자의 음성을 녹음 편집하여 사용하고 있는 실정이다. 합성음의 명료성과 자연성을 좌우하는 가장 중요한 요소로서는, 해당 언어에 대한 지식을 기본으로 한 음성학적 전처리와 인간의 발성에 기초한 운율정보, 및 합성방법과 합성의 기본단위가 있다.

음성학적 전처리 부분에서는 자연어처리(natural language processing) 기법을 적용하는 방법이 대종을 이루고 있으나 실시간 동작과 저장용량의 제약을 극복할 수 있는 방법은 아직 개발되지 않고 있다. 운율(prosody)부분에 있어서도 한국어 발성의 고유한 운율에 대한 연구 결과물이 거의 없는 상태에서 영어나 일본어의 운율 패턴을 그대로 또는 변형하여 적용하고 있는 실정이다. 한편, 합성 방식에 있어서는 기존의 LPC 계열의 보코딩 방식과 원음성에 기초하는 PSOLA 방식으로 대별할 수 있으나, 보코딩 방식을 쓸 경우에는 피치 패턴과 음성단편의 지속시간을 임의로 변화시킬 수 있고, 스펙트럼 포락(spectral envelope) 추정 파리미터와 내삽에 의해 음성단편간을 매끄럽게 접속시킬 수 있는 장점은 있으나, 보코딩 방식으로 음성을 코딩할 경우에는 재생되는 음성의 음질이 낮기 때문에, 저장된 음성단편을 디코딩하여 연쇄(concatenation)시킴으로써 만들어지는 합성음도 이들 보코딩 방식이 가지는 음질의 한계 이상 좋은 음질을 가질 수 없다. 원음성에 기초한 방법으로 PSOLA(pitch synchronous overlap and add)방식[3]이 있으나, PSOLA방식은 인접주기파형의 영향으로 인해 근사적인 스펙트럼 포락이 추정되며, 음소 또는 음절간의 스펙트럼의 왜곡으로 인한 음질의 열화를 해결하기가 매우 어렵다. 또한 음성 합성의 기본단위로 주로 많이 이용되고 있는 것은 다이폰이나 변이음 등이 있으나 음소간의 조음결합현상을 적절히 표현하지 못하는 경우가 있다.

본 논문에서는 지금까지 논한 무제한단어합성시스템의 문제점들을 극복하고 좀더 자연스러운 합성음을 만들기 위해 개발된 한소리 합성시스템에 대해서 논한다.

2. 한소리 시스템의 개요

시스템은 그림 1에서 보여주는 바와 같이 음성학적 전처리부, 운율발생부 그리고 합성부의 중요한 세 부분으로 나누어진다. 음성학적 전처리부에서는 과일이

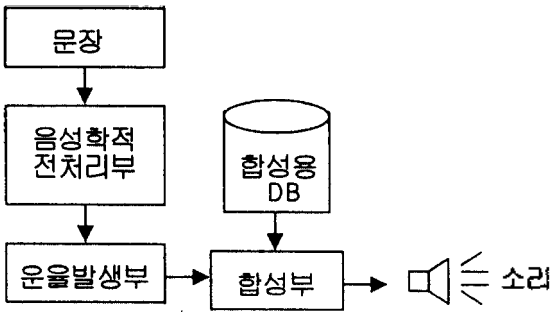


그림 1. 한소리 시스템의 구조

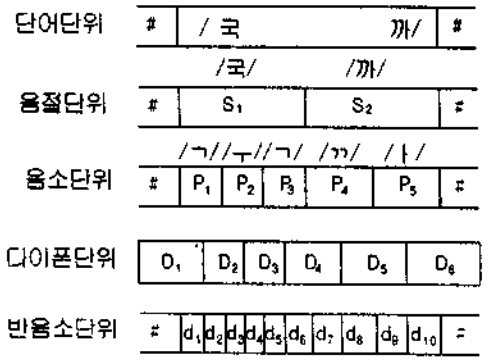


그림 2. 반음소의 개념 설명도(#는 휴지기호임)

나 키보드입력에 의한 문장을 성분분석하여 음운변동을 수행하고 운율정보발생을 위한 기본적인 구문분석을 수행한다. 다음 운율발생부에서는 전처리의 결과를 받아 한국어에 적합한 억양, 길이, 세기 등의 운율을 발생시킨다. 합성부에서는 합성합성의 기본단위를 가져와 연결시키므로써 문장을 만들어내는데, 운율구현 및 음소간의 interpolation을 동시에 수행한다.

2.1 합성 단위 (Synthesis Unit)

본 시스템에서는 새롭게 제안된 반음소(demiphone)를 합성의 기본단위로 사용한다. 반음소는 음소를 그것의 정상상태시점인 중점을 기준으로 해서 다시 양분함으로써 얻어진다. 음소를 양분하여 얻어진 두개의 반음소 중에서 먼저 것을 전반음소(initial demiphone), 나중 것을 후반음소(final demiphone)라고 한다. 반음소의 경계는 음소 및 다이폰의 경계와 일치한다. 따라서 전반음소와 후반음소들을 적당히 결합함에 따라 음소를 만들 수도 있고 다이폰을 만들 수도 있다. 이와 같은 성질 때문에 반음소는 음소와 다이폰의 장점을 동시에 가지게 된다. 다시 말하자면 음소와 마찬가지로 다루기 쉽고 메모리 양을 적게 필요로 하며, 다이폰과 마찬가지로 합성시 얻어지는 합성음성의 음질이 좋게 되는 장점이 있다. 그림 2는 반음소와 다른 음성단위를 비교할 수 있는 반음소의 개념 설명도이다.

2.2 음성학적 전처리(Preprocessing)

무제한 음성합성 시스템에 있어서 음성학적 전처리 단계는 문자열 정형화부(Text Preformatting Block), 문장구조 추출부(Parsing Block), 음운변동 처리부(Phonetic Recoding Block)로 세분될 수 있다.

문자열 정형화부는 문자열이 입력되면 숫자,알파벳, 많이 쓰이는 영어단어,약자등을 입력한 사전을 참조하여 문자열 속에 있는 모든 약어 숫자 특수기호 및 수식 수사처리를 하여, 발음 가능한 문자열과 제어문자 구독기호(Punctuative Symbol)들로만 구성된 정형화된 문자열을 생성하며, 1300여 단어의 발음예외사전의 탐색 및 치환과정을 수행한다.

문장구조 추출과정은 조사,어미,선어말어미 등으로 구성된 형식형태소 사전(약 400단어)과 관형사,부사, 불완전명사들로 구성된 실질형태소 사전(약 230단어)을 참조하여 정형화된 문자열에 대해 구문해석을 함으로써 수식, 피수식 관계 및 구, 절의 경계점을 검출하고, 구 및 절의 통사적 기능을 결정한다. 음운변동 처리부는 발음예외사전 탐색결과 형태소 분석결과 문장구조 추출결과 및 음운변동 알고리듬 을 이용하여 자소 단위의 변환을 수행한다[8].

2.3 운율(Prosody)

운율조절부에서는 운율의 기본 요소인 각 음소의 길이, 억양 그리고 세기를 조절한다. 운율발생모델로서는 그림 3에 보인 바와 같은 모델을 사용하였다. 이 모델에 따르면, 운율발생부에서는 첫째로 각 음소에 대한 길이를 할당한다. 본 연구에서는 음소의 길이가 음성학적 요인, 구문론적요인 그리고 발음속도에 의하여 변하는 것으로 모델링을 하여 길이조절규칙을 제안하였다. 즉, 임의의 음소의 길이는 전후의 음소의 음성학적 성질에 따라 적절히 줄어드는 것으로 모델링을 하였으며, 문장의 끝이나 억양구의 끝에서는 길이가 늘어나도록 하였다. 둘째로 억양에 있어서는 어절단위, 억양구단위 그리고 절단위의 억양조절규칙을사용하였는데, 이 조절규칙으는 기본적으로 어절단위의 피치패턴규칙과 baseline resetting규칙으

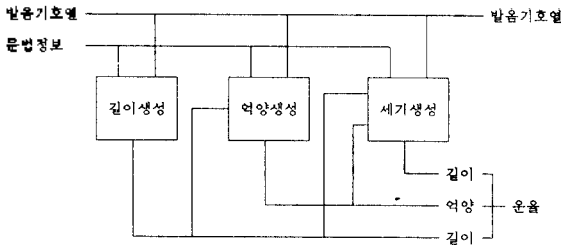


그림 3. 운율 생성도

로 이루어져 있다[10]. 다음 세번째로 각 음소의 세기 조절규칙에서는 먼저 음소별 기준세기를 설정한 후에 부여된 피치의 크기에 따라서 선형적으로 증감하는 방법으로 사용하였다.

2.4 합성 방식(Synthesis Method)

원음성(original speech) 중의 유성음 구간의 신호를 각 성문펄스(glottal pulse)에 의해 만들어지는 한 주기분의 음성파형에 해당하는 단위파형 (unit waveform 또는 wavelet) 들로 분해하는 주기파형분해 방식과 저장된 단위파형들 중 배치시키고자 하는 위치에 가장 가까운 단위파형을 선택하여 그것들을 서로 중첩시킴으로써 원음성의 음질을 그대로 가지면서도 음성단편의 지속시간(duration)과 피치주파수(pitch)를 임의대로 조절할 수 있게 하는 시간 왜곡식 단위파형 재배치방식을 합성 방식으로 사용한다.

주기적 음성을 그것의 스펙트럼 포락 함수의 시간영역 함수인 임펄스 응답과, 주기적 음성과 주기가 같고 평탄한 스펙트럼 포락을 가진 주기적 피치펄스열 신호로 디콘볼루션한 다음 이포크 검출 알고리즘(epoch detection algorithm)[7]과 같은 시간영역에서의 피치검출 알고리즘을 이용하여 주기적 피치펄스열 신호나 시간영역의 음성파형으로부터 피치펄스들의 위치를 구한 후 피치펄스가 한 주기구간당 하나씩 포함되도록 피치펄스열 신호를 주기적으로 분할하고, 유효지속시간에 따라 파라미터 연장과 영샘플 추가를 한 후, 이 피치펄스신호들을 그 주기구간 동안의 임펄스응답과 도로 콘볼루션시키면 단위파형이 구해진다. 그림 4는 주기파형을 분해하는 과정을 보여준다[9].

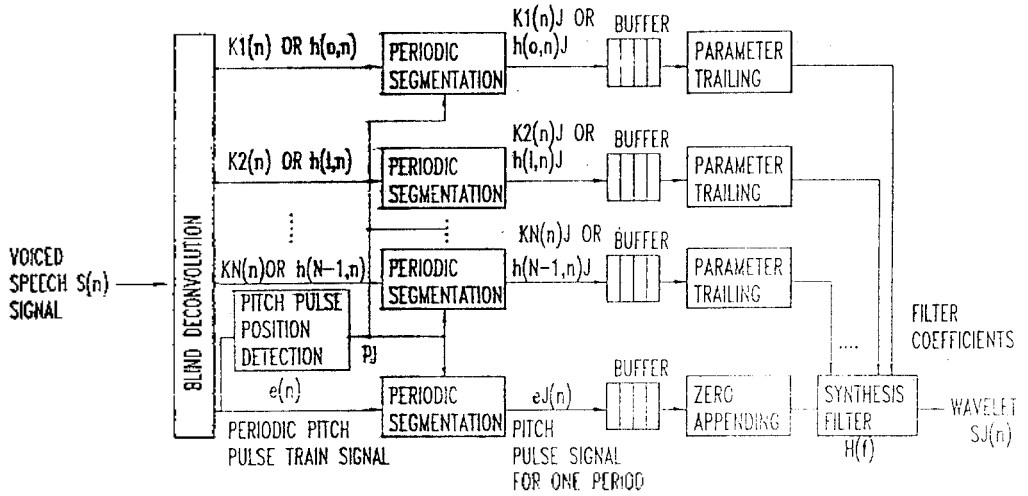
3. 합성 시스템의 평가

합성시스템의 평가방법으로는 객관적인 방법과 주관적인 방법이 있다. 본 논문에서는 주관적인 합성음 평가방법인 MOS방법을 사용하여 자연도 평가를 수행하였다. 물론 이해도의 평가는 MOS평가방법을 사용하지 않고 별도로 수행하여야 하겠으나 MOS 평가속에 이해도의 평가도 어느 정도 이루어진다고 가정하였다.

합성시스템의 평가용 문장으로서 임의로 추출된 2개의 문장을 추출하였는데 다음과 같다

(문장1) 가을은 참 이상한 계절이다

(문장2) 조금 차분해진 마음으로 오던 길을 되돌아 볼 때, 푸른 하늘 아래서 시름시름 앓고 있는 나무들을 바라볼 때, 산다는 게 뭘까 하고



(그림 4) 주기파형 분해 방법의 블록선도

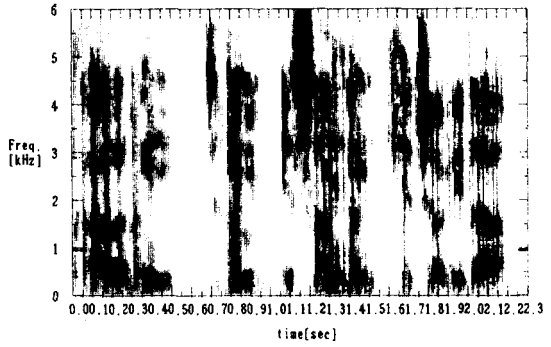


그림 5. 문장 1의 합성음의 스펙트로그램

문득 혼자서 중얼거릴 때, 나는 새삼스레 착해지려고 한다.

평가용 문장으로 선택된 위의 문장들을 살펴보면, 단문과 혼합문으로서 한국의 문장구조의 다양한 형태를 반영하고 있어, 합성시스템의 평가에 적절함을 알 수 있다. 그림 5는 문장 1에 대한 합성음의 스펙트럼을 보여주고 있다.

평가는 자연음을 들려주고 이를 5점 기준으로 한 후, 합성음을 들려주어 0-5사이의 점수를 주도록 하여 평균한 점수이다. 평가는 유일하게 판매되고 있는 합성시스템 '가라사대'합성기를 동시에 평가하도록 하여 본 음성합성시스템과 비교하도록 하였다 표 1은 본 음성합성시스템의 MOS평가 결과를 보여주고 있다.

표 1. 합성음의 평가

	가라사대	한소리
문장 1	3.2	4.3
문장 2	2.8	4.1

위의 표를 살펴보면, 제안된 합성시스템 한소리의 음질이 상당히 자연스러움을 알 수 있으며, 기존의 합성기인 가라사대에 비하여 우수함을 알 수 있다

4. 결 론

이 논문에서는, 음성인식, 기계번역, 음성합성으로 이루어지는 자동통역전화 시스템의 음성합성부분에 관하여 설명하였다. 음성합성시스템의 음질을 결정하는 주요 요소들은 합성의 기본단위, 합성 방법, 음운학적 전처리방법 및 운율정보이며, 이 논문에서는 현재 개발중인, 반음소(demiphone)를 음성합성의 기본단위로 하고, 형식형태소만으로 음성학적 전처리를 실행하며, 개선된 한국어 운율정보가 적용되고, 피치조절을 가능케 하는 음성단편 조합 방식을 합성방식으로 하는 무제한 음성합성 시스템(한소리)을 소개하였다. 청취실험결과 본 합성시스템을 통한 합성음이 자연스러움을 확인할 수 있었다.

앞으로, 억양구 단위에 대한 규칙과 억양구를 구별하기 위한 문장분석기술, 다양한 억양형태를 포함하는 억양제어규칙 그리고 합성단위를 연결함에 있어서 포만트를 부드럽게 연결시키는 기술등에 대한 연구가 필요하겠고, 이러한 기술이 확보되면 좀더 자연스러운 합성음이 가능할 것이다.

참고 문헌

[1] J.B.Allen ,L.R. Rabiner, "A Unified Approach to Short-Time Fourier Analysis and Synthesis," Proceedings IEEE, 65(11), pp. 1558-1564, Nov. 1977.

[2] S.Roucos, A.Wilgus, "High Quality Time-Scale Modification for Speech", Proc. ICASSP, 1985.

[3] F.J.Charpentier and M.G.Stella, "Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation," IEEE Int. Conf. Acoust., Speech, Signal Processing, 1986.

[4] T.Yazu and K.Yamada, "The Speech Synthesis System for an Unlimited Japanese Vocabulary," IEEE Int. Conf. Acoust., Speech, Signal Processing, 1986.

[5] C.d'Alessandro , J.S. Lienard,"Decomposition of he Speech Signal into Short-Time Waveform Using Spectral Segmentation," IEEE Int. Conf. Acoust., Speech, Signal Processing, 1988.

[6] 이종락,"반음소:새로운 음성합성 및 인식단위," 음성통신 및 신호처리워크샵논문집, 1993년 8월.

[7] 김재인, 이종락, "성문 폐색시점 검출에 관한 연구," 음성통신 및 신호처리워크샵논문집, 1993년 8월.

[8] 강용범, 안치홍, " 무제한 음성합성 시스템을 위한 운장구조 추출에 관한 연구," 음성통신 및 신호처리 워크샵논문집, 1993년8월.

[9] 김용인, 박용규, 이종락, "단위파형 재배치에 의한 음성합성 방식," 음성통신 및 신호처리워크샵 논문집, 1993년 8월.

[10] 김 진영, 성 굉모, "한국어의 억양에 관한 연구," Korean-Japan Joint Symposium on Acoustics, pp. 292-297, 1991.