포만트 합성방식에 의한 한국어 문자/음성 변환에 관한 연구(I)

김민년°, 최진산°, 손일권*, 이준우*, 배건성°
° 경북대학교 전자공학과 * LG전자

A Study on the Korean Text-to-Speech Conversion Using the Formant Synthesizer(I)

Min Youn Kim°, Jin San Choi°, Ill Kwon Son*, Joon Woo Lee*,and Keun Sung Bae°
° Dept. of Electronics, Kyungpook Nat'l Univ. * LG Electronics Co.

요약

본 연구에서는 음소단위의 포만트 합성방식을 이용하여 한국어의 규칙합성에 대해 실험하였다. 포만트 합성방식으로는 Klatt가 제안한 직/병렬 합성기를 수정하여 사용하였으며, 운율정보를 나타내는 피치값의 제어는 Fujisaki 모델을 이용하였다. 합성에 사용되는 각 파라미터들이 합성음의 음질 및 파형에 미치는 영향을 분석할 수 있도록 합성 파라미터와 음성파형 및 스펙트로그램을 화면에 나타내고 마우스를 이용하여 파라미터 값을 사용자가 적절히 변경한 후 합성할 수 있는 포만트 방식의 합성 Tool을 개발 하였으며, 이를 이용하여 한국어 문자/음성변환 시스템을 지속적으로 연구하고자 한다.

I. 서론

인간과 기계와의 정보전달(Man-Machine Interface) 수단으로써 사람의 음성을 이용하려는 연구가 꾸준히 계속되어 왔으며 이러한 노력은 최근 multimedia의 보급에서 보다 많은 필요성이 요구 되고 있다[1]. 음성을 이용한 인간과 기계와의 정보전달 방법에는 분석/합성과 문자/음성변환 방법이 있으나 다양한 음성을 만들어 내지 못하는 분석/합성 방식과는 달리 문자/음성변환(TTS:Text-to-Speech) 시스템에서는 각종 파라미터의 변환에 의한 다양한 음성의 합성이 가능하고 어휘에 제한이 없다는 장점 때문에 국내외에서 많은 연구가 진행되고 있다. 이때 문자/음성변환 시스템의 합성음은 올바른 정보전달 능력을 나타내는 명료성과 인간의 발성과의 유사함을 나타내는 자연성으로 평가되어 지는데, 음성합성의 영역이 넓어지고 보편화됨에 따라 인간의 음성과 같이 자연스러운 합성음에 대한 요구가 증가되고 있다.

본 연구에서는 합성단위로 베이타베이스의 양이 최소인 음소단위를 사용하여, 다양한 형태의 음성합성이 용이한 포만트 방식을 이용하여 한국어 문자/음성 변환시스템을 구현하고자 하였다. 합성시스템으로는 Klatt가 제안한 직/병렬 포만트 합성시스템을 사용하였으며 음소간의 파라미터 접속을 위해서 비선형적인 평활화 방법을, 운율제어는 Fujisaki가 제안한 모델[6]을 사용하였다.

II. 규칙에 의한 음성합성

규칙에 의한 음성합성은 언어의 기본 단위인 음소, 음절 등의 조합에 의해 음성신호를 합성해야 하므로 음소, 음절 등의 연결시 상호 조음현상의 처리 및 자연스러운 운율처리 등이 한국어에 맞게 고려되어야 한다. 따라서 이러한 처리를 위하여 문자/음성 변환기는 크게 언어 처리부와 음성합성 처리부로 나누어진다.

본 연구에서 구현한 문자/음성 변환기에서 언어 처리부에서는 정서법으로 쓰여진 문장을 입력받아 문장분석을 한 후, 입력된 문장을 발음규칙을 적용하여 실제로 발음되는 형태의 문장으로 변환시킨다. 그 다음 입력문장을 음소별로 분석하여 각 음소에 해당하는 파라미터 값을 포함하는 베이타베이스의 index를 찾는다. 음소별로 분류할 때 각 음소를 무성자음, 유성자음, 유성음 및 묵음으로 분류하여 각 음소의 정보에 포함시킨다.

음성합성 처리부에서는 각 음소의 위치 및 종류에 따라 지속시간을 제어하고, 다음으로 각 음소에 해당하는 index를 사용하여 합성에 필요한 파라미터를 베이타베이스로부터 불러온 후 문장단위로 기본주파수를 제어한다. 에너지 제어는 음절과 단어 단위로 제어하고, 그 이외의 파라미터는 각 음소단위로 제어하여 합성에 필요한 파라미터 파일을 만든다. 이 파라미터 파일을 사용하여 직/병렬 포만트 합성기를 통하여 음성을 합성한다. 그림 1은 문자/음성 변환기의 블럭선도를 나타낸 것이다.

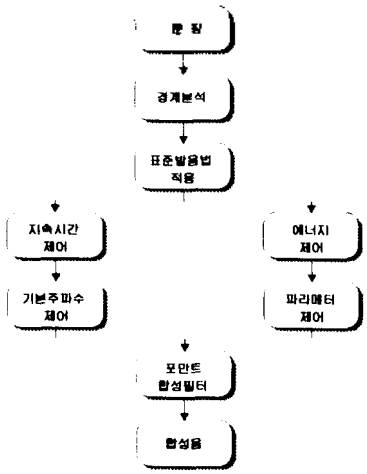


그림 1. 문자/음성 변환기의 블럭선도

1. 언어 처리부

규칙합성 시스템의 언어처리 과정은 문자형태의 입력을 받아 음운법칙을 사용하여 발음하는 형태의 문자로 변환 후, 합성단위의 파라미터 열로 바꾸는 과정으로서, 여기에는 단순한 발음표기 변환 뿐만 아니라 음성합성시 필요한 문장의 음조에 관련된 정보도 분석하는 과정이 포함되어 있다. 한글문장을 키보드로 부터 받아들여서 제어에 필요한 문장의 종류를 입력된 문장의 구두점에 따라 평서문, 감탄문, 의문문으로 구분하고, 2바이트 상용조합형으로 표현한 입력문장을 실제 발음하는 형태로 변환하기 위해 각 음절을 초, 중, 종성의 3바이트로 분리하고 한국어 표준발음법에 따라 받침의 발음, 음의 동화, 경음화, 자음접변, 구개음화 등을 처리한다. 이 때 예외적인 경우와 소리의 첨가 등은 간단한 예외사전을 구성하였다.

2. 음성합성 처리부

1) 지속시간 제어

한국어는 극단적 강세시간언어와 음절시간언어의 중간형태지만, 음절의 지속시간은 음절을 구성하는 음소의 종류나 갯수보다는 어절안의 음절수와 단어, 구, 절, 문장의 경계에 의해 크게 결정되는 경향이 있다. 그리고 음절을 구성하는 각 음소의 지속시간은 음절의 장단 비율과 동비례하지 않는 특성을 가지고 있다. 즉, 이해도와 자연성을 유지하기 위해서는 자음의 최소시간, 고유지속시간은 보장되어야 하며 따라서 자음보다는 모음이 음절의 장단에 더 밀접한 관계를 가지고 있다. 음절내에서는 자음의 지속시간이 뒤따르는 모음의 영향을 받는 특성이 있다. 또 앞, 뒤 음절과의 연결방법이 음소의 지속시간에 영향을 준다.

본 연구에서는 각 음절 앞뒤의 음절과의 연결환경에 따라 각 음소의 고유지속시간을 적절히 수정하여 제어하는 지속시간 제어방식[3]을 사용하였다. 지속시간 제어는 먼저 어절의 전체 지속시간을 구한 뒤, 어절내의 각 음절의 지속시간을 구한다. 그리고 단어, 절, 문장의 경계점 정보를 이용하여 휴지시간 처리를 한다. 음절들의 지속시간이 결정되면 Klatt의 분절음 지속시간 규칙을 참고하여 만든 규칙에 따라, 각 음절을 구성하는 음소의 지속시간을 구한다.

2) 기본주파수 작성규칙

한국어 표준발음에서는 기본주파수 패턴에 의해 의미가 달라지는 어휘의 쌍은 존재하지 않는다. 같은 단어를 구성하는 음절구조에 따른 기본주파수 패턴의 변동이 한국어를 위한 강세, 그리고 개인의 발성 습관 등도 변동을 일으키는 큰 요인으로 작용한다. 본 연구의 문자/음성변환 시스템에서는 Fujisaki 모델을 사용하여 평서문을 기준으로 하여 기본주파수를 제어 하였다.

3) 에너지의 제어

음의 에너지는 소리의 크기를 나타내는 것으로 성대안의 성대가 열리는 정도와 성문을 통과하는 공기의 양과 속도로서 나타난다. 에너지는 음성의 인지도에 영향을 미치는 중요한 요소로서 단어나 문장 내에서 강조하려고 하는 부분에서 크게 나타나는 등 문장내의 운율성분과 관계가 있으나 합성음의 자연성과 명료성에는 영향을 미치는 주된 요인은 아니다. 음절의 에너지는 일반적으로 앞 부분이 크며, 자음보다 모음이 크다는 경향이 있다. 이러한 에너지의 제어는 비선형 평활화 방법을 사용하여 음절단위로 제어하였다[8].

4) 포만트 주파수 및 대역폭 제어 방법

음성합성의 단위가 음소인 문자/음성 변환기에서는 포만트와 대역폭의 제어는 합성음의 명료성에 커다란 영향을 미친다. 따라서 음소단위의 포만트 및 대역폭의 파라미터 제어는 신중히 고려되어야 한다. 포만트 트랙의 변화는 갑작스럽게 변하지 않고 시간에 따라 완만한 궤적을 나타낸다. 이러한 완만한 궤적을 모델링하기 위하여 본 연구에서는 비선형 평활화 방법으로 포만트 주파수 및 대역폭의 궤적을 조절하였다.

III. 포만트 합성을 위한 베이타베이스

1. 음성신호의 분석 과정

음소단위의 베이타베이스를 작성하기 위하여 음성신호를 표준발음을 배운 20대의 남성화자로부터 음절단위인 460개의 음

성을 10kHz로 샘플링하여 수집하였다. 수집한 음성 데이타로 부터 각 음소를 hand labeling에 의해 분리한 후 각 음소에 해당하는 대표 포만트의 주파수를 추출하여 베이타베이스 작성에 사용하였다. 포만트 추출을 위해 사용된 음성신호의 분석조건은 다음과 같다.

LPC order : 12

Window length : 20ms

Preemphasis factor : 0.95

Window type : Hamming

위에서 얻어진 LPC 계수로 만들어 지는 전달함수의 다항식에서 root solving 방법으로 pole을 구하여 각 음소의 대표적인 포만트 주파수 및 대역폭을 구하였다[9].

2. 베이타베이스 작성

베이타베이스의 작성 방법은 합성단위에 따라 여러가지로 나누어진다. 여기서 음성의 합성단위는 출력음성을 구성하기 위해 연결시키는 합성파라미터의 구성요소를 일컫는다. 따라서 베이타베이스의 작성 방법은 음절외에 어절이나 단어, 음소, 구 등이 있을 수 있다. 어절이나 구, 단어 단위의 합성은 많은 양의 데이타를 필요로 하므로 메모리 양이 많이 필요하다는 단점이 있다. 음절, 반음절, 음소단위를 사용하면 적은 양의 베이타베이스로 다양한 음성을 합성할 수 있지만 합성단위를 연결해 음성을 발생시키는 방법이 복잡해진다. 그러나 적은 베이타베이스로 다양한 운율정보를 추출해 적용한다면 매우 실용성이 뛰어난 문자/음성 변환기를 구현할 수 있다. 따라서 본 연구에서는 음소단위의 베이타베이스를 이용하여 문자/음성 변환 시스템을 구현하고자 하였다.

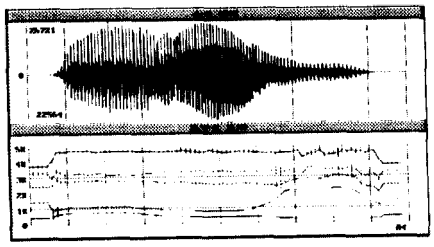
음소단위에 들어가는 파라미터는 표 1에서 추어진 한국어 음소의 음성신호로 부터 분석실험을 통하여 대표 파라미터를 뽑아 베이타베이스를 작성하였다. 각 음소는 초성 18개, 중성 21개, 종성 7개로 나누어 베이타베이스를 작성하였는데 음소를 초성, 중성, 종성으로 나눔으로써 합성에 필요한 음소에 대한 처리가 손쉬워지고 파라미터들도 변화시킬 수 있어 베이타베이스 작성이 쉬워진다.

표1. 실험에 사용된 음소표

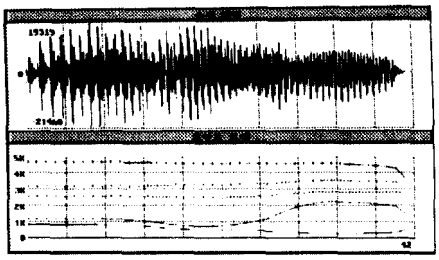
초성(18개)	ㄱ ㄲ ㄴ ㄷ ㄸ ㄹ ㅁ ㅂ ㅃ ㅅ ㅆ ㅈ ㅉ ㅊ ㅋ ㅌ ㅍ ㅎ
중성(21개)	ㅏ ㅐ ㅑ ㅒ ㅓ ㅔ ㅕ ㅖ ㅗ ㅘ ㅙ ㅚ ㅛ ㅜ ㅝ ㅞ ㅟ ㅠ ㅡ ㅢ ㅣ
종성(7개)	ㄱ ㄴ ㄷ ㄹ ㅁ ㅂ ㅇ

Ⅳ. 실험 및 검토

합성시에 사용되는 파라미터의 조절은 합성음질을 결정하는 중요한 요소라고 할 수 있다. 기본주파수, 에너지, 포만트와 대역폭 등이 복합적으로 조절이 잘 되어야만 좋은 합성음을 얻을 수 있다. 특히 원음의 포만트 궤적을 얼마나 잘 모델링할 수 있는가에 의해 명료성이 좌우되므로 포만트 궤적의 평활화 방법은 매우 중요한 것이다. 본 연구에서 사용한 평활화 방법이 얼마나 원음의 포만트 궤적과 유사하게 모델링할 수 있는가를 알아 보기 위하여 유성음만 있는 단어를 사용하여 비교하였다. 그림 2에서는 "아오이"라는 무의미 단어에 대해 본 연구에서 사용한 포만트 궤적과 실제 음성에서 구한 포만트 궤적을 비교하였는데 상당히 유사한 것을 볼 수 있다.



(a) Original speech

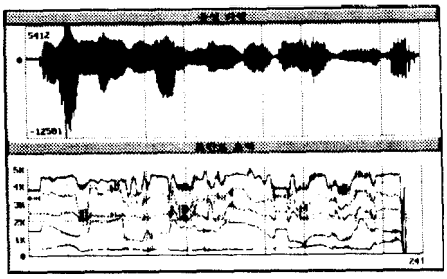


(b) Synthetic speech

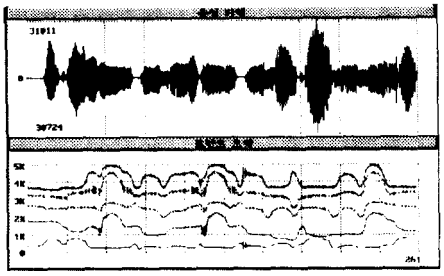
그림 2. 원음 및 합성음 "아오이"의 포만트 궤적 비교

그림 3은 "대한민국은 민주공화국이다."의 원음과 음소단위의 포만트 방식으로 합성된 합성음을 보인 것이다. 합성시 사용한 여기신호는 Rosenberg가 제안한 모델[7]을 사용하였고, glottal open quotient는 70%를 사용하였다.

또한, 본 연구에서는 문자/음성변환 시스템의 지속적인 연구를 위해 합성 파라미터의 변화에 따라 합성음의 변화를 실험할 수 있는 tool을 개발하였다. 그림 4에서와 같이 표시모드에서는 음성파형, 스펙트로그램, 포만트 궤적등을 보여주고, 합성모드에서는 사용자가 마우스의 조작으로 파라미터 값을 바꾸어 다양한 합성실험을 할 수 있다.

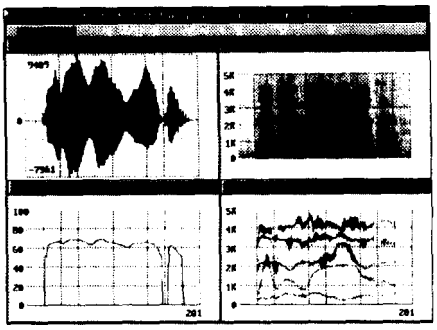


(a) Original speech

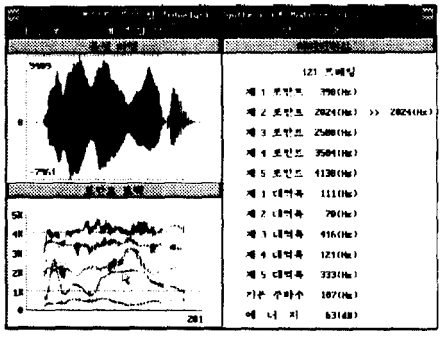


(b) Synthetic speech

그림 3. "대한민국은 민주공화국이다."의 원음과 합성음



(a) Display mode



(b) Modify and Synthesis mode

그림 4. 합성파라미터 표시·갱신 도구

V. 결론

음소단위의 포만트 음성합성 방식을 이용하여 한국어의 문자/음성변환 시스템을 구현하여 실험하였다. 포만트 합성방식으로는 Klatt가 제안한 직/병렬 합성기를 수정하여 여러가지 형태의 여기신호와 파라미터 값을 사용자가 조절할 수 있도록 함으로써 다양한 음질을 갖는 합성음을 얻을 수 있도록 하였다. 합성할 때 음소 사이의 포만트 및 에너지 값 등의 파라미터 변화를 적절히 접속하기 위해 비선형적인 평활화 방법을 사용하였으며, 운율정보를 나타내는 피치값의 제어는 Fujisaki 모델을 이용하였다.

합성 실험결과에 의하면 어느정도 명료성은 나타났지만 명료성 및 자연성 모두 다 만족스럽지 못했다. 명료성의 문제점은 한국어 표준 발음에 대한 포만트 정보를 분석 및 합성실험을 통해 정확히 추출하여 베이타베이스를 구축함으로써 해결할 수 있다고 생각된다. 자연성의 개선을 위해서는 먼저 음소 사이의 파라미터 접속에 관한 규칙정보가 합성시에 좀 더 구체적으로 고려되도록 하여야 하며, 한국어의 복합적인 운율정보에 대해 많은 연구가 이루어져야 한다.

참고 문헌

[1] S. E. Levison, J. P. Olive, and J. S. Tschirgi, "Speech Synthesis in Telecommunications," *IEEE Commun.*, Vol. 31, No 11, pp.46-53, 1993.

[2] D. H. Klatt "Review of text-to-speech conversion for English," *J. Acoust. Soc. Am.*, Vol. 82, No. 3, pp. 737-793, 1987.

[3] 이용주, "통신처리를 위한 음성정보 변환기술 개발," ETRI 연구보고서, 1990.

[4] 조철우, "포만트 합성법을 이용한 한국어 규칙합성에 관한 연구," 박사학위논문, 1988.

[5] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, Vol. 67, No.3, pp.971-995, 1980.

[6] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Ja.*, pp.233-242, April 1984.

[7] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, Vol. 87, No.2, pp.820-857, 1990.

[8] 손일권, "포만트 합성방식에 의한 음소단위의 한국어 규칙합성," 석사학위논문, 1994

[9] 이준우, "음성 및 EGG 신호 분석에 의한 포만트 추정 및 음성합성," 석사학위논문, 1994