論文 91-28B-10-2

합성음의 자연도 향상을 위한 포만트 궤적 중첩 방법

(Formant Locus Overlapping Method to Enhance Naturalness of Synthetic Speech)

安 承 權* 成 宏 模**

(Seung Kwon Ahn and Koeng Mo Sung)

要 約

본 논문에서는 포만트를 이용한 반음절 단위의 한국어 text-to-speech 시스템을 구현할 때, 포만트 궤적 정보들을 효과적으로 중첩하여 합성음의 자연도를 향상시킬 수 있는 새로운 방법을 제안한다. 한국어 반음절의 포만트 궤적을 임의의 갯수의 선형적 변화 구간으로 분할한 후, 각 구간 경계점의 포만트 정보 및 구간 길이들로 반음절 데이타 베이스를 구성한다. 이렇게 구성된 데이타 베이스를 이용하여 음성을 합성할 때, 각 반음절들의 포만트 궤적을 중첩하여 연결하는 방법을 제안 하므로써, 합성음의 포만트 궤적을 인간의 조음 메카니즘에 보다 가깝게 모델링할 수 있게한다.

제안된 방법을 이용하여 한국어 text-to-speech 시스템을 구성하고, 합성된 음의 포만트 궤적을 자연음의 포만트 궤적과 비교하여 유사성을 입증하며, 기존의 방법보다 자연음에 가까운 음성을 합성할 수 있음을 보였다.

Abstract

In this paper, we propose a new formant locus overlapping method which can effectively enhance a naturalness of synthetic speech produced by demisyllable based Korean text-to-speech system. At first, Korean demisyllables are divided into several number of segments which have linear formant transition characteristics. Then, database, which is composed of start point and length of each formant segments, is provided. When we synthesize speech with these demisyllable database, we concatenate each formant locus by using a proposed overlapping method which can closely simulate human articulation mechanism.

We have implemented a Korean text-to-speech system by using this method and proved that the formant loci of the synthetic speech are similar to those of the natural speech. Finally, we could illustrate that the resulting spectrograms of proposed method are more similar to natural speech than those of conventional method.

*正會員, 金星社 中央研究所
(GoldStar Central Research Lab.)
**正會員, 서울大學校 電子工學科
(Dept. of Elec. Eng., Seoul Nat'l Univ.)
接受日字 : 1991年 6月 18日

I. 서 론

Text-to-speech 시스템은 작은 음운 단위를 기본 단위로하여, 언어학적 지식에 근거한 규칙들을 적용하므로써 임의의 문장을 음성으로 변환시키는 시스

템이다. 이러한 text-to-speech 시스템을 구현하는 방법으로는 LPC나 LSP 등의 coding 계수를 이용하는 방법과 성도의 공진 주파수인 포만트 정보를 이용하는 방법등이 있다.[1]-[4] 전자의 방법은 자연음으로 부터 성도 필터 계수를 coding 기법에 의해 추출하고, 이를 decoding하여 음성을 합성해내므로 대체로 양호한 음질을 얻을 수 있다. 그러나, 음절 길이의 조절 및 음운현상 처리를 위한 parameter의 조작이 불편하고, all-pole model로는 비음 처리가 불가능하다는 한계도 있다. 반면, 포만트 합성 방법은 포만트 정보의 추출과정이 어렵고 많은 시간이 소요되지만, data량을 줄일 수 있고 포만트 정보의 조작이 간편하기 때문에, 성도의 발생 메카니즘에 상응하는 여러 규칙들을 이용하여 보다 자연스러운 음성을 합성해 낼 수 있다.

한국어에 대한 text-to-speech 시스템도 최근 많은 연구자들에 의해 여러가지 방법으로 구현된 것들이 발표되고 있지만, 아직까지 합성음의 자연도에 있어서 많은 개선이 요구된다.[1][6]-[8] 따라서, 본 논문에서는 반음절 단위를 이용한 한국어 포만트 합성기를 제안하고, 특히, 합성음의 자연도 향상을 위하여 반음절 단위의 포만트 정보들을 연결할 때, 이들을 자연음과 더욱 유사하게 모델링할 수 있는 연결 방법에 대해서 중점적으로 기술한다. 먼저, 합성 단위로는 합성음의 질과 database의 양 사이의 절충점(trade-off)을 고려하여 CV 및 VC 결합에 의한 포만트의 전이 정보(inforation of formant transition)를 포함하는 반음절 단위를 선택한다. 즉, 한국어의 반음절에 대한 포만트 궤적 정보를 기본 단위로 한다. 이 기본 단위들을 database화할 때, 포만트 궤적 정보를 각 반음절 단위 내의 모든 frame들에 대하여 모두 저장하게 되면 정보량이 방대해지므로, 각 반음절을 선형 전이구간(linear transition segment)으로 분할하여 그 경계 부분의 정보만을 저장하므로써 data량을 감축 시킨다. 이 반음절 단위의 포만트 정보들을 운율 규칙에 의하여 연결하여 음절을 형성할 때, 자연음의 음절 길이보다 길게 합성되는 현상 및 한음절 내에서 모음의 안정구간의 포만트 특성이 자연음의 포만트 특성과 차이가 발생되는 현상이 생길 수 있다. 이러한 현상들을 없애기 위해 두 반음절 단위의 포만트 궤적을 중첩시킨 후, 포만트 스위칭 및 interpolation 방법을 이용하여 포만트 궤적을 연결하는 새로운 방법을 제안하므로써 합성음의 자연도 향상을 도모한다.

이러한 방법들을 필자들이 이미 발표한 text-to-speech 시스템에 적용시켜 합성음을 만든 뒤, 자연음과 스펙토그램 비교 및 기존 방법에 의한 스페트로그램과의 비교 분석 결과, 제안된 포만트 궤적 중첩 방법에 의해 합성된 음이 기존 방법에 의한 것보다 훨씬 자연음에 가깝게 합성되었음을 보인다.

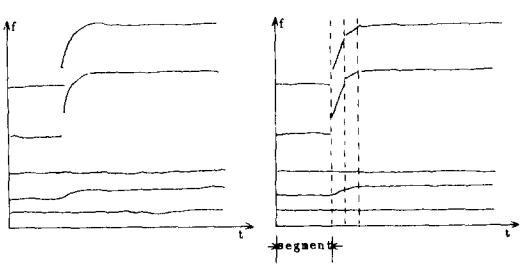
Ⅱ. DATABASE 구성 및 포만트 궤적 분할

합성에 이용된 기본 database는 반음절 단위의 포만트 궤적정보로 구성된다. 포만트는 1차부터 4차까지를 이용하며, 각 포만트 정보는 중심 주파수 및 대역폭을 포함한다. 이때, 병렬합성(parallel systhesis)이 필요한 파열음, 마찰음 및 무성 자음의 경우는 각 포만트의 power 정보들도 함께 저장된다. 그밖에, 피치궤적, 에너지 envelope 및 유성음/무성음의 분류 정보들도 database에 포함된다. 반음절 단위의 포만트 정보들을 database화 하기 위해서, 본 논문에서는 각 반음절의 스펙트로그램상의 포만트 궤적을 관찰하여, 모음의 안정구간 전부분(CV 부분) 및 안정구간 후부분(VC부분)의 선형적 전이 구간을 기준으로 각각 분할(segmentation)한 후, 각 segment의 시작점의 포만트 정보와 각 segment의 길이 및 경사도만 저장하므로써 database의 크기를 줄일 수 있도록 한다. 즉, 다음과 같은 규칙에 의해 포만트 정보가 분할된다.

i) 한 음절은 두개의 반음절로 분리된다.
- 반음절#1 : 초성+전이구간+모음의 안정 구간
- 반음절#2 : 모음의 안정구간+전이구간+종성

ii) 각 반음절은 임의의 갯수의 segment로 분할된다.

iii) 각 segement 내에서 포만트 궤적은 선형 특성을 갖는다.



(a) '미/mi/'의 포만트 궤적 (b) (a)의 궤적의 분할

그림 1. 반음절 '미/mi/'의 포만트 궤적 분할
Fig. 1. Segmentation of the formant loci of the demisyllable /mi/.

이러한 분할 방법의 예로써 그림1에 '미/mi/'라는 반음절의 포만트 분할을 보였다.

한국어에 있어서 반음절의 갯수는 모든 초성과 중성의 조합 및 중성과 종성의 조합 갯수를 더한 것이 되어 약 400여개가 된다. 그러나, 한국어의 음소는 음운학적 특성상 조음 위치가 비슷한 몇개의 군(group)으로 분류가 가능하며, 조음 위치가 유사한 음소들의 그 포만트 궤적도 유사하므로 본 논문에서는 표1과 같이 조음 위치가 비슷한 음소들을 grouping 하므로써 database의 양을 더욱 감소시킬 수 있었다.[5]

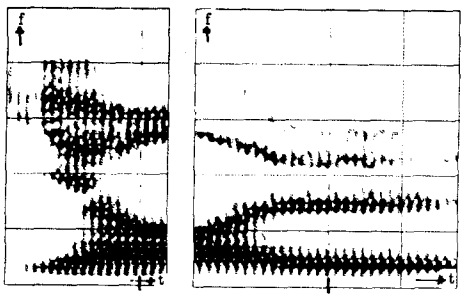
표 1. 조음 위치에 따른 한국어 음소군 분류
Table 1. A grouping of korean phonemes according to the articulation position.

Group	Initial sound	Final sound
1	/g/, /k/, /k^h/+V	V+/g/, /ŋ/
2	/n/, /d/, /t/, /t^h/+V	V+/n/, /d/
3	/l/+V	V+/l/
4	/r/+V	
5	/m/, /b/, /p/, /p^h/+V	V+/m/, /b/
6	/s^h/, /s/+V	
7	/j/, /c/, /c^h/+V	
8	/h/+V	

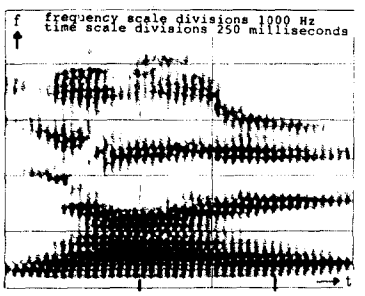
(V : 모음)

Ⅲ. 중첩에 의한 반음절들의 연결

반음절 단위의 한국어 합성에 있어서 두개의 반음절들(CV, VC)을 연결해서 하나의 음절을 형성할 경우, 길이가 실제 자연음보다 부자연스럽게 길어지는 현상이 생긴다. 이를 해결하기 위해 기존의 방법들에서는 frame을 subsampling 하는 방법들을 사용하기도 하지만, 이렇게 하면 합성음의 명료도가 떨어질 뿐 아니라, 모음의 포만트 변화 경사(slope of formant transition)가 급하게 되고, 전이구간의 길이(transition period)도 줄어들게 되어 자연도도 떨어지게 된다.[7]-[11] 그림2에 이러한 예로써 두 반음절 '여/yeo/'와 '얼/eol/'의 스펙트로그램들과 이들 두 반음절을 연결하여 만들고자하는 음절 '열/yeol/'의 스펙트로그램을 보인다. 여기서 두 반음절을 더한 길이가 원하는 음절의 길이보다 훨씬 길어진 것을 알 수 있으며, '여'및 '얼'의 각각의 모음의 안정 구간의 스펙트로 그램과 '열'의 모음의 안정 구간의 스펙트로그램이 다름을 볼 수 있다. 이러한 현상이 나타나는 이유는, 자연음의 경우 전후 자음의 영향때문



(a) '여/yeo/' (b) '얼/eol/'



(c) '열/yeol/'

그림 2. 반음절 '여', '얼' 및 음절 '열'의 스펙트로그램
Fig. 2. Spectrograms of demisyllables '/yeo/', '/eol/' and syllable '/yeol/'

에 모음이 안정된 상태에 이르지 못하거나, 모음의 포만트 특성이 앞에 오는 자음의 영향으로 일시적으로 변화 하였다가 시간이 지남에 따라 안정된 상태에 도달하려다가도 뒤에 오는 자음의 영향으로 또다시 포만트가 전이되는 경우가 있기 때문이다.

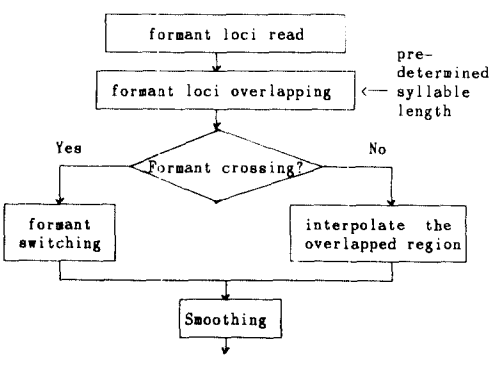
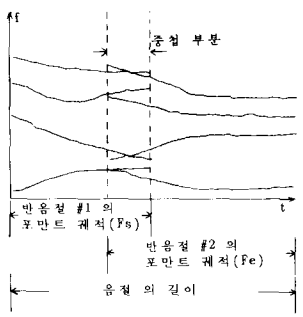
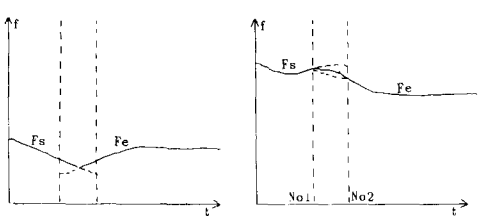


그림 3. 중첩에 의한 반음절 연결 방법 흐름도
Fig. 3. Flowchart of demisyllable concatenating method by overlapping.

따라서, 본 논문에서는 이러한 발음 특성을 효과적으로 모델링할 수 있도록 그림3에 보인것과 같은 중첩에 의한 반음절 연결 기법을 제안하는 것이다. 즉, 합성시 database로 부터 읽어온 CV 및 VC 부분의 포만트 궤적들을 운율 제어 부분에서 결정된 원하는 음절 길이에 맞추어서 그림4(a)와 같이 중첩된다. 여기서, 각각의 포만트 궤적에 대해 CV부분 및 VC 부분의 포만트 궤적의 교차 여부를 조사하여, 교차되는 경우는 교차점에서 CV 부분의 포만트 궤적과 VC부분의 포만트 궤적을 스위칭 시켜주며(그림 4(b)), 교차되지 않을 경우는 중첩 부분의 CV 부분과 VC 부분의 포만트 궤적을 interpolation 시킨다(그림 4(c)).



(a) 궤적 중첩



(b) 스위칭 (c) interpolation

그림 4. CV, VC의 포만트 궤적 중첩
Fig. 4. Formant loci overlapping of CV and VC parts.

이때 interpolation 하는 방법은 다음 식(1)에 의한다.

$$F(j)=Fs\times\frac{No2-j}{No2-No1}+Fe\times\frac{j-No1}{No2-No1},$$
$$No1\leqq j\leqq No2, \qquad (1)$$

여기서,

j : frame index,
No1 : 중첩 부분의 시작 frame index,
No2 : 중첩 부분의 끝 frame index,

이러한 방법으로 포만트 궤적을 재구성한 후, 각 segment의 연결부분 및 포만트 궤적 switching 부분의 불연속점을 보정하기 위해 다음의 식(2)에 따라 포만트 궤적을 smoothing 시킨다.

$$Fsm(j)=\frac{1}{9}\sum_{k=j-4}^{k=j+4}F(k),\ for\ all\ j, \qquad (2)$$

F : 포만트 궤적,
Fsm : smoothing된 포만트 궤적,
j : frame index,
frame length : 5 msec.

Ⅳ. 시스템 구현 및 평가

Ⅱ장에서 기술된 방법으로 구성된 한국어 데이타베이스를 바탕으로, Ⅲ장의 포만트 중첩 방법에 의해 연결된 포만트 궤적을 이용하여 그림5의 구성을 갖는 새로운 한국어 text-to-speech 시스템을 구성하였다. 전체 시스템은 크게 linguistic process, synthesis process, Korean database 및 합성기로 구성된다. linguistic process에서는 특수기호, 숫자 등을 한국어 문자로 바꾸어 주며, 장음표기, 액센트 처리 및 발음기호로의 변환등을 행한다. synthesis process에서는 피치, 에너지 및 duration 조절에 의한 운율 제어를 행하며, 이때 database로 부터 원하는 음성의 포만트

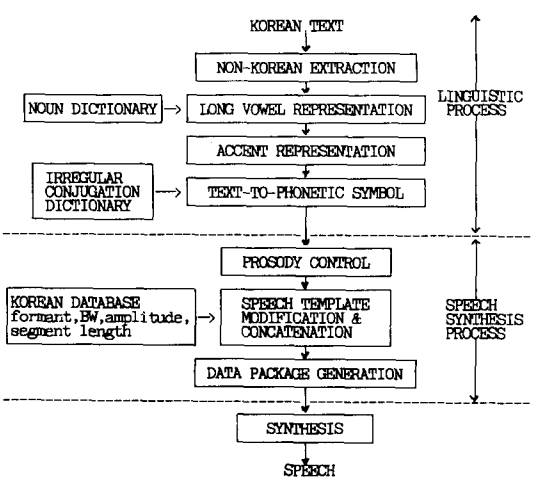
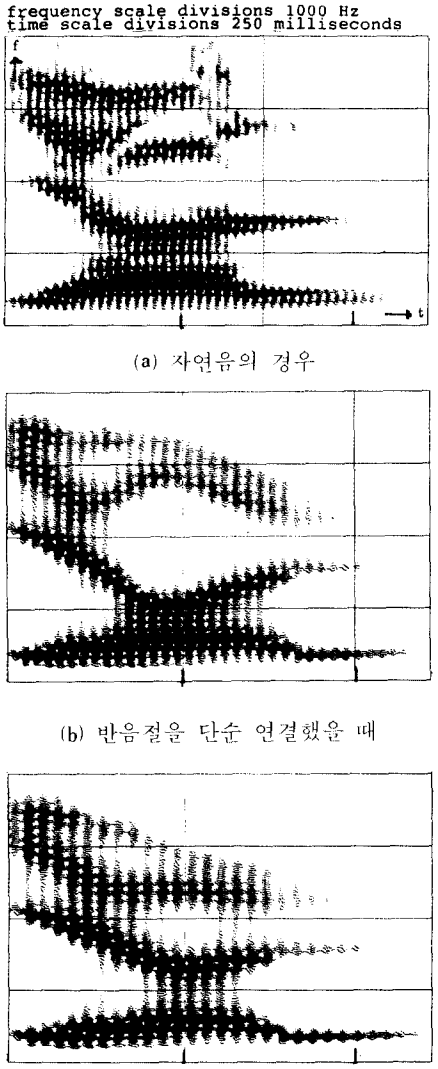


그림 5. 합성 시스템의 구조
Fig. 5. Block diagram of the text-to-speech system.

정보를 읽어서 이들을 연결시켜 전체 포만트 궤적을 구성한다. 또한, 이 과정에서 추출된 parameter 들을 합성기에서 읽을 수 있는 형태로 나열되어 datapackage로 구성된다. 합성기는 이러한 정보를 이용하여 최종적으로 합성음을 내는데, 본 시스템에서는 Klatt 합성기를 한국어의 특성에 맞게수정하여 사용한다.[12] 전체 시스템의 구성 및 각 부분의 자세한 구현 과정은 이미 발표된 reference1을 참조하기 바란다.[1]

제안된 방법에 의한 합성음을 평가하기 위해 그림 6에 음절 '열'에 대한 자연음, 기존의 방법에 의한 합성음 및 제안 방법에 의한 합성음의 스펙트로그램을 각각 나타내어 비교하였다. 이 그림에서 볼 수 있듯이 자연음의 음절 길이와 동일하게 만들었을 때 기존 방법에 의한 포만트의 경사들은 자연음에 비해 훨씬 급하게 만들어짐을 볼 수 있으며, 제안된 방법에 의한 합성음들의 포만트 경사들이 상대적으로 자연음에 가깝게 되므로써 합성음의 자연도를 향상시킬 수 있음을 보였다. 즉, 기존의 frame subsampling 방법에 의한 합성음은 모음의 포만트 전이 구간의 경사를 급하게 하고, 전이 구간의 길이도 줄일 수 밖에 없으므로 부자연스럽게 되지만, 제안된 방법에 의한 합성음은 이러한 문제를 해결함과 동시에, 전후 자음의 영향에 의한 모음의 안정구간(stable region)에 대한 보상도 허주므로써 자연음에 가깝게 합성됨을 알 수 있다.



(a) 자연음의 경우

(b) 반음절을 단순 연결했을 때

(c) 제안된 궤적 중첩을 했을 때

그림 6. 음절 '열'에 대한 스펙트로그램 비교
Fig. 6. Spectrogram comparison for the syllable '/yeol/'.

V. 결 론

본 연구에서는 반음절을 기본단위로 한 한국어 text-to-speech 시스템에서, 기존의 방법보다 자연도를 향상시킬 수 있는 포만트 연결방법을 제안하였다. 반음절 단위의 포만트 정보를 연결하여 음절을 만들 때, 각 반음절 포만트 궤적을 중첩하여 운율 규칙에 따라 원하는 길이로 전체 음절 길이를 맞추고, 각 포만트 궤적의 crossing 여부에 따라 포만트 switching 및 interpolation을 적절히 행하므로써, 음절 길이의 부정확성이나 포만트 궤적의 상이등에 의한 자연도의 저하를 방지하였다. 또한, 이러한 방법들을 액센트, 피치패턴 및 휴지 삽입(pause insertion)등의 운율 처리를 하여 구성한 한국어 text-to-speech 시스템에 구현한 결과, 기존의 방법보다 자연음에 가깝게 합성되었음을 스펙트로그램 비교를 통하여 확인하였다.

본 연구에서 제안된 방법들은 반음절 단위의 합성기 뿐 아니라, 합성 단위가 다르더라도 포만트 정보를 이용한 합성 시스템에서는 모두 유용하게 적용될 수 있을 것이다. 그러나 본 연구 결과는 음절 및 단어들 사이에서의 자연도 향상에 중점을 둔 것이므로, 전체 합성 문장의 자연도 향상을 위해서는 한국어의 운율 규칙에 대한 더욱 체계적인 연구가 이루어져야하고, 또한, 자연도의 평가를 위한 객관적 청취실험 방법등을 연구하여 보다 객관적인 평가를 할 수 있어야 한다.

參 考 文 獻

[1] Seung-Kwon Ahn and Koeng-Mo Sung,

"The Rules in a Korean text-to-speech system," *Proc.ICSLP 90* vol. 2, pp. 777-780, Nov. 1990.

[2] H. Sato, "Japanese text-to-speech conversion system," *Review Elec. Commun. Lab., Nippon Telegraph and Telephone Corp.*, vol. 32, no. 2, 1984.

[3] Yousif A. EL-IMAN, "An unristicted vocabulary Arabic speech synthesis system," *IEEE Trans. Acoust., Speech, Signal processing*, vol. 37, no. 12, pp. 1829-1845, Dec. 1989.

[4] Neal B. Pinto, et. al., "Formant speech synthesis: Improving production quality," *IEEE Trans. Acoust., Speech, Signal processing*, vol. 37, no. 12, pp. 1870-1886, Dec. 1989.

[5] 허웅, 국어 음운학, 정음사, pp. 205-206, 1984.

[6] 김병수, 윤기선, 박성한, "음절 단위를 이용한 한국어 음성 합성," 전자공학회 논문지, 제27권 제1호, pp. 143-150, 1990년 1월.

[7] 이성준 외, "MPLPC 부호화된 반음절을 사용한 무제한 단어 한국어 음성합성," 전자공학회 논문지, 제27권 제9호, pp. 1-10, 1990년 9월.

[8] 윤기선, 박성한, "반음절 단위를 이용한 한국어 음성합성에 관한 연구," 전자공학회논문지, 제27권 제10호, pp. 138-145, 1990년 10월.

[9] Joseph Olive, "A scheme for concatenating units for speech synthesis," *Proc. IEEE ICASSP*, pp. 568-571, 1980.

[10] Herbert E. Wolf, "Control of prosodic parameters for a formant synthesizer based on diphone concatenation," *Proc. IEEE ICASSP*, pp. 106-109, 1981.

[11] F.J. Charpentier and M.G. Stella, "Diphone synthesis using overlap-add technique for speech waveforms concatenation," *Proc. IEEE ICASSP*, pp. 2015-2018, 1986.

[12] D.H. Klatt, "Software for a cascade/parallel formant synthesizer," vol. 67, no. 3, Mar. 1980.

著 者 紹 介

安 承 權(正會員)

1957年 10月 20日生. 1980年 2月, 1982年 2月 서울대학교 전자공학과 학사, 석사. 1982年 3月~현재 (주)금성사 중앙연구소 근무 1991年 7月 서울대학교 전자공학과 박사과정 수료. 주관심분야는 음성신호 처리, 패턴인식 등임.

成 宏 模 (正會員) 第27卷 第3號 參照
현재 서울대학교 전자공학과 부교수