1999年度 大韓電子工學會 秋季綜合學術大會 論文集 第 22 卷 第2號 99/11

표본화율 변환을 이용한 합성음의 운율제어

이 현 구, 홍 광 석
성균관대학교 전기전자 및 컴퓨터공학부 휴먼컴퓨터연구실

Prosody Control of the Synthetic Speech using Sampling Rate Conversion

Hyeon-Gu Lee, Kwang-Seok Hong
HCI Lab., School of Electrical, Electronics and Computer Engineering,
Sung Kyun Kwan University
email : kshong@yurim.skku.ac.kr

Abstract

In this paper, we presents a method to control prosody of the synthetic speech using sampling rate conversion technique. In prosody control, the conventional methods perform overlap and add. So the synthetic speech has a distortion and the voice quality is not satisfied. Using sampling rate conversion technique, we can get high quality of the synthetic speech. Also we can control various talking speeds according to speaker's patterns.

I. 서 론

최근 휴먼 컴퓨터 인터페이스의 핵심 기술로 음성 합성이 각광을 받고 있으며 음성 합성 기술은 이제 임의의 문장을 무제한으로 합성해 내는 단계에 이르렀다. 그러나 아직 합성음의 음질은 자연 음성에 비해 떨어지며 이는 자연 음성이 가지고 있는 운율 정보가 완벽하게 처리되지 못한 이유이다.

운율은 지속 시간, 억양, 세기 등으로 구성되는 언어적 정보를 말하며 합성음의 자연성에 큰 영향을 미친다.

운율을 제어하기 위해 기존의 대부분의 연구에서는 합성 단위의 음편을 시간축 혹은 주파수축에서 중첩, 삽입, 제거를 통해 조절하므로 각각의 음편들의 접합 부위에서 음질의 왜곡을 피할 수가 없거나 연산이 복잡한 단점이 있다.

이러한 문제를 해결하기 위해 본 논문에서는 음성의 표본화율을 변환하므로써 합성음의 음질 왜곡이 없이 운율을 제어하는 방법을 제안한다.

연결 합성에 있어 합성 단위를 미리 녹음하여 데이타베이스로 구축하게 되는데 이 경우 합성 단위는 일률적인 샘플링 주파수로 표본화된다. 따라서 합성시 개별 합성 단위별로 표본화율을 변경시켜 전체 합성음의 운율을 제어한다.

입력 신호에 대해 표본화율을 증가시킨 다음 저역 통과 필터를 거쳐 다시 표본화율을 감소시킨다. 이때 저역 통과 필터의 계수값은 필터 설계시에 미리 정해지며 합성시에는 운율 패턴에 따라 해당 필터를 통과시킨다.

본 논문에서는 한국어는 물론 영어, 숫자, 특수 기호 등에 대한 무제한 합성 시스템을 구현하고, 합성음의 운율을 제어하기 위해 표본화율 변환 기법을 적용하여 합성음의 자연성이 개선되었음을 알 수 있었고 또한 합성음 전체의 출력 속도를 제어하므로써 화자의 발성 속도에 따른 음색 변환이 가능함을 확인할 수 있었다.

II. 표본화율 변환

표본화율을 원하는 비율로 자유롭게 변환시키기 위해서는 먼저 표본화율을 증가시킨 다음 저역 통과 필터를 거쳐 다시 표본화율을 감소시켜야 한다. 그림 1에 표본화율 변환 과정을 나타 내었다.

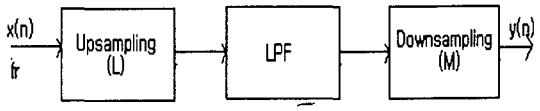


그림 1 표본화율 변환 과정
Fig. 1 Mechanism of sampling rate conversion

입력 표본화율 f_r 은 각 입력 샘플 사이에 $L-1$개의 0을 첨가하므로써 L배 증가하여 표본화율을 높이고(upsampling), 이득이 L이고 stopband의 차단주파수가 $\min(\pi/L, \pi/M)$인 FIR 저역 통과 필터를 거친다. 그런 다음 매 M샘플마다 하나씩 제거시켜 표본화율을 M배 감소시킨다(downsampling).[1]

저역 통과 필터 설계시에 전체적인 표본화율을 낮출 경우에는 인접 파형간에 엘리어싱(aliasing)이 발생하지 않도록 고려하여 설계하여야 음질의 왜곡을 막을 수 있다.[2]

III. 운율 제어

운율은 음의 고저(pitch), 세기(amplitude), 길이(duration)의 세가지에 의해 결정된다.[3] 고저는 성대의 진동 속도에 의해 결정되는데, 성대의 잔동 속도가 빠르면 높은 소리가 생성되고 느리면 낮은 소리가 생성된다. 세기는 음의 크고 작은 정도를 나타낸다. 길이는 음의 명료도에 영향을 주게되는데, 어휘의 의미에 대한 변별 기능을 가진다.

한국어에 있어 모음의 길이는 어휘 의미 구별의 기준이 되므로 음장(vowel length)이라 한다. 즉 같은 모음이라 하더라도 의미에 따라 장모음과 단모음으로 구별되어 생성하여야 한다.

또한 모음끼리의 상대적인 음성학적 길이를 가지게 되는데 이를 그림 2에 나타내었다.

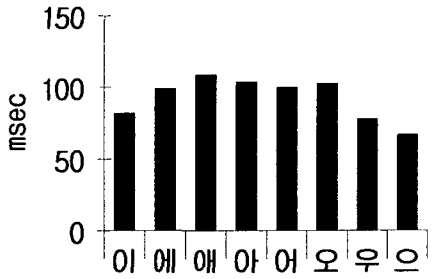


그림 2 모음의 음성학적 길이
Fig. 2 Phonetic length of Vowel

하나의 문장을 발화할 때 문장 내부에서 끊어 읽게되는데 이러한 끊기의 단위를 말토막(rhythm unit)이라 하고 말토막은 운율 구분의 기본이 된다. 말토막 경계의 수와 위치는 문장의 길이, 말의 속도와 스타일, 문법 구조와 의미 구조 등 여러 요인들의 상호 작용으로 결정된다.

한국어에서 억양(intonation)은 핵억양과 말토막 억양으로 구분된다. 말마디의 마지막 음절에 얹히는 억양을 핵억양이라 하고, 말토막을 단위로 말마디 억양보다 작은 패턴으로 부과되는 것이 말토막 억양이라 한다. 이들 억양을 자세히 살펴보면 다음과 같다.[4]

1. 핵억양 종류

(1) 낮은수평조(Low Level)
단정적이고 냉정한 태도를 전달하며, 앞 음절보다 조금 더 낮게 발음된다.
(2) 가운데수평조(Mid Level)
퉁명스런 태도를 전달하며, 앞 음절보다 더 높게 발음된다.
(3) 높은수평조(High Level)
관심이나 흥미를 전달하며, 앞 음절보다 훨씬 더 높게 발음된다.
(4) 낮내림조(Low Fall)
단정적이긴 하지만 부드럽고 친절한 태도를 전달하

며, 앞 음절보다 더 높게 발음한다.

(5) 높내림조(High Fall)

관심이나 놀람을 전달하며, 앞 음절보다 훨씬 더 높게 발음된다.

(6) 온오름조(Full Rise)

크게 놀람이나 의심하는 태도를 전달하며, 앞 음절보다 약간 낮은 높이로 발음된다.

(7) 낮오름조(Low Rise)

권유나 부탁을 전달하며, 앞 음절보다 약간 낮은 높이로 발음된다.

(8) 내리오름조(Fall Rise)

화난 태도를 전달하며, 앞 음절보다 더 높은 높이로 발음된다.

(9) 오르내림조(Rise Fall)

경멸하는 태도를 전달하며, 앞 음절보다 약간 더 낮은 높이로 발음된다.

2. 말토막 억양 종류

(1) 오름조(Rising)

친근한 표현에 사용되며, 말토막의 끝 음절을 제외한 나머지 음절들은 같은 높이로 발음되고 끝 음절은 나머지 음절들보다 더 높게 발음된다.

(2) 수평조(Level)

사무적인 표현에 사용되며, 말토막의 모든 음절이 같은 높이로 발음된다.

(3) 내림조(Falling)

말토막의 마지막 음절이 가장 낮게 발음되고 나머지 음절들은 같은 높이로 발음되거나 차례로 앞 음절보다 조금씩 낮은 높이로 발음된다.

(4) 오르내림조(Rise Falling)

말토막의 두번째 음절이 첫 음절보다 높게 발음되고 나머지 음절들은 차례로 낮게 발음된다.

IV. 실험 및 결과

합성 단위는 반음절을 사용하였다. 반음절 데이타베이스 구축을 위하여 소음이 45-50dB 정도되는 사무실 환경에서 여성 1인 화자가 발성한 음성을 short-term 분석을 통하여 피치의 위치를 얻어내고, 스펙트로그램을 동시에 관찰하여 수작업으로 반음절 단위로 분절하였다. 이를 토대로 반음절의 유형(CV/VC), 전체 샘플 갯수, 지속 시간, 피치 갯수, 평균 피치값, 최대 진폭 등으로 데이타베이스를 구축하였다. 입력 문장이 들어오면 언어 처리부를 통해 합성에 필요한 단위열로 변환되며 운율 정보에 따라 표본화율 변환부를 통해 최종적으로 합성음이 출력된다. 합성 시스템을 그림 3에 나타내었다.

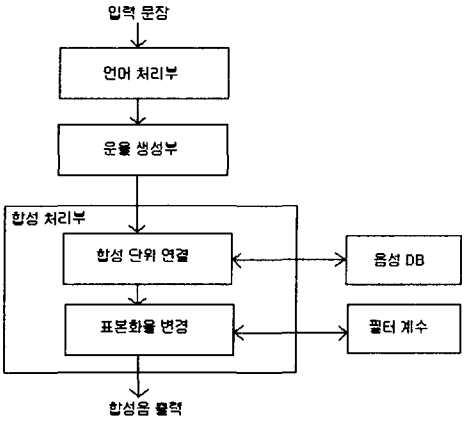


그림 3 합성 시스템

Fig. 3 Synthesis system

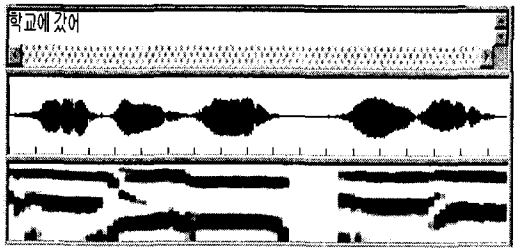
표본화율 변환을 통하여 문장 내에서의 말토막의 상대적인 길이가 운율 정보에 따라 변경됨을 알 수 있었고 아울러 주파수 패턴이 변화하므로 전반적인 피치가 변경됨을 알 수 있었다. 세기의 조절은 합성 단위가 가지고 있는 최대 진폭을 이용하여 에너지의 상대적인 비를 통해 제어가 가능하였다.

또한 TD-PSOLA 합성방식으로도 합성음을 출력해 보았는데 이 경우는 피치의 급격한 변화에 음질이 다소 왜곡됨을 알 수 있었고 전반적인 운율 제어에 있어 본 논문에서 제안한 표본화율 변환을 이용한 운율 제어가 원음성의 음성정보를 잃어 버리지 않고 효율적으로 제어됨을 확인할 수 있었다.[5]

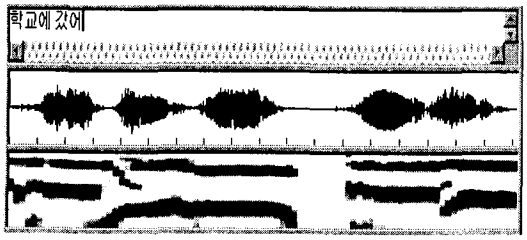
그림 4에 /학교에 갔어/에 대한 반음절 연결 합성음과 반음절 단위 TD-PSOLA 방식의 합성음 그리고 본 논문에서 제안한 방식에 의한 합성음의 파형과 스펙트

로그램을 나타내었다.[6]

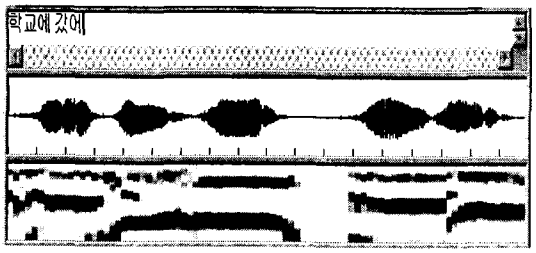
(c)의 경우 합성 단위의 샘플링 주파수 11.025 KHz에 대해 5% 줄인 것으로 저역 통과 필터의 passband ripple은 0.1dB, stopband attenuation은 60dB, passband frequency는 0.0382π, stopband frequency는 0.05π로 설계하여 적용한 것이다. 따라서 마치 10.474 KHz로 샘플링한 효과를 볼 수 있었다. (d)의 경우는 5% 늘인 것으로 passband frequency를 0.0363π로 설계하여 적용한 것으로 11.576KHz로 샘플링한 효과를 보았다.



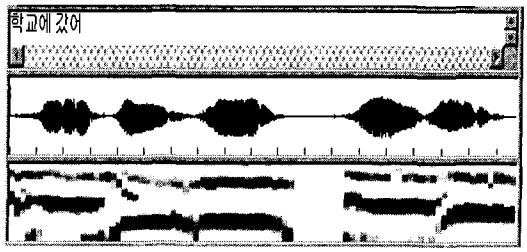
(a) 반음절 연결 합성음(/학교에 갔어/)



(b) 반음절 단위 TD-PSOLA 합성음(/학교에 갔어/)



(c) 표본화율 변경을 이용한 합성음
(/학교에 갔어/ 5% Downsampling)



(d) 표본화율 변경을 이용한 합성음
(/학교에 갔어/ 5% Upsampling)

그림 4 합성 결과
Fig. 4 Result of synthesis

V. 결 론

본 논문에서는 음성 합성에 있어 원음성의 음성 정보를 왜곡시키지 않고도 운율을 제어할 수 있는 표본화율 변환을 이용한 방식을 제안하였다. 실험 결과 명료성과 자연성이 개선된 합성음을 얻을 수 있었으며 또한 합성음의 발성 속도를 자유롭게 조절할 수 있음을 확인하였다. 향후 좀 더 자연스런 합성음을 내기위해 체계화된 운율 정보를 적용하고자 한다.

참고문헌

[1] R.E.Crochiere and L.R.Rabiner, "Optimum FIR Digital Implementation for Detection, Interpolation, and Narrow Band Filtering", IEEE Trans. Acoust.,Speech, Signal Processing, Vol. ASSP-23, No. 5, pp.444-456, 1975.

[2] Alan V.Oppenheim and Ronald W.Schafer, "Discrete-Time Signal Processing",PrenticeHall, 1989.

[3] 배주채, "국어음운론 개설", 신구문화사, 1996.

[4] 이호영, "국어음성학", 태학사, 1996.

[5] Thierry Dutoit, "High Quality Text-to-Speech Synthesis of the French Language", 1993

[6] 이현구 외 3인, "지능적 휴먼-컴퓨터 인터페이스를 위한 무제한 음성합성 시스템 구현", 대한전자공학회 멀티미디어 연구회 창립학술대회 논문집, 1999.