지속시간 변경에 의한 다중음성 합성에 관한 연구

김명*, 서지호, 배명진
숭실대학교 정보통신공학과
*숭실대학교 컴퓨터학과

A Study on the Multiple-Speech Synthesis using the Duration Control

Jin Ming*, Seo JiHo, Bae MyungJin
Department of Information and Telecom & *Computer Science
Soongsil University
mjin1978@hanmail.net, mjbae@ssu.ac.kr

요약

다중음성 합성시스템은 단일 화자의 음성을 입력받아 다양한 음색의 다중음성으로 합성을 해주는 음성합성 시스템이다. 기존의 다중음성 합성시스템의 출력인 다중 합성음은 피치만 변경된 음성으로 원 음성과 동일한 지속시간을 갖게 된다. 따라서 피치 변경된 음성간의 구분이 어렵게 되며 이러한 사항을 개선하고자 본 논문에서는 피치와 지속시간 변경에 의한 다중음성 합성시스템에서 관한 연구를 하였다. 본 논문에서는 시간 영역에서의 지속시간 변경법인 PSOLA방식을 적용하여 피치 변경된 음성의 지속시간을 변경하였다. 지속시간 변경을 적용한 다중음성 합성시스템을 이용하면 한 사람의 응원 목소리로 여러 사람이 응원하는 효과음을 낼 수 있는 합성기로 사용할 수 있고 영화의 효과음, 핸드폰의 음성 메시지 서비스 등에서 용이하게 사용될 것으로 예상하고 있다.

1. 서론

다중음성 합성시스템은 입력된 음성의 운율 즉 피치와 지속시간을 조절하여 다중음성으로 들려준다. 음성을 합성하는 경우 운율정보를 합성음에 반영하면 보다 명확한 의미전달이 가능해 진다. 본 논문에서는 AMDF방식을 이용한 피치시점 검출법을 사용하였으며 검출된 피치 주기에 의하여 고조파 스케일링 방식을 이용하여 주파수 측에서 피치를 변경함으로서 깨끗한 음질을 유지할 수 있었고 스펙트럼 왜곡율도 감소시킬 수 있었다. 또한 운율조절에 있어서 지속시간을 변경하여 줌으로서 화자간의 발성 특성을 구분할 수 있으며 지속시간에 의한 화자 구분도 가능하게 되고 지속시간의 자유로운 변경에 의하여 언어장애인의 발음교정이나 어학학습 등 여러 분야에 이용할 수 있을 것이다.

본 논문은 모두 5장으로 구성되어 있으며, 각 장의 내용은 다음과 같다. 제2장에서는 지속시간 변경에 의한 다중음성 합성시스템의 구성을 설명하였고 3장에서는 다중음성 합성시스템에서 사용된 운율조절 기법에 대하여 설명하였다. 여기에는 피치 변경법과 지속시간 변경법이 있다. 제4장에서는 실험을 거쳐 나온 결과에 대하여 설명하였고 제5장에서 결론을 맺었다.

2. 다중음성 합성시스템의 구성

다중음성 합성시스템에서 지속시간 변경은 기존의 다중음성 합성방식의 문제점을 보완하여 합성음의 명료성과 각각의 피치 변경된 음성간의 구분을 명확히 하였다. 그림 1은 본 논문에서 제안한 지속시간 변경에 의한 다중음성 합성 방식에 대한 블록도이다. 지속시간 변경에 의한 다중음성 합성시스템은 3개의 서브 블록으로 나뉘게 된다. 그림에서 (A)부분

운 피치 검출 과정이고 (B)는 각각의 음성신호에 대한 변경율에 따른 피치 변경 과정이고 (C)는 본 논문에서 적용한 지속시간 변경 과정이다.

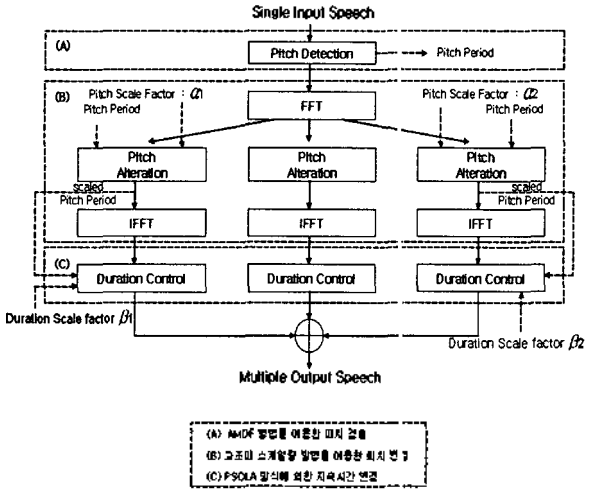

그림 1. 지속시간 변경에 의한 다중음성 합성시스템

3. 피치 및 지속시간 변경법

아래에 다중음성 합성시스템에서 사용된 피치 변경법과 지속시간 변경법에 대하여 설명하였다.

3.1 고조파 스케일링에 의한 피치 변경

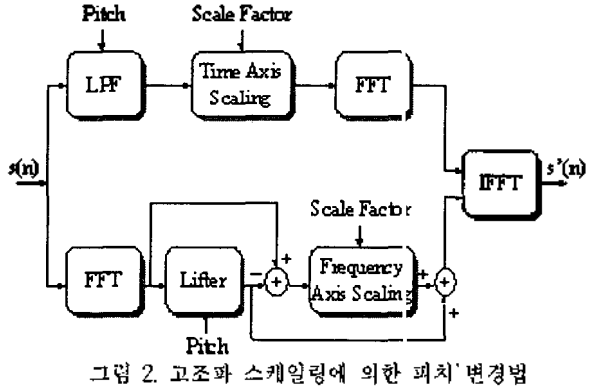

그림 2. 고조파 스케일링에 의한 피치 변경법

고조파 스케일링에 의한 피치 변경법은 주파수 영역에서의 피치 변경법으로서 주파수축 스케일링에 의해 기본 주파수를 변경하기 위해서는 음성 신호를 푸리에 변환하여 진폭 스펙트럼과 위상 스펙트럼으로 분리하여야 한다. 음성 신호의 푸리에 변환은 식 (3.1)에 의해 수행된다.

$$S(K) = \int_{-\infty}^{+\infty} s(n) e^{-j\frac{n}{2nN}k} dn \qquad (3.1)$$

푸리에 변환에 의해 얻어진 음성 스펙트럼은 식(3.2), 식(3.3)와 같이 진폭 스펙트럼과 위상 스펙트럼으로 나타낼 수 있다.

$$M(K) = 10 \log S^2(K) \qquad (3.2)$$

$$\phi(K) = \tan^{-1} \frac{Im[S(K)]}{Re[S(K)]} \qquad (3.3)$$

여기서 $Re[S(K)]$는 음성 스펙트럼의 실수성분이고, $Im[S(K)]$은 음성 스펙트럼의 허수성분을 나타낸다. 주파수 영역에서 기본 주파수를 변경하기 위해서 주파수축 스케일링을 사용할 수 있다. 주파수축 스케일링은 음성의 여기 스펙트럼에 대하여 수행하여야 하므로 여파기 스펙트럼과 여기 스펙트럼의 성분 분리는 기본 주파수를 변경하기 이전에 수행되어야 한다. 음성 신호의 기본 주파수를 높이는 것은 시간 영역에서는 피치 주기를 줄이는 것과 동일하며, 기본 주파수를 낮추는 것은 시간영역에서 피치 주기를 늘이는 것과 같다. 진폭 스펙트럼에서 근사적인 여파기 스펙트럼은 식 (3.4)에 의해 구할 수 있다.

$$F(K) = \frac{1}{K_0} \sum_{i=-\frac{K_0}{2}}^{\frac{K_0}{2}} M(K-i) \qquad (3.4)$$

여기서 K_0는 주파수 축에서의 기본 주파수에 해당하는 고조파 간격을 나타내며 다음 식 (3.5)과 같다.

$$K_0 = \frac{window\ size}{pitch} \qquad (3.5)$$

대수 진폭 스펙트럼 $M(K)$에서 근사 여파기 스펙트럼을 식 (3.6)과 같이 빼면 평탄화된 여기 스펙트럼 $E(K)$를 분리할 수 있다.

$$E(K) = M(K) - F(K) \qquad (3.6)$$

이 신호에 대하여 주파수 영역에서 스케일링을 함으로써 기본 주파수를 변경한다. 이 때 스케일링 율은 시간축 스케일링 계수의 역수를 사용하여야 한다.

$$E'(K) = E(K \times \rho^{-1}) \qquad (3.7)$$

여기서 ρ^{-1}은 주파수축 스케일링 율이다. 기본 주파수를

높이기 위해서는 고조파의 간격을 ρ^{-1}만큼 늘이고 기본 주파수를 낮추기 위해서는 고조파의 간격을 ρ^{-1}만큼 줄인다.

3.2 PSOLA방식에 의한 지속시간 변경

본 논문에서는 다중음성 합성 시스템의 지속시간 조절을 위하여 양질의 합성음을 얻을 수 있는 시간영역 합성기법인 PSOLA 합성방식을 적용하였다. PSOLA 알고리즘은 우선 피치주기 단위로 음성 파형을 분해한 다음 분해된 피치 단위에 윈도우 함수를 곱해서 단구간 ST(Short Term)신호의 열로 만들고 분해된 단위를 반복 삽입하여 지속시간을 늘이거나 삭제하여 조절한다. 신호의 재결합은 합성음의 지속시간과 피치를 고려하여 짧은 구간의 신호를 재배치하여 중첩 가산한다.

PSOLA 합성은 다음의 3 단계에 의하여 처리된다. 처음 입력 파형을 연속되는 분석 ST신호로 분할한다. 다음 각각의 ST 분석 신호를 ST 합성 신호로 변형하고 마지막으로 변형된 ST신호를 overlap adding 하여 최종 합성음을 만들어 낸다.

피치 동기 분석은 다음과 같이 이루어진다. 원래 음성 파형이 유성음인 경우에는 피치단위로 분해한 다음 윈도우 함수를 곱하여 ST신호의 열로 만든다. 무성음인 경우에는 10ms의 주기로 일정하게 분석한다. 분석 윈도우 함수에는 Hanning Window를 사용한다. 이런 윈도우 함수를 원래의 음성 샘플에 곱함으로써 다음 식 (3.8)과 같은 피치 단위로 분해된 샘플열들을 얻는다.

$$S_{analysis}(n) = W_{analysis}(m-n)S(n) \qquad (3.8)$$

식(3.8)에서 $S(n)$은 원 음성 파형을 표시하고 m은 m번째 피치를 의미하며 $S_{analysis}(n)$는 피치주기 단위의 ST 신호이고 $W_{analysis}(n)$는 분석 윈도우 함수이다. 음성 신호를 느리게 혹은 빠르게 할 때는 각각 분석 ST 신호의 선택적인 반복이나 제거에 의하여 변형된 피치마크에 ST signal을 씌워 합성함으로써 속도를 조절할 수 있다. 이렇게 재배열된 ST 신호에서 겹쳐지는 부분을 더해주면 된다. 그림 3과 그림 4는 피치주기 단위의 PSOLA합성방식을 적용하여 지속시간을 변경하는 예를 나타낸 것이다.

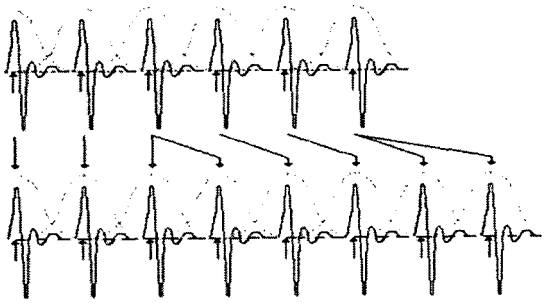

그림 3. PSOLA 합성방식에 의한 지속시간 확장 방법

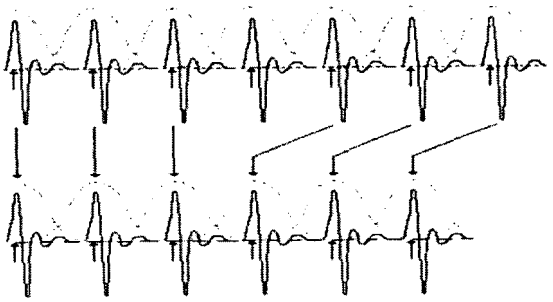

그림 4. PSOLA 합성방식에 의한 지속시간 압축 방법

4. 실험 및 결과

컴퓨터 시뮬레이션을 위하여 상용화된 16비트 AD/DA 변환기를 인터페이스로 사용하여 8kHz의 표본율로 데이터를 입력하였다. 분석 프레임의 길이는 256샘플로 처리하였다. 처리결과의 성능을 측정하기 위해 다음의 대표적인 노래와 문장을 연령층이 다양한 남녀 5명이 각 5번씩 발성하여 시료로 사용하였다.

발성 1 : 응원가 필승 코리아
발성 2 : 생일 축하 노래
발성 3 : 여기는 음성통신 연구실입니다.
발성 4 : 인수네 꼬마는 천재 소년을 좋아한다
발성 5 : 아름다운 가을입니다

그림 5는 음성 시료 /응원가 필승 코리아/에 대하여 피치 및 지속시간을 변경한 스펙트로그램이다. 그림에서 (a)는 원 음성의 스펙트로그램이고 (b)는 기본주파수를 80%, 지속시간 60%로 변경한 스펙트로그램이며 (c)는 기본주파수를 120%, 지속시간을 140%로 변경한 음성의 스펙트로그램이다. 그림 6에서는 원음성과 다중 합성음의 파형 및 스펙트로그램을 나타내었다.

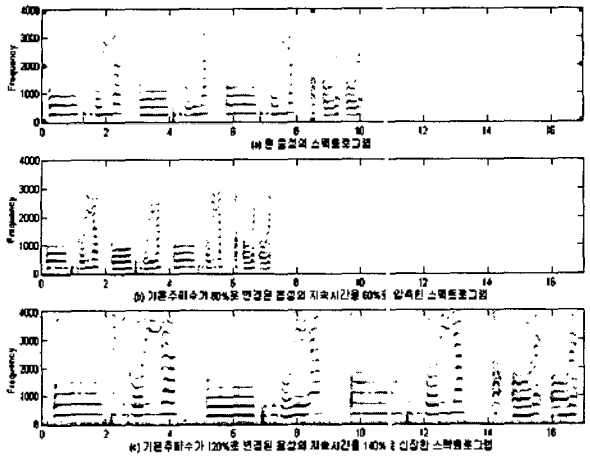

그림 5. 음성의 스펙트로그램

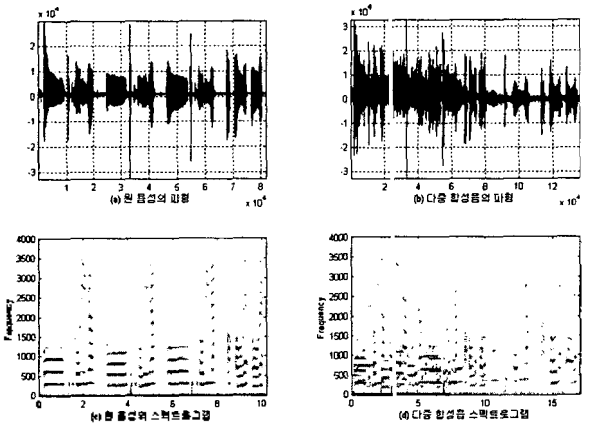

그림 6. 다중 합성음의 스펙트로그램

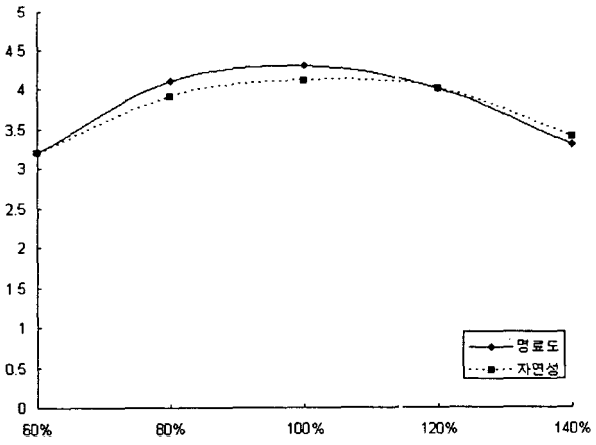

그림 7. 지속시간 변경율에 따른 평균 MOS

그림 7은 본 논문에서 사용한 5가지 발성시료에 대하여 지속시간 변경율에 따른 평균 MOS(Mean of Score)를 보여주었다. 다중음성 합성시스템에서 지속시간을 변경한 결과로부터 보면 지속시간 변경율이 80%에서 120%사이일 경우 합성음의 자연성과 명료도는 높은 MOS를 가지게 되고 80%이하 혹은 120% 이상의 변경율을 취하였을 경우 자연성과 명료도 방면에서 상대적으로 낮은 MOS값을 가진다는 것을 알 수 있다.

5. 결 론

본 논문에서는 지속시간 변경에 의한 다중음성 합성 시스템에 대한 연구를 하였으며 시스템의 출력인 합성음의 명료도 MOS는 3.9, 자연성 MOS는 3.8을 기록하였다. 따라서 지속시간 변경된 다중합성음은 지속시간 변경을 적용한 다중음성 합성시스템을 이용하면 한 사람의 응원 목소리로 여러 사람이 응원하는 효과음을 낼 수 있는 합성기로 사용할 수 있고 영화의 효과음, 핸드폰의 음성 메시지 서비스 등에서 용이하게 사용될 것을 예상하고 있다.

향후 연구 방향으로는 음절 단위의 지속시간 변경 기법에 대한 연구가 이루어지야 할 것이다. 또한 다중음성 합성시스템의 실시간 구현에 있어서 합성음의 개수에 따른 계산의 과부하를 줄일 수 있는 최적화된 알고리즘을 개발하여 시스템의 기능을 강화시키고 본 시스템의 상용화를 추진하는 것이다.

참고문헌

1. H. Valbret, E. Moulines and J. Tubach, "Voice transformation using PSOLA technique", Speech Communication, vol.11, no.2-3, pp.175-187, 1992.
2. MyungJin Bae, "Digital Speech Synthesis", Published by Dong-Young, 1999.
3. HyungBin Park, MyungJin Bae, "On a Detection of Pitch Point for Voice Color Conversion", J. Acoust. Society, Korea, Vol. 19, No.1, pp. 1, 49-152, July 7 8, 2000.
4. Y.H. JUNG, J.K. KIM, W.R. JO, and M.J. BAE," A Study on Real Time Pitch Alteration of Speech Signal." WSEAS, Vol.-Signal Processing System and Imaging, pp.457-461, July 7-10, 2003.