포먼트 이동과 스펙트럼 기울기의 변환을 이용한 음색 변환

손성용(ICU), 한민수(ICU)

<차 례>

1. 서론
2. 기존의 음색 변환
  2.1. 선형다변회귀 모델을 이용한 음색 변환
  2.2. 포먼트 이동에 의한 음색 변환
3. 제안된 알고리즘에 의한 음색 변환
  3.1. 포먼트의 위치 변환
    3.1.1. 포먼트 피크 추정 알고리즘
    3.1.2. 포먼트 변환
  3.2. 스펙트럼 기울기 변환
  3.3. 삼각창(Triangular Window)을 이용한 TD-PSOLA
    3.3.1. 기존의 PSOLA 방법
    3.3.2. 삼각창을 이용한 TD-PSOLA
4. 실험 및 결과
5. 결론

<Abstract>

Voice Color Conversion Based on the Formants and Spectrum Tilt Modification

Song-Young Son, Min-Soo Hahn

The purpose of voice color conversion is to change the speaker identity perceived from the speech signal. In this paper, we propose a new voice color conversion algorithm through the formant shifting and the spectrum-tilt modification in the frequency domain. The basic idea of this technique is to convert the positions of source formants into those of target speaker's formants through interpolation and decimation and to modify the spectrum-tilt by utilizing the information of both speakers' spectrum envelops. The LPC spectrum is adopted to evaluate the position of formant and the information of spectrum-tilt. Our algorithm enables us to convert the speaker identity rather successfully while maintaining good speech quality, since it modifies speech waveforms directly in the frequency domain.

* Keywords: Voice Color Conversion, formant shifting, LPC, Interpolation, decimation, TD-PSOLA

1.서 론

음색 변환(Voice Color Conversion)이란 화자 음성의 개인적인 특성을 수정하거나 치환하는 기술을 말한다. 즉 임의의 화자가 발성한 음성을 다른 화자가 발성한 것처럼 들리도록 하는 것이다. 이런 음색 변환 기술은 최근 문서음성변환(Text-to-Speech) 시스템의 급증하는 수요로 인하여 그 중요성이 커지게 되었다. 문서음성변환 시스템에 쓰이는 데이터베이스의 구축은 매우 많은 시간과 비용을 요구한다. 때문에 이미 구성된 데이터베이스 이외의 음성을 출력하기 위해서 또 다른 데이터베이스를 구성한다는 것은 비효율적이다. 이러한 비효율성을 해결하는 데 음색 변환 기술은 필수적이다. 또한 발성구조의 결함으로 발성되는 비정상적인 음성을 보다 쉽게 알아들을 수 있도록 음색 변환을 사용할 수 있으며, 보안이나 신분보호를 위한 음성변조에도 사용될 수 있다.

음색 변환의 핵심이 되는 화자 음성의 개인적인 특성은 크게 음향학적 요소와 운율적인 요소로 나누어진다. 음향학적인 요소는 포먼트 주파수(Frequency), 포먼트 대역폭(Bandwidth), 스펙트럼 기울기(Spectrum Tilt)와 성문 파형(Glottal Waveform) 등이 있으며 운율적 요소는 주기(Pitch), 발성 지속시간(Duration), 에너지(Energy) 등이 있다[1][2]. 이러한 여러 요소들을 원시 화자에서 목적 화자로 수정하거나 변환해줌으로써 음색 변환을 수행할 수 있다.

초기의 음색 변환은 벡터양자화(Vector Quantization) 방법과 LPC 파라미터를 많이 이용하였다. 그러나 이 방법은 변환된 LPC 파라미터와 LPC 잔차신호를 이용하여 합성할 경우 두 파라미터 사이의 불일치로 왜곡이 발생하고 음질이 저하되는 단점을 가지고 있다[3].

본 논문에서는 이러한 단점을 보완하기 위해 LPC 계열(LPCC, LSF 등)로 변환하지 않고 주파수 영역에서 포먼트와 스펙트럼 기울기를 변환함으로써 음색 변환을 수행할 것이다.

제시한 방법은 분석 단계와 변환 단계, 마지막으로 합성 단계로 구분이 된다. 첫 번째 단계인 분석 단계는 LPC 스펙트럼으로부터 포먼트 정보와 스펙트럼 기울기 정보를 추정하며, 변환 단계에서는 분석 단계에서 추정되어진 정보를 이용하여 주파수 영역에서 목적 화자로의 변환이 수행되고, 마지막 합성 단계에서는 역푸리에 변환과 후처리 과정으로 운율정보인 주기를 변환시켜준다.

본 논문은 2장에서 두 가지 기존의 음색 변환을 소개하고 3장에서 제안된 포먼트의 변환, 스펙트럼 기울기의 변환, 주기변환 알고리즘에 대하여 설명하였다. 4장에서 실험 환경 및 결과를 제시하고 5장에서는 결론을 맺는다.

2. 기존의 음색 변환

2.1. 선형다변회귀 모델을 이용한 음색 변환

기존의 음색 변환 중 하나인 선형다변회귀 모델은 DTW (Dynamic Time Wapping)[4]로 시간 정렬을 수행한 후 벡터양자화와 선형다변회귀를 이용해 음색 변환을 수행한다. 여기서 핵심이 되는 회귀분석이란 독립변수(Independent variable)로부터 종속변수(Dependent variable)를 예측하기 위하여 사용되는 회귀 방정식(Regression equation)이라는 두 변수 사이의 구체적인 함수 관계를 규명하는 데 이용되는 통계적 분석 방법을 말한다. 이때 독립변수와 종속변수의 관계가 선형이라면 선형회귀(Linear regression)라고 하며 다수의 변수를 가지는 선형회귀를 선형다변회귀라고 한다[5]. 이러한 선형다변회귀를 이용하여 LPC 켑스트럼을 원시 화자에서 목적 화자로 변환시키게 된다. 변환된 LPC 켑스트럼과 잔차신호를 컨볼루션하여 신호를 합성할 때 두 파라미터간의 왜곡된 관계 때문에 음질의 저하가 발생한다. 이러한 음질 저하는 LPC 계열의 특성 파라미터를 이용할 경우 대부분 발생하며 그로 인해 양질의 변환된 음성을 얻기 힘들다.

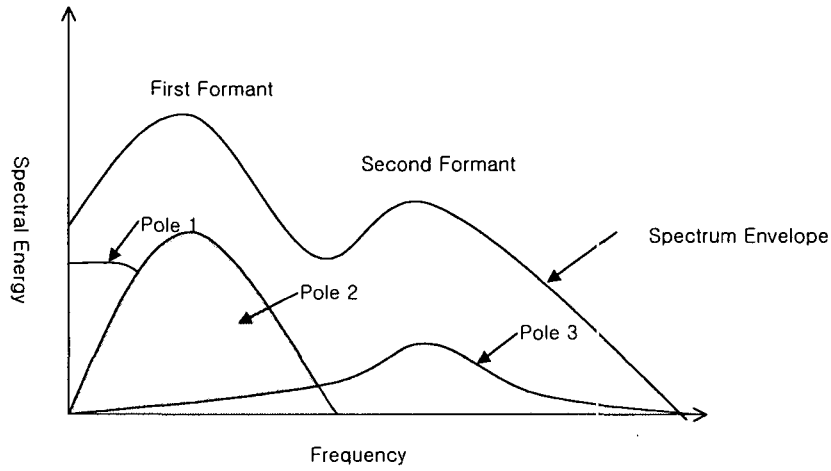
2.2. 포먼트 변환에 의한 음색 변환

음성신호의 포먼트는 극점의 집단에 의해 특징지어진다는 사실에 근거하여, 포먼트를 구성에 중심이 되는 극점의 위치를 원시 화자로부터 목적 화자로 이동함으로써 음색 변환을 수행하는 것이다[6]. 변환 과정은 성도 변환 함수 $S(z)$으로부터 극점을 구하고 얻어진 극점 중 포먼트 형성에 중심이 되는 극점을 찾아내어 이동시키는 것이다. 또한 스펙트럼의 기울기도 변환시켜 준다.

(1) $$S(z) = \frac{1}{1 + \sum_i \alpha_i z^{-i}} \qquad \text{for i=0,1... P}$$

(2) $$1 + \sum_i \alpha_i z^{-i} = 0 \qquad \text{for i=0,1... P}$$

위의 과정에서 변환된 스펙트럼을 역 푸리에를 시킴으로써 변환된 음성을 얻게 된다. 이러한 알고리즘의 장점은 주파수 영역에서 직접 음성신호를 변경시킴으로써 신호의 왜곡을 줄일 수 있다는 점이다.



<그림 1> 극점에 의해 구성된 포먼트

3. 제안된 알고리즘에 의한 음색 변환

음성의 특징을 잘 나타내는 요소 중 하나는 성도의 특성이다. 성도는 성문으로부터 입술까지의 일종의 파이프 모양으로 이루어진 구조이다. 음성은 이 파이프의 모양에 따라 여러 종류로 변하게 된다. 즉 화자의 성별, 나이 및 신체적 특성, 발성하고자 하는 음소에 따라 달라지게 된다. 이런 원리는 파이프의 길이와 굵기에 따라 다른 소리가 생성되는 파이프 오르간에서 쉽게 이해할 수 있다[7]. 성도의 특성은 스펙트럼의 윤곽, 즉 포먼트에 잘 나타난다. 이러한 포먼트는 음성의 특성뿐만 아니라 화자의 특성 또한 잘 나타낸다[8]. 제1포먼트는 성도의 길이의 특성을 나타내 준다. 그러나 제1포먼트의 경우 성문의 특성이 포함되어 있기 때문에 때에 따라서 정확하게 찾지 못할 경우도 있다. 비음 또한 어느 포먼트 추정 알고리즘이라도 변환 함수 안에 영점이 존재하기 때문에 특별한 문제를 나타낸다. 비음 안에서 극점은 비강과 구강의 공명현상이다. 근처의 영점 때문에 F_2는 크기가 매우 줄어 종종 F_2와 일치하는 피크가 존재하지 않을 경우가 발생한다. 스펙트럼의 기울기 또한 음성의 특징을 나타내는 요소 중 하나이다. 기울기가 급할수록 탁한 음성이 발성되며 목의 촉촉한 정도에 따라 기울기가 변화한다. 이러한 음성의 특성을 나타내는 두 가지 요소를 원시 화자로부터 목적 화자로 변환시켜 줌으로써 음색 변환을 수행할 것이다. 제안된 음색 변환 알고리즘은 크게 세 가지 단계로 나눌 수 있다. 첫 번째 단계인 분석 단계에서는 원시 화자와 목적 화자의 음성에서 한 주기의 길이를 갖는 프레임 단위로 LPC 켑스트럼을 추출하고 DTW를 이용하

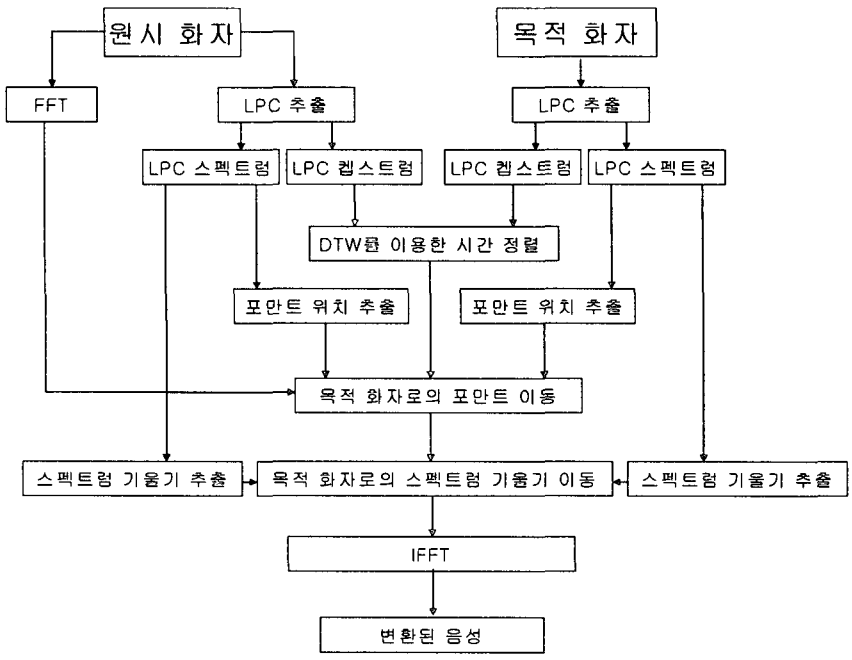
여 시간축 정렬을 수행한다. 또한 LPC 스펙트럼을 이용하여 포먼트 정보와 스펙트럼의 기울기 정보를 추출한다. 다음 단계인 변환 단계에서는 원시 화자의 음성 신호를 푸리에 변환하여 나온 결과에 대해서 분석 단계에서 추출한 포먼트 정보를 이용하여 포먼트의 위치를 목적 화자의 포먼트 위치로 이동시키고 스펙트럼 기울기 또한 목적 화자의 기울기로 변환시켜 준다. 마지막 단계인 합성 단계에서는 포먼트 위치와 기울기가 변환된 스펙트럼을 역 푸리에 변환을 해줌으로써 변환된 음성을 얻고 TD-PSOLA (Time Domain Pitch Synchronous Overlap and Add)를 수행하여 피치 정보를 변환시켜 준다.

3.1. 포먼트의 위치 변환

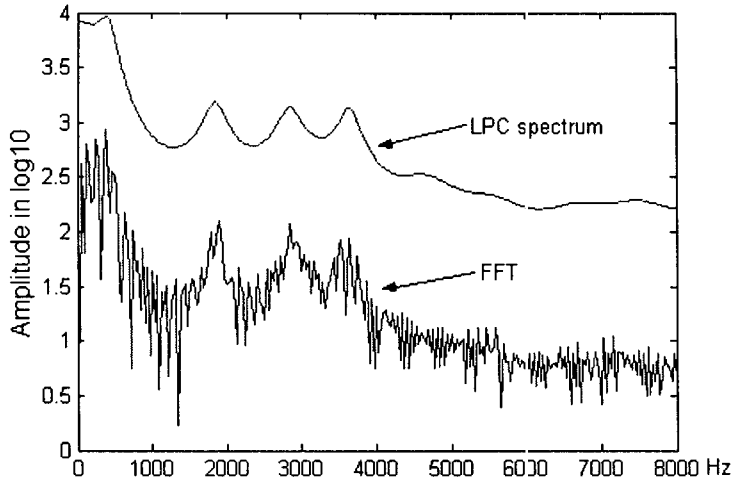
3.1.1. 포먼트 피크 추정 알고리즘

- 피크 피칭을 이용한 포먼트 추정 알고리즘.

모음 /a/를 대상으로 각 스펙트럼 포락선을 비교하면 LPC 스펙트럼과 FFT 켑스트럼의 스펙트럼이 비교적 스펙트럼 포락선을 잘 따라감을 알 수 있다. 특히 LPC 스펙트럼의 포락이 극점을 보다 잘 모델링함으로 LPC 스펙트럼을 이용하여 포먼트를 예측하는 것이 바람직하다.



<그림 2> 제안된 알고리즘의 구성도



<그림 3> FFT와 LPC 스펙트럼 포락선

위의 그림에서 보는 바와 같이 LPC 스펙트럼은 포먼트를 잘 따라가는 것을 볼 수가 있다. 우리는 LPC 스펙트럼으로부터 식 (3)과 (4)를 이용하여 포먼트의 위치를 추정할 수 있다.

(3)
$$a(i) = LS(i-1) - LS(i)$$
$$b(i) = LS(i) - LS(i+1)$$

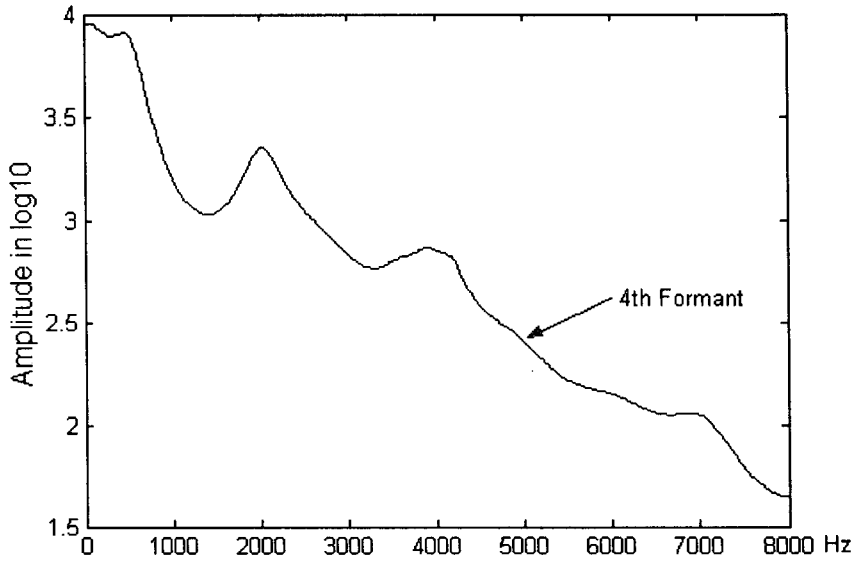
(4)
$$\text{if}(\ a(i)<0 \ \text{ and } \ a(i)*b(i)<0\ )$$

여기서 $LS(i)$은 LPC 스펙트럼의 i번째 값을 의미하며 $a(i)$는 과거값과 현재값의 차, $b(i)$는 현재값과 미래값의 차를 의미한다. 만약 식 (4)를 만족한다면 LPC 스펙트럼의 i번째 값이 최대값을 갖는 것이다. 즉 포먼트 위치를 의미한다. 그러나 위의 방법으로 포먼트의 위치를 구할 경우에 포먼트 위치를 찾지 못하는 경우가 발생한다.

<그림 4>와 같이 피크가 존재하지 않는 포먼트는 앞뒤 프레임의 피크점 2개씩, 총 4개의 피크값의 평균으로써 대체하여 수행하였다.

3.1.2. 포먼트 변환

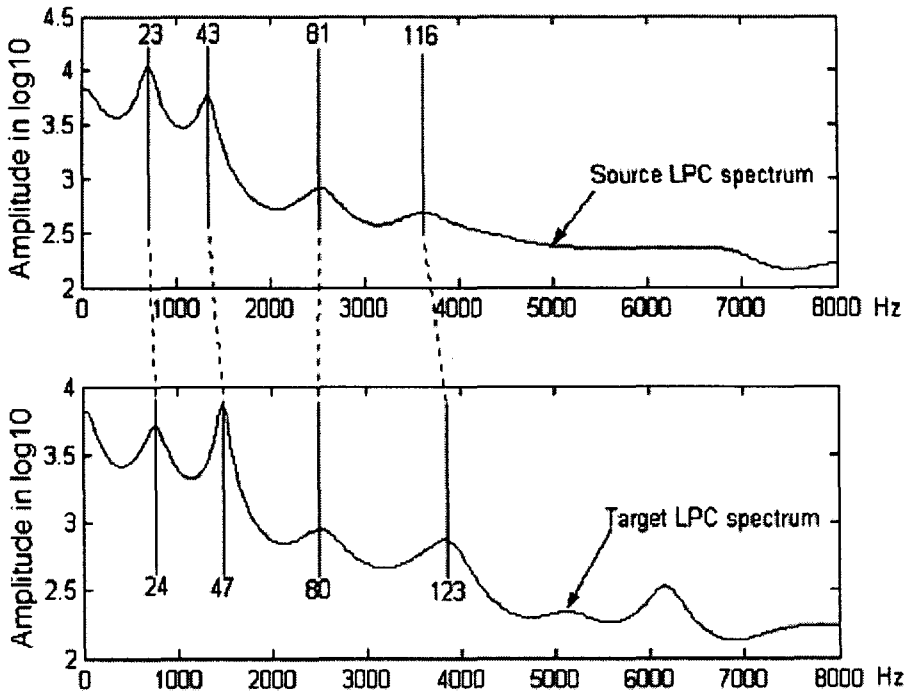
원시 화자로부터 목적 화자로의 포먼트 이동은 삽입(Interpolation)과 삭제(Decimation)의 선형적인 연산으로 이루어진다. 본 논문에서는 원시 화자와 목적 화자의 포먼트 위치 정보를 이용하여 제4포먼트까지 변화를 시켰다.



<그림 4> 포먼트가 피크를 가지지 않는 경우

먼저 LPC 스펙트럼에서 포먼트의 위치정보를 구하였으며, 얻어진 포먼트의 위치정보를 원 신호의 푸리에 변환 값에 직접 적용하여 포먼트의 위치를 변경하였다. 이 과정에서 푸리에 변환은 실수(Real) 값과 허수(Image) 값에 동시에 적용이 된다.

<그림 5>에서 원시 화자의 제1포먼트의 위치는 23포인트이고 목적 화자의 제1포먼트의 위치는 24포인트이다. 그러므로 원시 화자의 포먼트의 위치를 23포인트에서 24포인트로 변환시켜 주기 위해서 0포인트와 23포인트의 중간포인트인 12포인트에서 11포인트와 13포인트의 평균값을 12포인트에 삽입을 해주고 이전의 12포인트는 13포인트로 이동함으로써 삽입을 해 줄 수 있다. 또한 원시 화자의 제2포먼트에서 제3포먼트까지는 삭제를 해 주어야 한다. 삭제는 삽입과 연산이 비슷하다. 즉 삭제해야 할 포인트를 계산한 후 이를 두 포먼트 사이의 포인트 개수에서 나누어주게 되면 몇 번째마다 포인트를 삭제해 주어야 하는지 알 수가 있다. 위의 과정을 수행하기 위한 알고리즘을 단계별로 정리하면 다음과 같다.



<그림 5> 원시 화자의 포먼트와 목적 화자의 포먼트 위치 비교

<단계 1>

삽입 연산을 수행할 것인지 삭제 연산을 수행할 것인지를 결정한다.

그 때 if($SFP(i) \geq TFP(i)$)이면 삽입 연산을 수행하며 이외의 경우는 삭제 연산을 수행한다. 여기서 $SFT(i)$는 i번째 포먼트의 위치를 나타낸다.

<단계 2>

삽입이나 삭제해야 할 포인트 개수를 결정한다.

(5) $$N(i) = | \quad SFP(i) \quad - \quad TFP(i) \quad |$$

<단계 3>

해당 포먼트 사이에 몇 번째마다 삽입 또는 삭제를 해야하는지를 결정한다.

(6) $$S(i) = \frac{SFP(i) \; - \; SFP(i-1)}{N(i)}$$

여기서 $SFP(i-1)$은 이전 포먼트의 위치를 의미한다.

<단계 4>

삽입 또는 삭제를 수행한다.

for $n = SFP(i) - SFP(i+1)$

1)삽입의 경우

if($m == SFP(i) + n * N(i)$){

(7) $x'(m + Count) = \frac{x(n-1) + x(n+1)}{2}$

$Count++$

(8) $x'(m + Count) = x(n)$}

(9) $x'(m + Count) = x(n)$

여기서 $Count$는 현재까지 삽입한 개수를 의미한다.

2)삭제의 경우

if($m == SFP(i) + n * N(i)$){

$n = n+1$}

(10) $x'(m) = x(n)$

(11) $else\ x'(m) = x(n)$

3.2. 스펙트럼 기울기 변환

스펙트럼의 기울기는 화자의 운율 정보를 나타내는 중요한 요소 중 하나이다. 스펙트럼의 기울기의 경사가 급할 경우 탁한 소리의 특성을 가지며 경사가 만한 경우 웅웅거리는 소리의 특성을 가지게 된다.

원시 화자와 목적 화자의 LPC 스펙트럼으로부터 각 포먼트의 최고점/최저점을 구하고 이를 이용해 각 포먼트에서 다음 포먼트까지의 변화율, 즉 기울기를 구하여 신호의 스펙트럼을 바꾸어 준다.

(12) $$\tau(i) = \frac{TPV(i)}{SPV(i)}$$

여기서 SPV(i)는 i번째 원시 화자의 최고점/최저점의 크기이고, TPV(i)은 i번째 목적 화자의 최고(최저)점의 크기이다. 또한 $\tau(i)$은 원시 화자와 목적 화자의 최고(최저)값의 비를 의미한다. $\tau(i)$을 이용하여 다음 최고(최저)값으로 이동하는 기울기를 구할 수 있다.

(13) $$\beta = \frac{\tau(i-1) - \tau(i)}{SP(i-1) - SP(i)}$$

여기서 SP(i)는 원시 화자의 최고점/최저점의 위치를 의미한다. 위의 식 (13)에서 $SP(i-1) - SP(i)$으로 나누어주는 이유는 한 포인트 이동할 때의 변화량을 측정하기 위해서이다. 최종적으로 얻어지는 β를 이용하여 스펙트럼의 기울기를 변화시킨다.

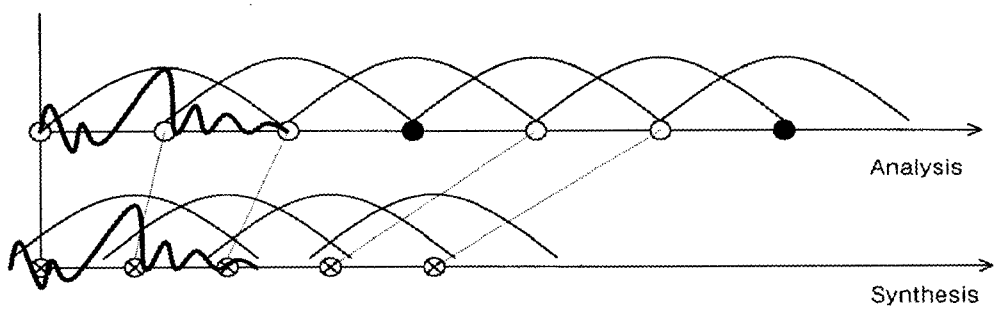
(14) $$x'(n) = x(n) * (\tau(i) + n * \beta)$$

여기서 $x(n)$은 n번째 신호의 값을 의미한다.

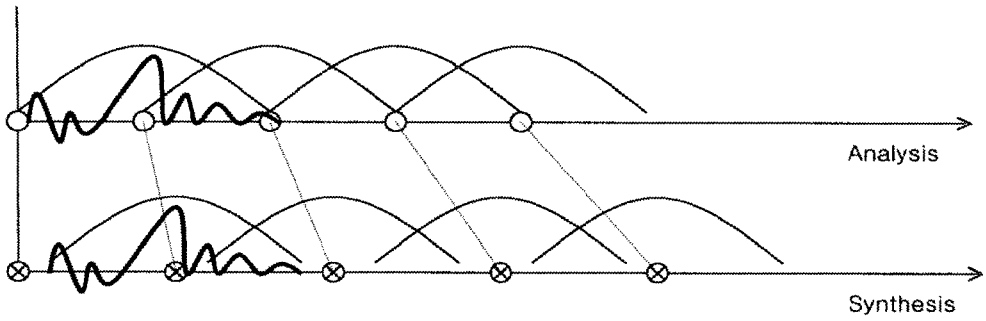
3.3. 삼각창(Triangular Window)을 이용한 TD-PSOLA

3.3.1. 기존의 PSOLA 방법

PSOLA 방법은 CNET의 France Telecom에서 개발되었다. PSOLA는 미리 녹음된 합성 단위를 부드럽게 연결시키고 지속시간과 피치를 조절하여 원하는 합성음을 얻는 방식이다[9][10]. 원 음성을 유성음 구간에서는 피치에 동기화하며, 무성음의 경우는 일정 간격으로 해닝창을 씌워서 짧은 구간의 신호로 분리하게 된다. 이렇게 분리된 신호를 이용하여 피치나 지속시간을 조절할 수 있다. 즉 합성음의 피치를 높일 경우는 짧은 구간 신호들을 재결합할 때 배치 간격을 좁히고, 피치를 낮출 경우는 짧은 구간 신호들의 배치 간격을 넓혀 재결합한다.



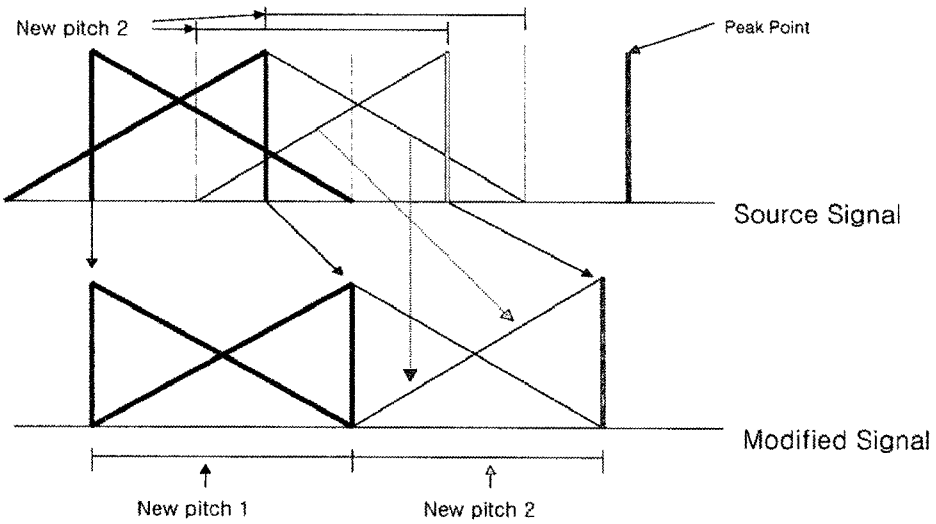
<그림 6> 피치를 감소시킬 경우



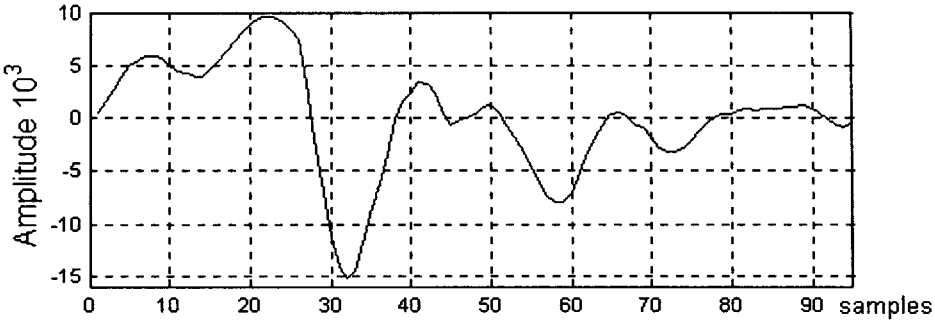
<그림 7> 피치를 증가시킬 경우

3.3.2. 삼각창을 이용한 TD-PSOLA

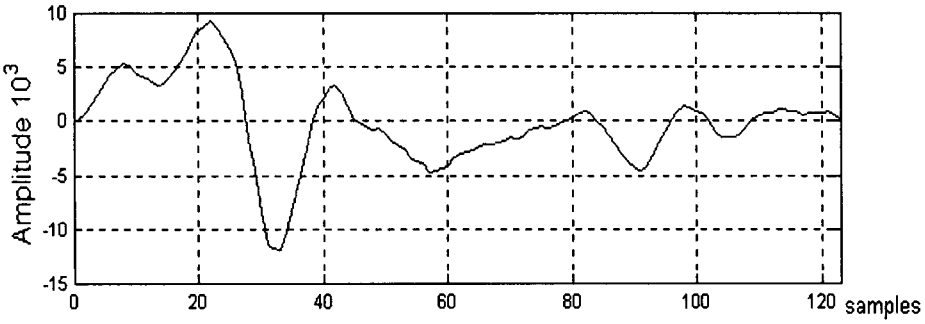
해닝창을 이용하여 PSOLA를 수행하였을 경우 합성되는 신호의 크기가 달라진다. 해닝창끼리 더해지는 부분의 창의 크기가 일치하지 않기 때문이다. 이러한 문제를 해결하기 위해 삼각창을 이용하여 피크 양쪽에 씌워지는 창의 크기를 서로 다르게 하여 합성된 신호의 크기가 변화하지 않도록 하였다.



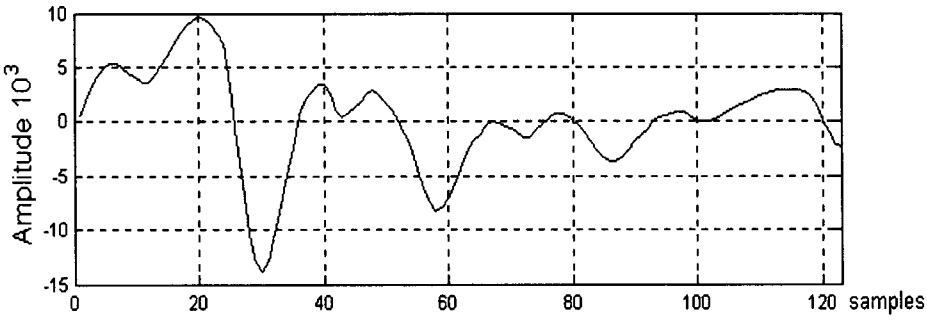
<그림 8> 피크 양쪽에 다른 길이의 삼각창을 이용한 경우



(a) 원시 화자의 신호



(b) 피크 양쪽에 같은 크기의 해닝창을 사용했을 경우



(c) 피크 양쪽에 다른 크기의 삼각창을 사용했을 경우

<그림 9> 수행 결과 비교

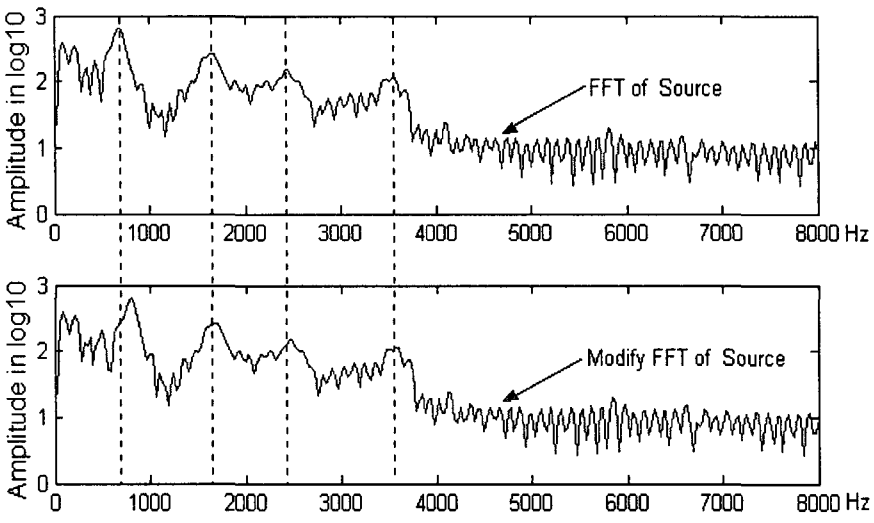
4. 실험 및 결과

실험 데이터로는 무향 녹음실 환경에서 수집한 남성화자 2명의 문장을 사용했다. <표 1>은 실험에 사용된 실험 조건을 나타내며, DTW를 수행하기 위해 사용되는 거리의 척도는 유클리디안(Euclidean) 거리를 사용하였다.

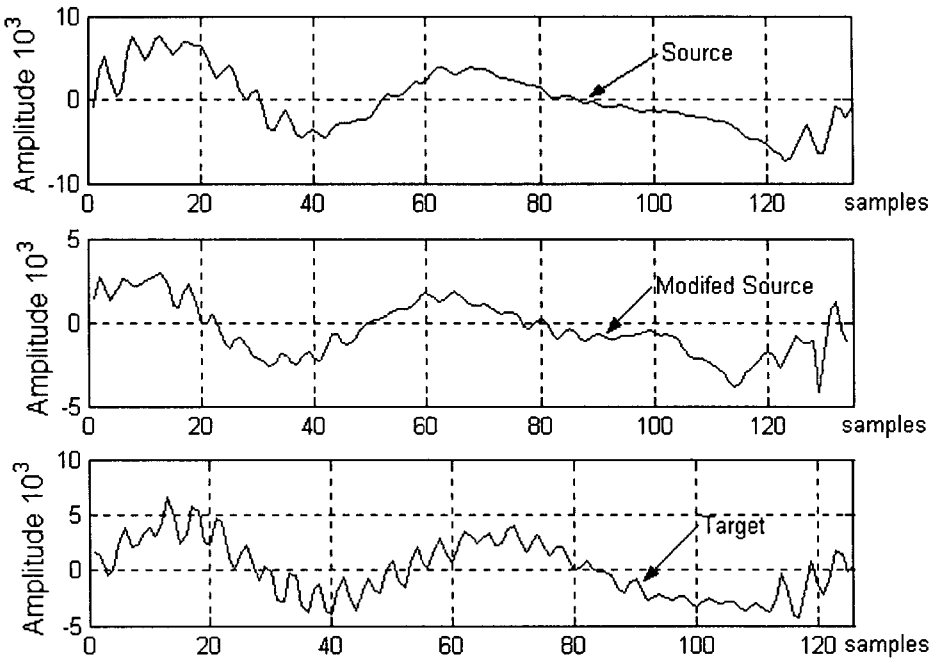
<그림 10>은 한 프레임, 즉 한 주기를 분석하여 나타낸 것이며 한 주기 신호의 푸리에 변환한 값의 크기에 로그를 취한 결과를 나타낸 것이다. 원시 화자의 포먼트를 대상 화자의 포먼트로 이동한 결과를 보여준다. <그림 10>에서는 제1포먼트에서의 이동이 많았으며 제2, 제3, 제4포먼트는 거의 이동하지 않았다.

<표 1> 실험 조건

샘플링 주파수	16kHz	**변환된 포먼트 수**		4개(F1~F4)
LPC 분석방법	자기상관 방법	**창함수**	**분석**	해닝 창함수
LPC 차수	18차		**PSOLA**	삼각 창함수
LPC 켑스트럼 차수	18차	**푸리에 분석 포인트**		512 포인트
프레임 크기	한 주기			



<그림 10> 포먼트 위치 이동 결과



<그림 11> 변환된 파형 비교

5. 결 론

본 논문에서는 원시 화자의 음성을 목적 화자로의 음성으로 변환하는 기법인 음색 변환을 수행하였다. 제시한 새로운 알고리즘은 LPC를 특징요소로 하여 음색 변환을 수행했을 경우, 잔차신호와 변환된 LPC를 컨볼루션할 때 발생되는 왜곡으로 인한 음질 저하가 나타나게 된다. 이러한 단점을 개선하고자 음색 변환을 주파수 영역에서 수행하였다. 알고리즘에 사용되는 특징 요소로는 성도의 특성을 나타내는 포먼트와 스펙트럼 기울기를 사용하였으며 운율 정보를 나타내는 음성의 피치주기는 PSOLA를 수행하여 변환하였다. 실험 결과는 LPC를 이용하여 음색 변환을 수행한 음질보다는 주파수 영역에서 변환을 수행했을 때가 좀 더 좋은 음질의 결과를 얻을 수가 있었다.

향후 계획으로는 성문의 특성 및 포먼트 정보를 대용량의 데이터로부터 벡터 양자화를 수행하여 음색 변환에 사용하는 것이다. 또한 주파수 영역에서의 하모닉스 성분을 목적 화자의 것으로 변환하기 위해 벡터양자화를 이용할 것이다.

참 고 문 헌

[1] M. Abe, S. Nakamura et al., "Voice Conversion though Vector Quantization", *Proc. ICASSP,* pp.665-658, 1988.

[2] H. Balbret, E. Moulines, J. P. Tubach, "Voice Transformation using PSOLA technique", *Speech Communication,* Vol. 11, pp.175-187, 1992.

[3] 권홍석, 배건성, "선형다변회귀모델과 LP-PSOLA 합성 방식을 이용한 음성변환", *한국음향학회지,* 20권, 3호, pp.15-23, 2001.

[4] S. Chiba, H. Sakoe, "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans. Acoust. Speech signal Process.,* Vol. ASSP-26, pp.43-49, 1978.

[5] B. Carnahan, H. A. Luther, J. O. Wilkes, *Applied Numerical method*, John Wiley & Sons, Inc., 1969.

[6] H. Mizuno, M. Abe, "Voice conversion algorithm based on piecewise linear conversion rule of formant frequency and spectrum tilt", *Speech Communication,* Vol. 16, pp.153-164, 1995.

[7] L. R. Rabiner, R. W. Schafer, "*Digital Processing of Speech Signal*", Prentice Hall International. Inc., 1978.

[8] H. Kuwabara, K. Ohgushi, "Contributions of vocal track resonant frequencies and bandwidths to the personal perception of speech", *Acoustica,* Vol. 63, pp.120-137, 1987.

[9] F. J. Charpentier, M. G. Stella, "Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation", *Proc. ICASSP'86 Tokyo*, pp.2015-2018, 1986.

[10] E. Moulines, F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication,* Vol. 9, pp.453-467, 1990.

접수일자: 2002년 11월 16일
게재결정: 2002년 12월 11일

▶ 손성용(Song-Young Son)
주소: 305-732 대전광역시 유성구 화암동 58-4번지 한국정보통신대학원대학교
소속: 한국정보통신대학원대학교(ICU) 음성/음향 정보 연구실
전화: 042) 866-6196
E-mail: thill@icu.ac.kr

▶ 한민수(Min-Soo Hahn)
주소: 305-732 대전광역시 유성구 화암동 58-4번지 한국정보통신대학원대학교
소속: 한국정보통신대학원대학교(ICU) 음성/음향 정보 연구실
전화: 042) 866-6123
E-mail: mshahn@icu.ac.kr