정현파 모델링을 이용한 폴리포닉 오디오 신호의 시간축 변화

Time-Scale Modification of Polyphonic Audio Signals Using Sinusoidal Modeling

장 호 근*, 박 주 성**
(Ho Keun Jang*, Ju Sung Park**)

*현대전자 시스템 IC, **부산대학교 전자공학과

(접수일자: 2000년 1월 15일; 수정일자: 2000년 9월 19일; 채택일자: 2001년 1월 29일)

본 논문에서는 폴리포닉 음과 같은 복잡한 스펙트럼을 갖는 오디오 신호를 정현파 성분으로 모델링하고, 이를 바탕으로 고음질의 시간축 변화된 음을 얻는 방법을 제안한다. 입력 신호는 옥타브 밴드 구조의 다중 해상도 필터 뱅크를 통과하고 여기에서 나온 각 서브밴드 신호로부터 정현파 성분이 추출된다. 서브밴드 신호의 정현파 분석시 정현파 성분을 추출하는 구간의 크기를 국지적인 신호의 특성에 따라 다르게 해 주는 동적 세그멘테이션 방법을 적용한다. 이렇게 함으로써 기존 정현파 모델링에서 신호의 천이 구간에서 발생하는 퍼짐 현상을 개선하고, 시간축 변화시에도 원래 음에 가까운 음질을 얻을 수 있다. 정현파 분석을 위한 스펙트럼 분석 도구로는 심리 음향 모델을 적용한 matching pursuit을 사용함으로써 정현파 성분의 갯수를 줄이고, matching pursuit의 반복 과정에 대한 합리적인 정지 조건을 제공할 수 있다. 정현파 성분으로 표현하기 어려운 신호의 잡음 성분은 원래 신호에서 정현파 성분으로 합성된 신호를 뺀 것으로 얻을 수 있으며, 스펙트럼 포락선 근사화 방법으로써 모델링된다. 본 논문의 알고리즘을 적용해 다양한 폴리포닉 음에 대해 실험한 결과 제안한 정현파 모델링 방법이 원래 신호의 음질을 잘 복원할 수 있고, 시간축 변화율이 큰 경우에도 신호의 천이 구간을 잘 표현할 수 있음을 확인하였다.

핵심용어: 정현파 모델링, 시간축 변화, 다중 해상도 필터 뱅크, 동적 세그멘테이션, Matching pursuit, 심리음향 모델
투고분야: 음악음향 및 음향심리 분야 (8.6)

This paper proposes a method of time-scale modification of polyphonic audio signals based on a sinusoidal model. The signals are modeled with sinusoidal component and noise component. A multiresolution filter bank is designed which splits the input signal into six octave-spaced subbands without aliasing and sinusoidal modeling is applied to each subband signal. To alleviate smearing of transients in time-scale modification a dynamic segmentation method is applied to subbands which determines the analysis-synthesis frame size adaptively to fit time-frequency characteristics of the subband signal. For extracting sinusoidal components and calculating their parameters matching pursuit algorithm is applied to each analysis frame of subband signal. In accordance with spectrum analysis a psychoacoustic model implementing the effect of frequency masking is incorporated with matching pursuit to provide a resonable stop condition of iteration and reduce the number of sinusoids. The noise component obtained by subtracting the synthesized signal with sinusoidal components from the original signal is modeled by line-segment model of short time spectrum envelope. For various polyphonic audio signals the result of simulation shows suggested sinusoidal modeling can synthesize original signal without loss of perceptual quality and do more robust and high quality time-scale modification for large scale factor because of representing transients without any perceptual loss.

Key words: Sinusoidal modeling, Time-scale modification, Mutiresolution fiter bank, Dynamic segmentation, Matching pursuit, Psychoacoustic model
***Subject classification:** Musical acoustics and Psychoacoustics* (8.6)

I. 서 론

오디오 신호의 시간축 변화는 신호의 주파수 성분을 변화시킴으로써 음정이나 음색을 변화시키는 주파수 변화와 더불어 음성 합성과 컴퓨터 음악의 사운드 분석/합성 분야에서 활발하게 연구되어 온 내용이다. 이것은 방대한 오디오 신호를 데이터 베이스화시켜 놓고 이를 빠르게 검색하고자 할 때, 외국어 어학 학습시 사람의 말하는 속도를 느리게 변화시켜 듣기 능력을 향상시키고자

책임저자: 장호근 (jonghk@hei.co.kr)
135-738 서울시 강남구 대치동 891 영동빌딩
현대전자 디지털미디어 IAV팀
(전화: 02-3459-3234; 팩스: 02-3459-5843)

할 때, 그리고 영상 미디어 제작시 비디오 신호와의 시간 동기를 위해 오디오 신호의 재생 시간을 조절할 때 등에 유용하게 쓰일 수 있다. 또한 컴퓨터 음악의 합성시 음질을 향상시키는데도 이용할 수 있다. 신디사이저나 노래방 기기 등에서는 연주되는 곡의 박자나 음정을 변화시키는 것을 기본적인 기능으로 제공하고 있는데, 여기에 사용되는 사운드 합성 방식은 미리 개별 악기음의 샘플을 메모리에 저장시켜 놓고 이를 바탕으로 입력되는 박자와 음정에 맞게 연주 음을 합성해내는 것이다[15]. 그러나 좀 더 고급화된 음질을 위해서는 미리 실제 연주된 음악 신호를 저장해서 재생시키는 것을 생각해 볼 수 있고, 박자나 음정을 변화시키고자 할 때는 오디오 신호 자체에 대한 변화 방법이 필요하다. 여기에 시간축 변화와 같은 알고리즘이 적용될 수 있다.

1980년대 중반에 나온 정현파 모델링[1-3]은 단구간 퓨리에 변환 (Short-time Fourier transform: STFT)을 통해 분석되는 값을 신호에 포함된 실제 정현파 성분의 주파수와 진폭으로 모델링함으로써 시간축 변화나 주파수 변화와 같은 신호 변형이 쉽고, 우수한 음질을 얻을 수 있는 방법으로 알려져 있다. 그러나 정현파 모델은 다음과 같은 몇 가지 문제점을 지니고 있다.

첫째, 정현파 모델은 신호의 특성이 안정적이고 주기적이라는 전제하에서 신호를 잘 모델링한다. 이것은 스펙트럼이 랜덤하게 변화하는 잡음과 같은 신호에는 잘 적용이 되지 않는다는 것을 의미한다. Serra[4]는 이러한 문제점을 해결하기 위해 신호 모델에 정현파 성분과 함께 잡음 성분을 추가하여 여기에는 다른 모델링 방법을 적용함으로써 신호에 포함된 잡음 성분이 신호의 복원이나 변형시 잡음의 특성을 그대로 유지하면서 음질에 기여할 수 있도록 하였다.

둘째, 정현파 모델링의 분석/합성 과정에서 일정한 크기의 과형 구간에서 분석/합성이 이루어짐에 따라 짧은 시간 내에 급격하게 신호의 특성이 변화하는 어택 (attack)과 같은 천이 구간 (transients)을 잘 표현할 수 없다. 이를 해결하기 위해 천이 구간만을 따로 분리시켜 파형을 그대로 복원할 수 있는 다른 방법을 사용해 모델링하는 방법이 있다[5-6]. 이렇게 하면 시간축 변화시에도 음질을 원래 음과 동일하게 할 수 있지만, 천이 구간에서 정현파 모델링과 전혀 다른 모델을 사용함으로써 복잡도를 증가시키고, 신호 변형시에도 천이 구간은 고정되어 있으므로 천이 구간이 많은 음악 신호의 경우 큰 폭으로 시간축을 변화시키면 속도 조절에 문제를 가져올 수 있다.

셋째, 지금까지 정현파 모델링에 대해 연구되어 온 것은 주로 음성 신호나 단일 악기음에 대한 것으로 이것을 여러 가지 악기음이 복잡하게 섞여 있는 폴리포닉 음에는 적용하기 어렵다. 음성 신호나 단일 악기음의 경우 정현파 분석 과정에서 필요한 윈도우의 크기, FFT 크기, 그리고 윈도우 적용 간격 등의 분석 파라메터들은 주로 신호의 피치 주파수로부터 결정되지만, 폴리포닉 음은 단일 피치 주파수를 정의할 수 없으므로 일정한 분석 파라메터를 결정하기 어렵다.

본 논문은 폴리포닉 음과 같은 복잡한 스펙트럼을 갖는 오디오 신호의 시간축 변화 방법에 대한 연구로 기존의 정현파 모델링을 바탕으로 하고 있다. 2장에서는 본 논문의 앞 단계로서 [8]번 논문에서 제안하고 있는 동적 세그멘테이션을 적용한 다중 해상도 정현파 분석 방법에 대해서 간략하게 설명한다. 이것은 폴리포닉 오디오 신호에 대한 정현파 분석시에 발생하는 윈도우 크기 문제를 해결할 수 있고, 시간축 변화시 천이 구간의 음질을 보존할 수 있다. 3장에서는 여러 개의 음이 섞여 있는 폴리포닉 오디오 신호의 스펙트럼에서 각 음에 대한 정현파 성분을 효과적으로 찾아내기 위한 방법을 기술한다. 4장에서는 폴리포닉 오디오 신호를 정현파 모델링하고 시간축 변화시킨 음을 생성시키는 전체 알고리즘을 설명하고, 여러가지 오디오 신호에 대한 실험 결과를 보여준다. 마지막으로 5장에서는 결론을 맺도록 한다.

II. 천이 구간을 고려한 다중 해상도 정현파 분석

2.1. 다중 해상도 정현파 분석

폴리포닉 음의 정현파 모델링에서 분석 윈도우 크기를 결정하기 어려운 문제를 해결하기 위한 하나의 방법은 신호를 대역폭이 제한된 여러 개의 서브밴드 신호로 나누고, 이 서브밴드 신호로부터 정현파 분석을 행하는 것이다. 필터 뱅크를 이용한 정현파 모델링 방법은 Levine[6]을 비롯한 몇 사람에 의해 소개되었는데, 그 중 Levine이 제안한 방법이 폴리포닉 오디오 신호를 분석하기에 적합하다고 알려져 있다. 본 논문에서는 이것을 바탕으로 그림 1과 같이 신호의 주파수 영역을 서로 다른 대역폭을 가지는 6개의 서브밴드로 나누었다. 각 서브밴드는 저주파수 대역으로 갈수록 상위 서브밴드의 대역폭보다 크기가 1/2인 옥타브 밴드 구조를 갖는다. 이와 같은 구조의 다중 해상도 필터 뱅크는 저주파수 대역에서는 주파수 해상도를 높이고, 고주파수 대역에서는 시간 해상도를 높이는 장점을 지닌다. Complementary 필터 뱅크의 구조는 그림 2와 같이 데시메이션 필터와 데시메이터로 간단하게 구현할 수 있다. 이 필터 뱅크의 고주파 출력 신호는 2배로 오버샘플링 되어 있고, 저주파 출력 신호는 데시메이션 필터에 의해 알리아싱이 방지되므로 알리아싱이 제거된 서브밴드 신호를 얻을 수 있다. 각 옥타브 신호는 complementary 필터 뱅크의 데시메이터에 의해 각각 1배, 2배, 4배, 8배, 16배, 32배로 데시메이션된 신호이다.

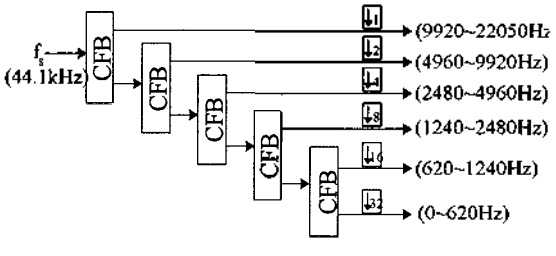


그림 1. 정현파 분석을 위한 다중 해상도 필터 뱅크
Fig. 1. Multiresolution filter bank for sinusoidal analysis.

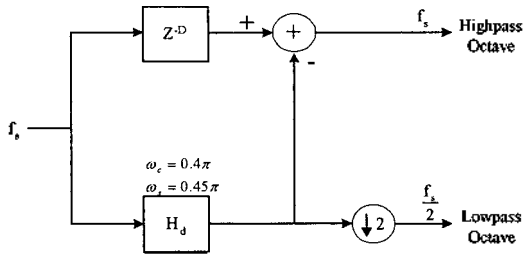


그림 2. Complementary 필터 뱅크
Fig. 2. Complementary filter bank.

2.2. 동적 세그멘테이션을 이용한 천이 구간의 모델링

시간축 변화된 신호의 음질이 손상되지 않도록 하기 위해서는 악기음의 어택과 같은 천이 구간이 잘 보존되어야 한다. 이것은 주파수 스펙트럼으로부터 모델 파라메터를 추출할 수 있는 잡음과는 다른 특성을 지니는 신호 성분이다. 정현파 모델이 천이 구간을 잘 모델링하지 못하는 것은 신호의 파라메터 분석과 합성이 일정한 시간 동안의 신호 구간에서 행해지기 때문이다. 천이 구간은 국지적으로 매우 큰 변화를 일으키므로 일정 구간을 잡아서 분석한 결과로부터 합성하게 되면 어택 부분은 시간상으로 퍼져버리는 결과를 가져오게 된다. 이것은 시간축 확장시 음질을 더욱 열악하게 만드는 문제이다. George[3]는 신호의 정적 구간 (stationary parts)과 더불어 천이 구간과 잡음 성분까지 정현파로 모델링하는 방법을 제안하였다. 이것은 천이 구간의 파형을 잘 표현할 수 있지만, 잡음 성분까지 정현파로 모델링되어 변형되므로 신호의 변형시 이를 따로 모델링하는 방법보다 더 나은 음질을 낸다는 장담을 할 수 없다. Hamdy[5]와 Levine[6]은 천이 구간을 따로 분리시켜 그 부분의 파형을 그대로 복원할 수 있는 다른 코딩 방법이나 모델링 방법을 적용하고, 시간축 변화시에는 복원된 파형을 단순히 시간축 상에서 이동시키는 방법을 사용하였다. 그러나 이 경우 천이 구간을 효과적으로 찾아내고 분리시키는 것도 문제가 되지만, 천이 구간에 다른 신호 모델을 적용하므로 시스템의 복잡도를 증가시키고 시간축 변화시 천이 구간에서는 변화가 일어나지 않으므로 속도 조절에 문제를 가져올 수 있다. 한편 Verma[14]는 천이 구간을 DCT (Discrete Cosine Transform) 변환시킨 신호에 대해 정현파 모델링을 행하였으나, 정현파 모델링이 원래 신호를 완벽하게 복원하지 못하므로 여러 번의 변환 과정을 거치게 되면 천이 구간의 파형을 그대로 복원하기 어렵게 될 수도 있다. Goodwin[7]은 천이 구간 근처에서 정현파 분석/합성이 이루어지는 프레임 크기를 작게 함으로써 정현파 모델을 그대로 적용시키면서도 효과적으로 천이 구간을 모델링하는 방법을 제안하였다. 이 방법은 천이 구간에서도 정현파 모델을 통해 신호가 표현되므로 시스템 구조가 간단해지고, 시간축 변화시에도 천이 구간의 특성을 살리면서 안정적인 속도 변화가 가능하다는 장점이 있다.

신호의 구간별 특성에 따라 서로 다른 프레임 크기를 할당하는 것을 동적 세그멘테이션이라고 한다. Goodwin이 제안한 동적 세그멘테이션 방법들은 분석 및 합성 프레임 크기를 결정하기 위해 정현파 모델링의 전 과정을 통해 원래 신호를 복원해야 하므로 계산량이 많고, 폴리포닉 음에는 잘 적용이 되지 않는다. 본 논문에서는 역퓨리에 변환 (IFFT)을 이용한 정현파 합성으로 계산량을 줄이고, 원래 신호와 합성 신호와의 오차 에너지를 정의하는 방법을 다르게 함으로써 천이 구간 근처에서만 프레임 크기를 작게 하고 나머지 부분에서는 주파수 해상도를 높일 수 있도록 프레임 크기를 크게 하는 동적 세그멘테이션 방법[8]을 적용하였다.

다중 해상도 필터 뱅크를 통과해서 나온 서브밴드 신호에 대해 동적 세그멘테이션을 이용하여 정현파 분석과 합성을 위한 구간별 프레임 크기를 결정한다. 그림 3은 하나의 서브밴드 신호에 대한 동적 세그멘테이션 결과를 나타낸다. 그림에서 보면 천이 구간 근처에서는 프레임 크기가 작고, 나머지 구간에서는 프레임 크기가 큰 것을 확인할 수 있다.

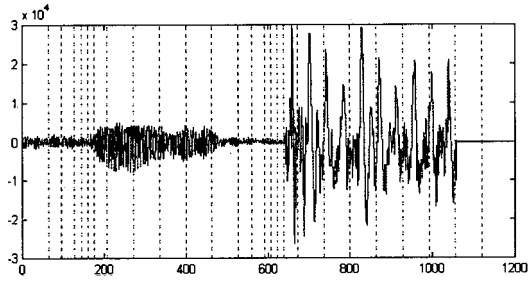


그림 3. 서브밴드 신호에 대한 동적 세그멘테이션 결과
Fig. 3. The result of dynamic segmentation for a subband signal.

III. Matching pursuit을 이용한 정현파 성분의 추출

각 서브밴드 신호에 대해 신호 구간마다 프레임 크기가 결정되면, 다음 단계는 각 프레임에서 스펙트럼 분석을 통해 정현파 성분을 찾아내고 정현파 성분에 대한 파라메터를 추출하는 것이다. 본 논문에서는 정현파 추출을 위한 스펙트럼 분석 도구로 matching pursuit[9-10]을 사용한다. Matching pursuit은 반복 과정을 통해서 스펙트럼 상에서 가장 중요한 정현파 성분만을 차례로 찾아냄으로써, 단순한 피크 선택 방법보다 여러 개의 음의 주파수 성분이 섞여 있는 폴리포닉 음의 정현파 분석에 유리하다. 또한 이 과정은 FFT를 통해 구현될 수 있으므로 계산량 측면에서도 장점을 지닌다. Matching pursuit과 더불어 심리음향 모델에 기초한 주파수 마스킹 효과를 적용하면 matching pursuit의 반복 과정에 대한 합리적인 정지 조건을 제공하고, 정현파의 수도 줄일 수 있다.

3.1. Matching pursuit을 이용한 정현파 성분의 추출

Matching pursuit은 시간 제한된 신호를 전개 함수 (expansion function)라고 부르는 임의의 신호의 선형 조합으로 나타내는 것이다. 전개 함수는 overcomplete dictionary라고 부르는 하나의 벡터 공간을 형성하며, 이것은 일반적으로 신호를 전개시키는 orthogonal basis보다 많은 수의 벡터를 포함한다. Matching pursuit에서 전개 함수를 찾는 방법은 다음과 같이 설명할 수 있다.

어떤 신호 $x[n]$에 대한 overcomplete dictionary가 D이고, 여기에 속한 임의의 벡터를 $d_m[n]$이라 할때 atching pursuit의 i번째 반복 과정에서는 다음의 식에서 나타나는 $r_{i+1}[n]$의 L_2 놈 (norm)을 최소로 하는 전개 함수 $d_m^i[n]$을 찾게 된다.

$$r_{i+1}[n] = r_i[n] - \alpha_i d_m^i[n] \tag{1}$$

여기서 $r_i[n]$은 i-1번째까지의 반복 과정에서 신호 $x[n]$을 구성하는 전개 함수를 찾고 난 후에 남는 신호 (residual)이며, $d_m^i[n]$은 i번째 반복 과정에서 선택되는 전개 함수를 나타내고, α_i는 이 전개 함수와 관련된 전개 계수를 나타낸다. 첫 번째 반복 과정에서 $r_i[n]$은 $x[n]$이 될 것이다. 표기를 간략하게 하기 위해 $g_i[n] = d_m^i[n]$이라 할 때 $g_i[n]$은 다음과 같이 나타낼 수 있다.

$$g_i = \min_{g_i \in D} \| r_{i+1} \|^2 = \min_{g_i \in D} \| r_i - \alpha_i g_i \|^2 \tag{2}$$

$\| r_{i+1} \|^2$이 최소가 되기 위해서는 r_{i+1}과 g_i가 orthogonal 관계에 있어야 하므로,

$$< r_{i+1}, g_i > = < r_i - \alpha_i g_i, g_i > = (r_i - \alpha_i g_i)^H g_i = 0 \tag{3}$$

여기서 $\langle a, b \rangle = a^H b$이고, H는 복소 conjugate를 나타낸다.

식 (3)을 전개하면

$$(r_i - \alpha_i g_i)^H g_i = r_i^H g_i - \alpha_i^H \| g_i \|^2 = 0 \tag{4}$$

따라서, 모든 전개 함수의 크기가 1이라고 할 때 α_i는 다음과 같이 구해진다.

$$\alpha_i = \frac{\langle g_i, r_i \rangle}{\| g_i \|^2} = \langle g_i, r_i \rangle \tag{5}$$

이것을 r_{i+1}의 L_2 놈을 구하는 식에 대입하면 다음과 같다.

$$\begin{aligned} \| r_{i+1} \|^2 &= (r_i - \alpha_i g_i)^H (r_i - \alpha_i g_i) \\ &= \| r_{i+1} \|^2 - | \alpha_i |^2 \end{aligned} \tag{6}$$

따라서 $\| r_{i+1} \|^2$을 최소로 하기 위해서는 dictionary 벡터들과 r_i의 correlation에서 가장 큰 계수값을 가지는 전개 함수를 선택하면 되고, 이 때 이 계수값은 식 (5)와 같이 계산된다.

Matching pursuit은 반복 과정을 무한히 할 때 원래 신호를 그대로 복원할 수 있다. 일반적으로 유한값 R번째까지의 반복 과정으로부터 추출된 전개 함수들로부터 합성되는 신호는 다음과 같이 나타낼 수 있다.

$$\hat{x}[n] = \sum_{i=1}^{R} \alpha_i g_i[n] \approx x[n] \tag{7}$$

Matching pursuit을 정현파 모델링에 적용하기 위해서는 먼저 dictionary가 정의되어야 한다. 정현파 성분의 파라메터는 주파수, 진폭, 위상으로 나타낼 수 있으므로 이 세 가지를 모두 고려한 dictionary는 매우 많은 벡터를 포함하게 된다. 따라서 다음과 같이 주파수 파라메터만으로 표현되는 복소 지수 함수를 dictionary 벡터로 구성한다.

$$D = \left\{ d_m[n] = \frac{1}{N} e^{j\frac{2\pi}{K}mn}, n = 0,1,..N-1 \right\}, \quad m = 0,1,...,K-1 \tag{8}$$

여기서 N은 신호의 크기 (dimension), K는 dictionary의 벡터 개수를 나타낸다.

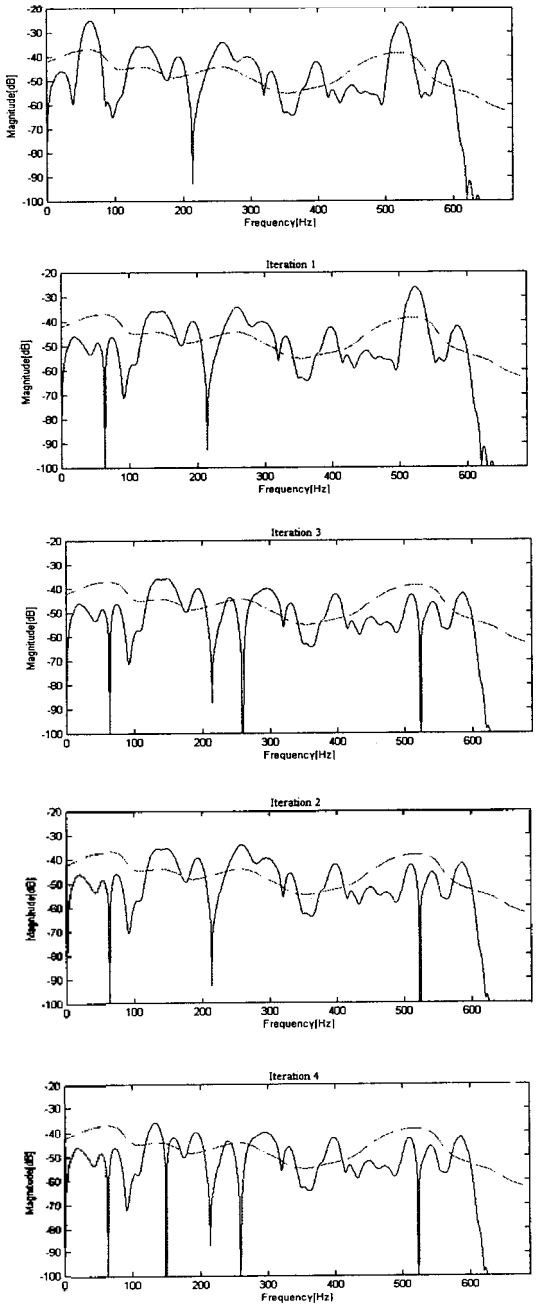
본 연구에서는 다중 해상도 필터 뱅크에서 나오는 각 서브밴드 신호에 대해 matching pursuit을 적용하여 정현파 성분을 추출한다. 각 서브밴드 신호는 동적 세그멘테이션에 의해 신호 구간별로 정현파 분석을 위한 최적의 프레임 크기가 결정되고, 이 분석 프레임에 matching pursuit을 적용하여 실제로 신호를 구성하는 가장 중요한 정현파 성분부터 차례로 추출한다. 여러가지 음악 신호에 대해 실험한 결과, 각 서브밴드 신호의 하나의 분석 프레임에서 15~20개 정도의 정현파 성분을 추출하면 원래 신호와 거의 구별할 수 없는 음질의 신호를 합성할 수 있음을 확인하였다.

3.2. 심리 음향 모델을 이용한 정현파 분석

MPEG (Moving Pictures Expert Group)와 같은 오디오 신호의 압축 코딩 방법에서는 심리 음향 모델의 중요한 결과인 주파수 마스킹 효과를 사용해 샘플값을 나타내는 비트 수를 줄이게 된다. 주파수 마스킹은 같은 시간에 발생된 신호 성분에서 큰 진폭을 가진 주파수 성분 주변에 위치한 약한 진폭의 주파수 성분은 강한 신호 성분에 의해 마스킹되어 소리가 들리지 않는 것을 말한다. 주파수 마스킹 효과는 정현파 분석에서 유효하게 사용될 수 있다. 먼저 스펙트럼에서 피크를 선택할 때 마스킹 임계치 이상의 피크만을 선택하면 사이드로브 (sidelobe) 피크나 음의 인지상 중요하지 않은 정현파 성분을 제거할 수 있으므로 정현파의 개수를 줄일 수 있다. 또한 마스킹 임계치를 구

하는 과정에서 정현파 성분과 잡음 성분에 대한 임계치를 따로 구하므로, 스펙트럼에서 나타나는 피크들에 대해 정현파 성분과 잡음 성분으로 나눌 수 있는 근거를 제공한다.

본 논문에서는 MPEG-AAC (Advanced Audio Coding) [11]에서 사용하는 심리 음향 모델을 정현파 분석에 적합하도록 변형시켜 마스킹 임계치를 구하였다. 심리 음향 모델에서는 주파수 대역을 사람 귀의 특성에 따라 나누어 놓은 임계 대역 (critical band)별로 마스킹 임계치가 계산된다. 그러나 정현파 분석을 위해서는 개개의 주파수에 대한 마스킹 임계치를 구하는 것이 더 정확하므로, 본 논문에서는 FFT결과에서 나오는 개개의 주파수에서 마스킹 임계치를 구하는 것으로 알고리즘을 변형하였다.



Matching pursuit을 이용한 정현파 성분 추출시 마스킹 임계치는 중요한 정현파 성분을 찾아내는 것뿐만 아니라, matching pursuit 알고리즘의 정지 조건을 제공한다. 마스킹 임계치보다 위에 있는 주파수 성분이 정현파 성분을 나타낸다고 볼 수 있으므로 반복적인 정현파 성분 추출 과정은 정현파 성분의 진폭이 마스킹 임계치 밑으로 내려갔을 때 정지한다. 이때까지 찾아낸 정현파 성분은 신호에 포함된 중요한 정현파 성분을 나타낸다고 볼 수 있다.

그림 4에는 마스킹 임계치를 적용한 matching pursuit을 통해 정현파 성분을 추출하는 반복 과정을 나타내었다.

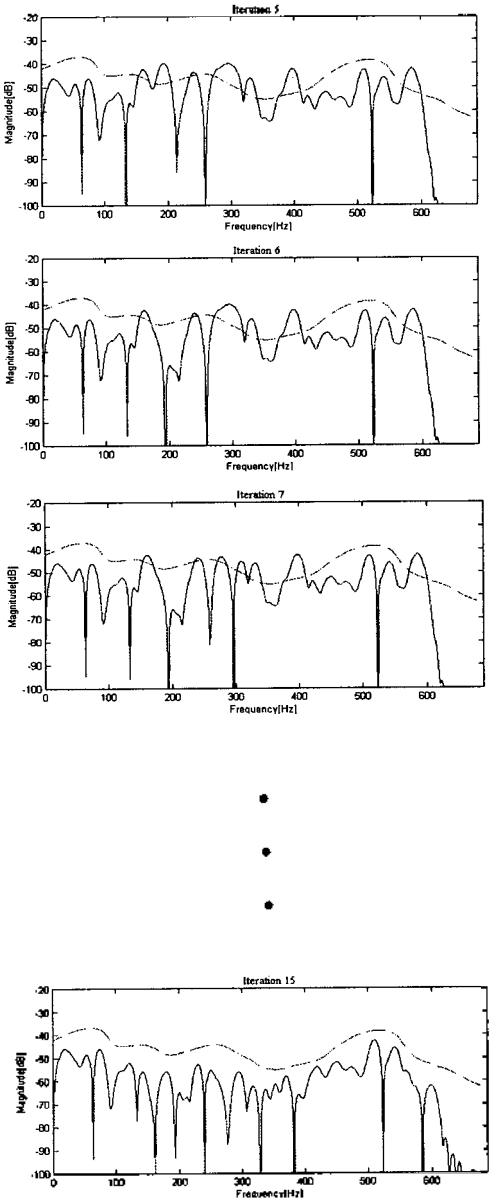


그림 4. 마스킹 임계치를 적용한 matching pursuit
Fig. 4. Matching pursuit with masking threshold.

IV. 실험 및 고찰

IV. 실험 및 고찰

4.1. 시스템 개요

그림 5는 본 논문에서 제안하는 폴리포닉 오디오 신호의 시간축 변화를 위한 정현파 모델링의 분석/합성 과정을 나타낸 것이다. 입력 신호는 다중 해상도 필터 뱅크를 통과해 6개의 서브밴드 신호로 나누어진다. 이 중 0~5kHz에 해당하는 4개의 서브밴드 신호에서만 정현파 모델링이 행해지고 나머지 밴드의 신호는 잡음 성분으로 취급된다. 이렇게 함으로써 음질에 크게 영향을 주지 않으면서도 계산량을 줄일 수 있다. 정현파 모델링이 행해지는 서브밴드 신호는 동적 세그멘테이션을 통해 구간별 프레임 크기가 결정된다. 프레임 구성이 정해지면 각 프레임 신호에서 matching pursuit을 이용한 스펙트럼 분석과 정현파 성분의 파라메터 추출이 이루어진다. 이것을 위해 각 프레임 신호에 대한 마스킹 임계치가 구해진다. 프레임별로 구해진 파라메터로부터 정현파 모델링의 피크 매칭 알고리즘을 통해 각 정현파 성분에 대한 연속된 파라메터 세트를 얻을 수 있다. 입력 신호에 포함된 잡음 성분을 얻기 위해서는 먼저 각 서브밴드 신호의 정현파 합성 신호를 생성시키고, 이것을 모두 더한 것을 입력 신호에서 빼주면 신호에 포함된 잡음 성분을 얻을 수 있다. 잡음 성분은 스펙트럼 포락선 모델[4]을 통해 파라메터화된다. 분석 결과로 나온 프레임 단위의 정현파 파라메터는 합성 과정에서 시간축 변화율에 따라 샘플 단위의 파라메터로 인터폴레이션되어 정현파 발진기를 통해 합성된다. 잡음 파라메터는 역퓨리에 변환에 필요한 복소수 값으로 변환되고 IFFT를 통해 하나의 합성 프레임 신호가 구해지면 인접 합성 프레임 간 overlap-add를 통해 전체 신호를 합성한다. 이때 시간축 변화 효과는 합성 프레임 크기를 조정함으로써 얻을 수 있다[4]. 마지막으로 정현파 합성 신호와 잡음 합성 신호를 더함으로써 입력 신호의 시간축 변화된 신호를 얻는다.

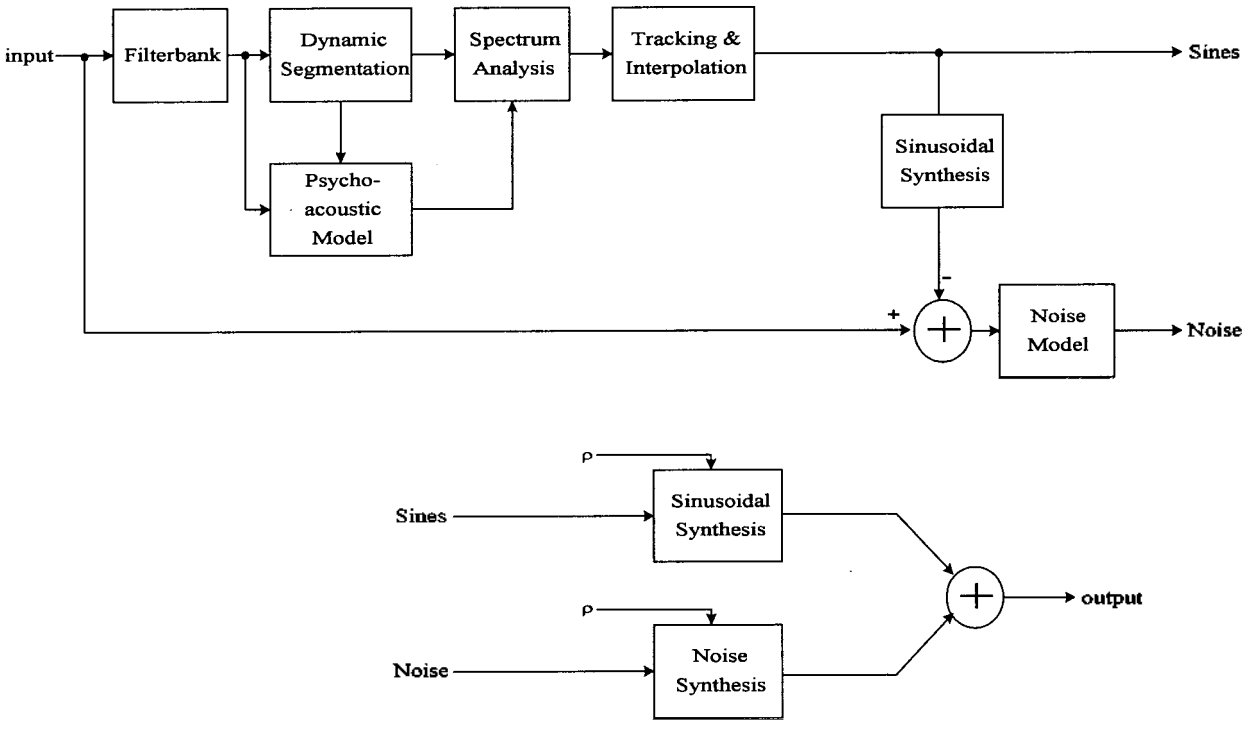


그림 5. 시스템 블록도
Fig. 5. System block diagram.

4.2. 모의 실험 결과

본 논문의 알고리즘을 검증하기 위해 실험에 사용된 오디오 데이터는 표 1과 같이 음성 신호에서 연주 음악에 이르기까지 다양한 종류의 음악 신호를 포함한다. 모든 신호의 샘플링 주파수는 44.1KHz이고 16비트로 표현된다. 정현파 분석/합성 과정에 필요한 몇 가지 중요한 입력 파라메터 값을 표 2에 나타내었다.

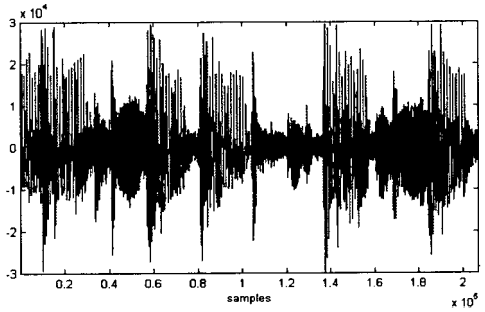
표 1. 실험에 사용된 오디오 데이터 목록
Table 1. List of audio data simulation.

곡 이름	종 류	길 이 (sec)
Orig1	팝 발라드	4.695
Figaro	클래식 성악곡	7.143
Pop	팝 랩음악	3.265
Titasub	영화음악 일부	2.389
Speech1	CNN 뉴스 일부	3.692
Speech2	여성 인사말	3.075

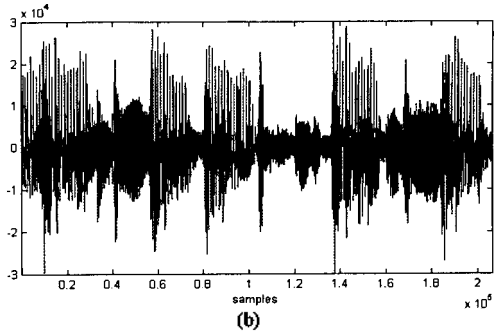
표 2. 정현파 모델링을 위한 중요 분석 파라메터
Table 2. Analysis parameters for sinusoidal modeling.

파라메터	값
정현파 성분 분석 주파수 대역	0~5KHz의 4개의 서브밴드
동적 세그멘테이션에 사용된 프레임 크기	32,64,96,128
주파수 매칭에 사용된 최대 주파수 차이	0.1-0.2
최대 FFT 크기	1024
잡음 성분 분석 크기 및 간격	11.6ms, 5.8ms

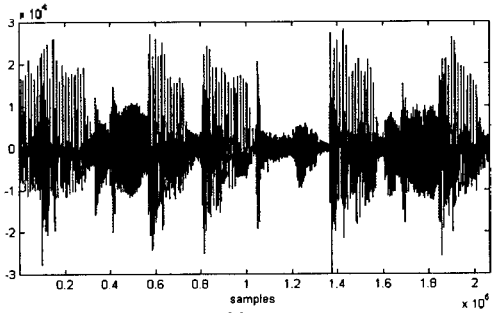
그림 6은 'orig1'음의 정현파 성분의 합성 신호, 잡음 성분의 합성 신호, 그리고 이 두 개의 신호를 더한 원래 신호의 추정 신호를 나타내고 있다. 청취 실험 결과 본 논문에서 제안한 정현파 모델링으로부터 합성된 신호는 원래 신호의 음질을 그대로 복원할 수 있음을 확인할 수 있었다.



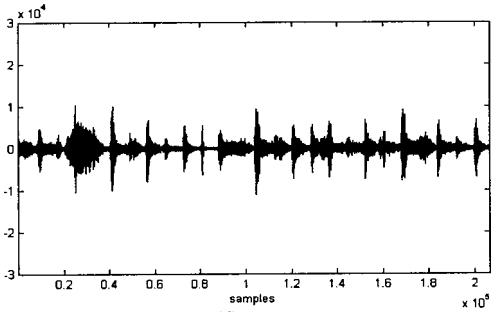
(a)



(b)



(c)



(d)

그림 6. 'orig1'음의 다중 해상도 정현파 모델을 통한 합성 신호
(a) 원래 신호 (b) 정현파 모델 복원 신호
(c) 정현파 성분 합성 신호 (d) 잡음 성분 합성 신호

Fig. 6. Synthesized signal for orig1 using multiresolution sinusoidal modeling,
(a) Original signal, (b) Synthesized signal,
(c) Sinusoidal component synthesis,
(d) Noise component synthesis.

그림 7은 'orig1'음을 각각 0.7배, 1.5배, 2.0배로 시간축 변화한 신호를 보여준다. 그림에서 보면 매우 큰 변화율인 2배 정도의 시간축 확장에도 천이 구간의 파형이 그대로 잘 살아있음을 볼 수 있다.

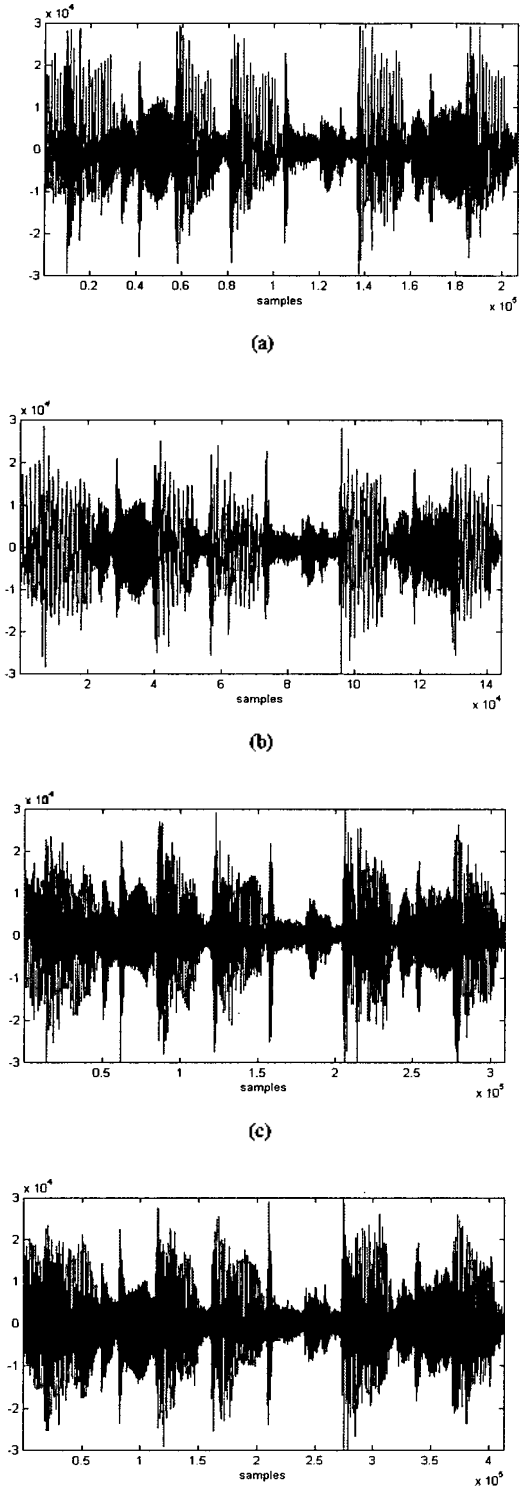


그림 7. orig1 음의 시간축 변화 신호
(a) 원래 신호 (b) 변화율 0.7
(c) 변화율 1.5 (d) 변화율 2.0

Fig. 7. Time-scale modified signals of orig1,
(a) Original signal, (b) Scale factor 0.7,
(c) Scale factor 1.5, (d) Scale factor 2.0.

본 논문에서 제안된 알고리즘이 실제로 얼마나 우수한 음질을 나타내는가를 살펴보기 위해 주관적인 음질 평가를 수행하였다. 음질 평가는 표 1에 나타난 테스트 음으로 부산대학교 전자공학과 대학원생 10명을 상대로 수행하였다. 평가 방법은 하나의 음에 대해 시간축 변화의 시간 영역 방법인 SOLA, 주파수 영역 알고리즘인 phase vocoder, 그리고 본 논문에서 제안된 알고리즘으로 시간축 변화를 수행했을 때 원래 음질과 비교해 가장 음질이 뛰어난 것을 선택하는 것이다. 시간축 변화율은 0.7배, 1.5배, 2.0배로 하였다. 먼저 원래 음을 들려준 다음, 각 변화율에 대해 세가지 알고리즘으로 변화된 음을 임의의 순서로 들려주고, 그 중 가장 나은 것을 선택하게 하였다. 만약 음질을 구별할 수 없으면 구별할 수 없는 음의 번호들을 복수로 기입하게 하였다. 표 3에는 이에 대한 결과를 나타내었다. 여기서 PV로 표시된 것이 phase vocoder, MR로 표시된 것이 본 논문에서 사용한 방법으로 합성한 것이다.

시간 영역의 변화 알고리즘인 SOLA는 음성 신호에 대해서는 우수한 결과를 보인다. 그리고 시간축 축소인 경우에는 폴리포닉 음에 대해서도 비교적 좋은 결과를 얻을 수 있음을 보여준다. 그러나 폴리포닉 음의 시간축 확장이나 음성 신호의 시간축 확장시 변화율이 1.5 이상인 경우에는 음질이 매우 떨어지는 것을 확인하였다

Phase vocoder의 경우는 시간축 축소인 경우에는 좋은 결과를 내지 못하는 반면, 모든 음에 대해 시간축 변화율이 2.0일 때 비교적 좋은 음질을 내는 것으로 나타나 있다. 이것은 phase vocoder가 정수배로 시간축 변화를 일으킬 때는 주파수 성분간의 위상 관계가 유지되므로 우수한 음질을 낼 수 있다는 기존의 연구 결과를 뒷받침해 주는 것이다[12-13].

한편, 본 논문에서 제안된 알고리즘은 음성 신호를 제외한 대부분의 폴리포닉 음에서 시간축 변화시킨 음이 우수한 음질을 내는 것을 확인할 수 있다. 그리고 음성 신호와 같이 주기성이 높은 신호의 경우 시간축 변화된 신호의 음질은 SOLA 방법이 정현파 모델링 방법보다 좀 더 우수한 결과를 나타내는 것을 알 수 있다.

표 3. 시간축 변화 알고리즘의 음질 비교 평가 결과
Table 3. The result of sound quality for different time-scale modification algorithms.

단위: %

변화율	0.7			1.5			2.0		
알고리즘	SOLA	PV	MR	SOLA	PV	MR	SOLA	PV	MR
Orig1	100	20	80	30	0	100	20	50	80
Figaro	90	40	90	20	40	90	20	70	50
Pop	70	30	80	70	30	70	0	80	80
Titasub	90	90	90	50	40	60	40	70	80
Speech1	60	0	80	80	20	20	0	90	20
평균	82	36	84	50	26	68	16	72	62

V. 결 론

본 논문에서는 폴리포닉 음과 같은 복잡한 스펙트럼을 갖는 오디오 신호를 정현파 모델링하고, 고음질의 시간축 변화된 음을 얻는 방법을 제안하였다. 본 논문에서 사용된 방법을 정리하면 다음과 같다.

첫째, 입력 신호를 다중 해상도 필터 뱅크를 통과시켜 서로 대역폭을 가지는 6개의 서브 밴드 신호로 나누고 각 서브 밴드 신호에 대해 정현파 분석을 행하였다. 이렇게 하면 광대역 오디오 신호에 대해 정현파 파라메터를 좀 더 정확하게 분석할 수 있고, 시간-주파수 해상도를 향상시키는 결과를 얻을 수 있다.

둘째, 합성되는 신호의 어택 부분을 최대한 살리기 위해 각 서브 밴드 신호에 대해 동적 세그멘테이션을 적용하여 정현파 합성시 인터폴레이션에 의해 일어나는 과형의 퍼짐 현상을 개선하였다.

셋째, 동적 세그멘테이션에 의해 나누어진 프레임들로부터 정현파 성분의 파라메터들을 추출하기 위해 심리음향 모델에 의한 각 주파수 성분에 대한 마스킹 임계치를 구하고 이를 이용하여 프레임 신호에 포함된 가장 큰 진폭을 가지는 정현파 성분부터 차례로 추출하는 방법을 사용하였다. 다중 해상도 필터 뱅크를 통과한 서브밴드 신호는 주파수 영역에서 확장되어 있으므로 이러한 matching pursuit 알고리즘을 적용함으로서 폴리포닉 오디오 신호에 포함된 다양한 악기음들의 정현파 성분을 효과적으로 추출할 수 있었다.

본 논문의 알고리즘을 이용해 다양한 폴리포닉 음을 실험한 결과 제안한 정현파 모델링 방법이 원해 신호의 음질을 잘 복원할 수 있고, 시간축 변화율의 큰 경우에도 신호의 천이 구간을 잘 표현할 수 있음을 확인하였다. 주관적인 음질 평가를 통해 여러 가지 시간축 알고리즘을 비교한 결과 폴리포닉 음에 대해 본 논문의 알고리즘이 기존의 방법보다 우수한 음질을 내는 것을 확인하였다.

참 고 문 헌

1. R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation", IEEE Trans. on Acoust., Speech and Signal Processing, vol. 34, no 4, pp. 744-754, 1986.
2. T. F. Quatieri and R. J. McAulay, "Speech transformations based on a sinusoidal representation", IEEE Trans. on Acoust., Speech and Signal Processing, vol. 34, no. 6, pp. 1449-1464, 1986.
3. E. B. George and M. J. T. "Smith, Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model", IEEE Trans. on Speech and Audio Processing, vol. 5, no. 5, pp. 389-406, 1997.
4. X. Serra, *A system for sound analysis/transformation/-synthesis based on a deterministic plus stochastic decomposition*, Ph. D. thesis, Stanford University, 1989.
5. K. N. Hamdy, A. H. Tewfik, "T. Chen, and S. Takagi, Time-sacle modification of audio signals with combined harmonic and wavelet representations", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, 1997.
6. S. N. Levine, *Audio representations for data compression and compressed domain processing*, Ph. D. thesis, Stanford University, 1998.
7. M. Goodwin, "Multiresolution sinusoidal modeling using adaptive segmentation", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp. 1525-1528, 1998.
8. 장호근, 박주성, "동적 세그멘테이션을 이용한 폴리포닉 오디오 신호의 정현파 모델링", 한국음향학회지, vol. 19, no. 4, pp. 58-68, 2000.
9. S. G. Mallat and Z. Zhang, "Matching Pursuit with time-frequency dictionaries", IEEE Trans. on Signal Processing, vol. 41, no. 12, pp. 3397-3415, 1993.
10. M. *Goodwin, Adaptive Signal Models*, Kluwer Academic Publishers, 1998.
11. *ISO/IEC 13818-7:Information technology-generic coding of moving pictures and associated audio information-part 7. Advanced Audio Coding, 1997.*
12. M. Puckette, "Phase-locked vocoder", Proc. of IEEE Workshop Appl. of Signal Processing to Audio and Acoustics, 1995.
13. J. Laroche and M. Dolson, "Phase vocoder: About this phasiness business", Proc. of IEEE Workshop Appl. of Signal Processing to Audio and Acoustics, 1997.
14. T. S. Verma and T. H. Y. Meng, "An analysis/synthesis tool for transient signals that allows a flexible sines+transients+noise model for audio", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp. 3573-3576, 1998.
15. 장호근, 권민도, 박주성, "사운드 합성을 위한 DSP의 설계 및 검증", 한국음향학회지, 제 17권, 제 3호, pp. 17-26, 1998.

▲ 장 호 근 (Ho Keun Jang) 1968년 1월 15일생

1993년 2월 : 부산대학교 전자공학과
1995년 2월 : 부산대학교 전자공학과 (공학석사)
2000년 2월 : 부산대학교 전자공학과 (공학박사)
2000년 2월~현재 : 현대전자 시스템 IC 선임 연구원
※ 주관심분야 : 오디오 신호 처리, DSP 설계

▲ 박 주 성 (Ju Sung Park) 1953년 12월 19일생
1976년 2월 : 부산대학교 전자공학과
1978년 2월 : 한국과학기술원 전기 및 전자공학과 (공학석사)
1978년 3월~1985년 7월 : 한국전자기술연구소
1985년 8월~1989년 7월 : University of Florida, Ph.D.
1989년 8월~1991년 3월 : 한국전자통신연구소 책임연구원
1991년 3월~현재 : 부산대학교 전자공학과 부교수
※ 주관심분야 : DSP 설계 및 응용, 사운드 합성, 반도체 소자 모델링