1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 1(s) 호

Sine 파를 이용한 오디오 신호 분석 및 합성

남승현, 홍진우*
배재대학교, 전자공학과
*한국전자통신 연구원, 실감통신 연구실

Analysis and Synthesis of Audio Signals using a Sinusoidal Model

Seung Hyon Nam, Jin-Woo Hong*
Dept. of Electronic Engineering, Paichai University
*Realistic Telecommunications Section, ETRI

요약

Sine 파를 이용한 오디오 분석과 합성은 고음질 저비트율 오디오 부호화에 매우 효율적인 방법의 하나로 알려져 있다. 본 논문은 sine 파를 이용한 오디오 분석과 합성에 중요한 sine 파 검출에 심리음향모델을 활용하는 방안을 제안하였다. 모의실험 결과, 심리음향모델을 사용한 경우 사용하지 않은 경우에 비해 합성에 사용되는 sine 파의 개수를 약 50% 정도 줄일 수 있었음을 알 수 있었다. 한편 오디오 신호의 attack 이나 nonstationarity 를 처리할 수 있는 방법이 sine 파 를 이용한 오디오 부호화에 필수적이라는 사실을 확인하였고 그에 대한 대처 방안을 제시하였다.

1. 개요

지난 10 여년 간 고음질 디지털 오디오 부호화는 사람의 청각특성을 이용한 심리음향모델의 활용에 힘입어 급격한 발전을 이루었다[1]. MPEG 오디오 표준은 이러한 연구의 산물로서 최대 약 12:1 의 압축비를 제공하고 있다. 고음질을 유지하면서 압축비를 높이는 시도는 앞으로 지속될 것이며 많은 진보가 기대된다. 기존의 오디오 부호화기에서 사용하고 있는 스펙트럼의 단순한 양자/부호화 방식은 근본적으로 오디오 신호의 중복성(redundancy)을 효과적으로 제거하지 못하므로 더 높은 압축비를 얻는데 한계를 지닐 수 밖에 없다. 보다 효율적인 부호화를 위해서 새로운 오디오 신호 모델이 필요하며 sine 파 모델은 이러한 모델의 하나로 최근 많은 관심을 끌고 있는 것 중의 하나이다[2]-[6].

Sine 파 모델은 오디오 신호를 주기적인 sine 파 성분들의 합으로 표현하는 방법으로 초기에 음성 신호용으로 제안되었으나, 배경 잡음과 비음성 신호에 강인한 특성으로 인해 오디오 신호의 분석, 합성, 부호화에도 활용되기 시작했다[3]-[6]. Sine 파 모델이 음성과 오디오 신호의 부호화에 효율적이기는 하지만 sine 파 만을 이용하는 경우 오디오 신호를 효과적으로 모델링할 수 없다. 예를 들면, 현악기음에 포함된 마찰음과 같은 잡음 성분을 표현하기 위해서는 비주기적인 잡음 성분이 필요하다. 이와 같이, 오디오 신호를 주기적인 sine 파 성분과 잡음성분으로 분류하거나[4], sine 파 성분과 잡음 성분 그리고 신호의 attack 에서 나타나는 일시적인 (transient) 성분으로 분류하는 방안[6]이 제안되었다.

본 논문에서는 오디오 신호의 성분 중 sine 파 성분의 검출/ 분류/추적/합성하는 문제에 초점을 맞추어 논의하고자 한다. 먼저 sine 파 모델을 이용한 오디오 신호의 분석 및 합성 모델을 설명하고 심리음향모델을 이용한 sine 파의 검출 및 합성 방식을 제안하고 모의실험 결과를 설명한다. 한편 sine 파 분석 및 합성과 관련하여 오디오 신호에 존재하는 attack 과 nonstatioanrity 에 대한 해결 방안을 제시하고자 한다.

2. Sine 파 모델

McAulay 와 Quatieri 에 의해 제안된 sine 파 모델은[2]

$$\tilde{s}[n] = \sum_{l=1}^{L} A_l[n]\cos(\omega_l[n]n + \theta_l) \qquad (1)$$

로 주어진다. 여기서 $A_l[n]$, $\omega_l[n]$, $\theta_l[n]$ 은 각각 l 번째 sine 파 성분의 진폭, 순간 주파수, 그리고 위상이다. 그림 1 은 sine 파 모델을 이용한 일반적인 오디오 신호 분석 및 합성 과정을 보여준다.

분석의 첫번째 과정은 입력 오디오 신호에 윈도우를 씌워 STFT(Short-Time Fourier Transform)을 취한 다음 스펙트럼의 sine 파를 검출하는 것이다. 분석 윈도우 $w_a[n]$ 은 홀수 길이로 정해지며

$$\sum_{n=-N}^{N} w_a[n] = 1 \qquad (2)$$

와 같이 정규화되고, 선형 위상 특성의 영향을 배제하기 위해 윈도우된 데이터가 FFT 버퍼의 중앙에 위치하도록 조정된다[4]. 충분한 주파수 해상도를 얻기 위해서 윈도우의 길이는 피치주기의 약 2.5 배 이상으로 설정된다.

일반적으로 오디오 신호 중의 sine 파 성분은 STFT 스펙트럼의 peak 로 해석된다[2]. 따라서 sine 파 검출은 STFT 스펙트럼의 peak 중 크기가 일정한 문턱값 보다 큰 것을 검출함으로써 이루어진다. 이때 주파수 해상도를 0.1% 정도로 높이기 위해서는 근본적으로 분석 윈도우의 길이를 길게 해야 한다. 그러나 이것은 시간 해상도를 떨어뜨리며 계산량을 급증시키기 때문에 비현실적이다. 따라서 FFT 크기를 적절한 수준으로 유지하며 주파수 해상도를 높이는 방법들이 활용되고 있다[4][5].

검출된 peak 들은 경우에 따라 피치 검출 과정을 거친다. 검출된 피치 정보는 다음에 전개될 peak 추적 과정을 단순하게 하거나 분석 윈도우의 길이를 적응적으로 변화시키는데 사용될 수 있다. 오디오 신호에서의 피치 검출은 음성 신호와는 달리 매우 까다롭다.

검출된 peak 들은 연속성을 점검하기 위해 과거 프레임에서 검출된 peak 들과 비교하는 birth-death 방식의 peak 추적 과정을 거친다[2]. Peak 추적 과정을 거쳐 프레임 간의 연속성이 찾아진 peak 들에 대해 프레임 간의 보간법이 사용되는데 진폭에 대해서는 1 차 보간이 위상에 대해서는 3 차 보간이 실행된다. 보간 후 k 번째 프레임에서의 오디오 신호 합성은

$$\tilde{s}^k[n] = \sum_{l=1}^{p} A_l^k[n]\cos\left(\phi_l^k[n]\right) \qquad (3)$$

과 같이 이루어진다. 여기서 $A_l^k[n]$ 과 $\phi_l^k[n]$ 은 보간 후 얻어진 l 번째 sine 파 성분의 진폭과 위상이다.

일반적으로 위상 정보는 음질에 크게 영향을 미치지 않는 것으로 추정되었지만 위상 정보가 생략될 경우 "reverberant"한 음질이 발생하기 때문에 고음질 합성을 위해서는 위상 정보가 필수적으로 사용되어져야 한다[4]. 또한 오디오 신호를 sine 파 성분과 잡음 등의 잔여 성분으로 분리하여 합성할 경우 sine 파 성분의 위상 일치성이 오디오 신호의 효과적인 분리에 매우 중요하므로 상당한 정확도의 위상 정보가 요구된다.

3. 심리음향모델을 이용한 sine 파 성분 검출 및 합성

Sine 파 성분을 검출하는데 여러 방법들이 제안되어 사용되어왔다. 가장 일반적인 형태는 단순히 일정한 스펙트럼 크기의 문턱값을 초과하는 모든 스펙트럼 peak 들을 검출한 다음 복잡한 peak 추적 방식을 거쳐 의미있는 peak 들을 산출하는 방법이다. 이 방법의 경우 매우 많은 수의 peak 들이 검출되기 때문에 신호의 압축을 목적으로 하는 오디오 신호의 부호화에 그대로 적용하기는 어렵다. 다른 방법으로 least square 최적화를 이용하여 SNR 을 최대로 하는 peak 검출 방법이 제안되었지만[3] 이 방법은 peak 검출 이후 남게 되는 잔여 스펙트럼이 거짓된 peak 로 인식되는 단점이 있다.

본 논문에서 제안하는 방법은 입력 오디오 신호로부터 심리음향모델을 이용하여 매스킹 레벨을 산출하고 이 매스킹 레벨을 넘는 스펙트럼 peak 들을 검출하는 방법이다. 오디오 신호 부호화 과정에서 매스킹 레벨은 어차피 계산되어져야만 하는 값이기 때문에 계산의 추가적인 부담은 없다고 볼 수 있다. 사용 가능한 심리음향모델은 기존의 MPEG 오디오의 모델 1 과 2 가 있으며[7], 모델 2 가 더 정교하고 정확한 매스킹 레벨을 산출하는 것으로 알려져 있으나 어느 방법이나 사용 가능하다.

매스킹 레벨의 산출을 위해서 별도의 STFT 과정을 거치지 않고 peak 검출을 위해 얻어진 STFT 스펙트럼을 사용한다. 그러나 일반적으로 peak 의 주파수 해상도를 높이기 위해 STFT 에서 사용하는 FFT 수는 윈도우 길이 보다 커지게 된다. 보간법을 이용하여 0.1%의 주파수 해상도를 얻는 경우 STFT 에서 요구되는 FFT 수는 윈도우 길이의 약 4 배 이상으로 설정되는데 이 값은 매스킹 레벨 산출을 위해 요구되는 FFT 수 보다 훨씬 큰 값이 된다. 따라서 peak 검출에서 사용하는 STFT 스펙트럼과 심리음향모델에서 사용하는 FFT 스펙트럼 사이의 적절한 조절이 필요하다.

심리음향모델을 이용한 peak 검출 방식은 검출되는 peak 의 수를 대폭 줄임으로써 peak 추적 과정을 단순화시키는 장점도 있다. Peak 의 수가 많아지면 birth-death 매칭 과정을 통한 peak 추적 과정에서 잘못된 매칭을 결론지을 수 있고 이는 위상의 불연속으로 이어져서 경우에 따라 음질이 오히려 저하되는 결과를 나을 수도 있다. 따라서 적은 수의 유효한 sine 파를 얻는 것은 매우 중요한 일이다. 한편 심리음향모델을 사용한 peak 검출의 또 다른 장점은 오디오 신호의 합성 과정에서 시간 영역의 합성 방법을 사용하는 경우 요구되는 많은 계산량을 크게 감소시킬 수 있다는 점이다. 일반적으로 계산량을 줄이기 위해 IFFT 와 overlap-add 를 이용한 주파수 영역에서의 합성을 사용하기도 하지만 이 경우 위상 정보의 손실로 인한 음질의 저하가 두드러진다. 또한 오디오 신호의 파형의 보존이 어려워 오디오 신호를 sine 파와 잡음으로 분리하여 부호화하는데 부적합하다. 따라서 오디오 신호를 sine 파와 잡음 성분으로 구분하여 부호화하고 음질을 높이기 위해서는 시간 영역에서 식 (3)을 이용하여 합성하는 것이 바람직하며, 이 경우 sine 파의 개수의 줄이는 것은 계산량을 감소에 결정적이다.

마지막으로, sine 파를 이용한 오디오 신호 분석과 합성에서 특히 고려해주어야 하는 사항은 오디오 신호에 내재된 attack 부분과 nonstationarity 특성들이다. 이것은 sine 파 모델이 본질적으로 stationary 하다는 점에서 더욱 문제가 된다. 문제 해결을 위해 분석 과정에서 transform coder 에서 일반적으로 사용되고 있는 dynamic window switching 방식[1]의 도입을 고려해 볼 수 있다. 합성 과정에서 합성 프레임은 작을수록 좋은 결과를 얻을 수 있으나 압축 효과가 떨어지는 단점이 있어 근본적인 해결이 되지 못한다.

4. 모의실험 결과

모의 실험을 통해 심리음향모델을 사용한 sine 파 성분 검출의 효과를 조사하였다. 심리음향모델로는 MPEG 오디오에서 사용되는 모델 1 을 활용하였다. 입력 오디오 신호는 44.1 kHz 로 샘플링 된 것이며 분석 윈도우의

길이는 2047 샘플, 합성 프레임의 길이는 1024 샘플로 고정하여 분석 윈도우가 약 50%의 중첩되도록 하였다. 그림 2 는 바이올린 소리에 대해 심리음향모델을 사용하지 않은 경우와 사용한 경우 peak 검출 결과를 보여준다. 여기서 최대 peak 의 수는 200 개로 제한되었다. 심리음향모델을 사용하지 않은 경우 200 개의 peak 를 모두 검출하였으나 심리음향모델을 사용한 경우 평균 50 개의 peak 를 검출하는 것을 확인할 수 있었다. 그러나 일반적으로 peak 매칭 과정에서 매칭이 이루어지지 않는 peak 들은 잘못된 파편으로 간주하여 제거하기 때문에 검출된 peak 모두가 합성에 사용되는 것은 아니다. 본 실험에서는 합성되는 sine 파의 수를 120 개로 제한하였다. 결과 그림 3 에서 볼 수 있는 것처럼 심리음향모델을 사용하지 않은 경우 거의 120 개에 가까운 sine 파가 합성에 사용되었지만 심리음향모델을 사용한 경우 약 30~70 개의 sine 파 만이 사용되었다. 그림 4 는 합성된 오디오 신호의 스펙트럼을 보여준다. 심리음향모델을 사용하지 않은 경우와 사용한 경우 스펙트럼 성분 상의 차이는 심리음향적으로 거의 의미없는 영역의 것임을 알 수 있다.

실제 청음 결과 두 경우의 음질 차이를 거의 느끼지 못하는 것으로 드러났으며 시간 영역에서의 신호 파형도 거의 유사함을 알 수 있었다. 다만 attack 의 처리가 원활하지 못하다는 사실을 확인할 수 있었다.

한편 분석 프레임을 2047 샘플로 고정시킨 상태에서 합성 프레임의 길이를 128, 256, 512, 1024 샘플로 변화시키면서 합성한 신호를 청음한 결과 합성 프레임의 길이에 따라 음질이 크게 좌우되지 않는 것을 알 수 있었다. 이것은 바이올린 음의 피치가 급격하게 변화하지 않기 때문으로 여겨진다. 그러나 그림 5 와 같이 피치가 많이 변화하는 경우 프레임 길이의 변화에 따른 음질의 차이가 두드러지게 나타났다. 높은 압축비를 유지하기 위해 프레임의 길이를 1024 샘플로 고정시키면서 문제를 해결하기 위해서는 피치 변화가 일정 수준을 넘는 경우에는 overlap-add 방식[8]을 활용하여야 할 것이다.

5. 결론

본 논문에서는 sine 파 모델을 이용한 오디오 신호 분석과 합성에 대해 살펴보고 심리음향모델을 활용한 sine 파 검출 방법에 대해 살펴보았다. 모의실험 결과, 심리음향모델을 사용한 경우 사용하지 않은 경우에 비해 합성에 사용되는 sine 파의 개수를 약 50% 정도 줄일 수 있었음을 알 수 있었다. 한편 모의실험을 통해 합성 프레임의 길이가 길어질수록 nonstationarity 나 attack 에 취약하므로 오디오 부호화를 위해서는 합성 프레임을 길게 유지하면서도 비주기적인 신호 성분을 효과적으로 처리할 수 있는 방법이 필수적이라는 사실을 확인할 수 있었다.

참고문헌

[1] K. Brandenburg and M. Bosi, "Overview of MPEG Audio: Current and Future Standards for Low-Bit Rate Audio Coding," *J. Audio Eng. Soc.*, Vol. 45, No. 12, Jan/Feb 1997.

[2] R. J. McAulay and T. F. Quatieri, "Speech analysis-synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-34, No. 4, pp. 744-754, Aug. 1986.

[3] E. B. George and M. J. T. Smith, "Analysis/Synthesis Overlap-Add Sinusoidal Modeling Applied to the Analysis and Synthesis of Musical Tones," *J. of Audio Eng. Soc.*, Vol. 40, No. 6m pp. 497-516, June, 1992.

[4] X. Serra, "Musical Sound Modeling with Sinusoids plus Noise," *Musical Signal Processing*, Swets & Zeitlinger Publisher, 1997.

[5] ISO/IEC JTC1/SC29/WG11 MPEG, CD 14496-3 Subpart 2: Parametric Coding, 1997.

[6] K. Hamdy, M. Ali, and A. Tewfik, "High Qaulity Audio Coding of Audio Signals with a Combined Harmonics and Wavelet Representation," *ICASSP-96*, Atlanta, GA.

[7] ISO/IEC JTC1/SC29/WG11 MPEG, International Standard IS-11172-3, Part 3: Audio, 1992.

[8] Digital Voice Systems, "Inmarsat-M Voice Codec-Version 2," Inmarsat-M specs, Inmarsat, Feb. 1991.

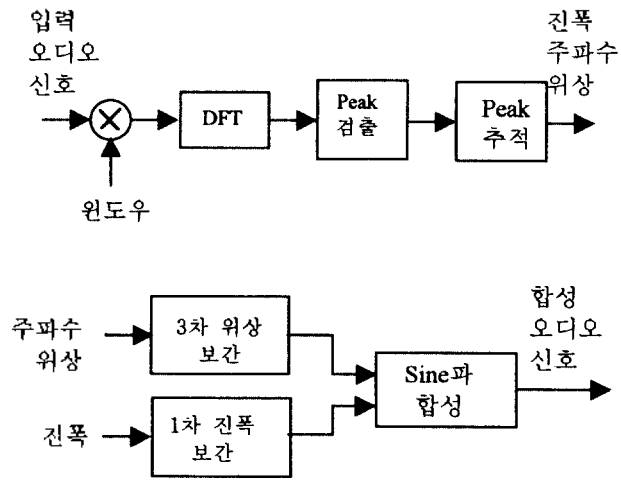

그림 1. Sine 파 모델을 이용한 오디오 신호 분석 및 합성

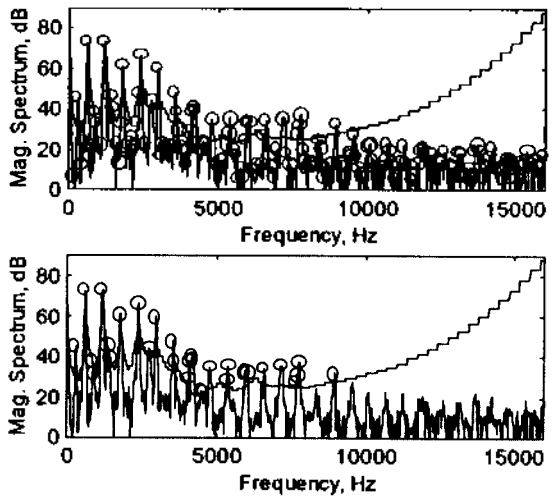

그림 2. Peak 검출: 심리음향모델을 사용하지 않은 경우(위), 사용한 경우(아래)

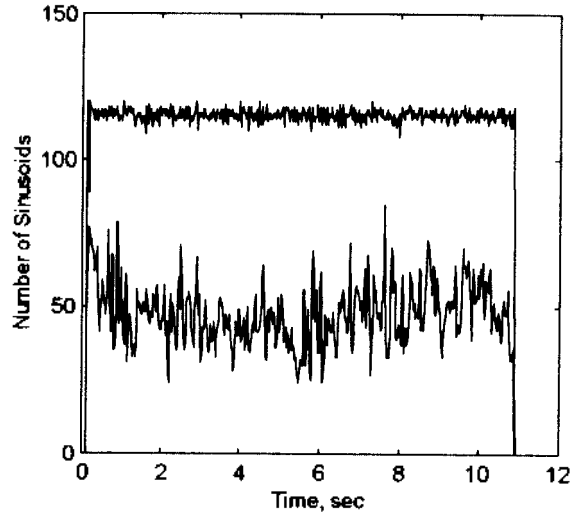

그림 3. 오디오 신호 합성에 사용된 sine 파의 개수: 심리음향모델을 사용하지 않은 경우(위), 사용한 경우(아래).

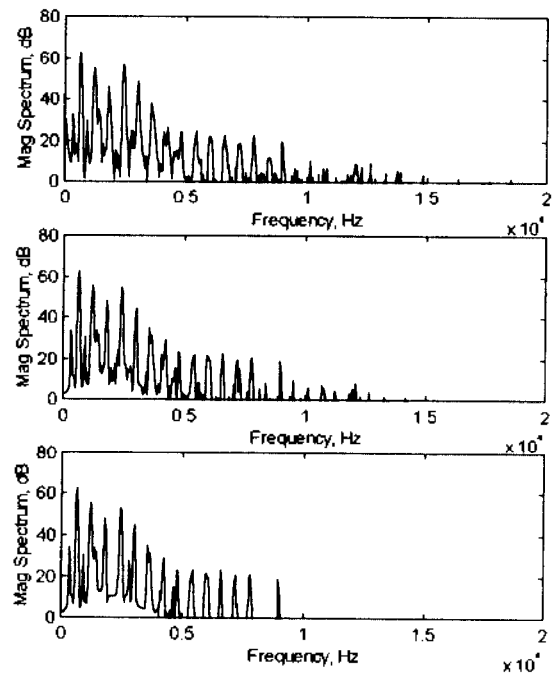

그림 4. 합성 오디오 신호의 스펙트럼 비교: 원음(위), 심리음향모델을 사용하지 않은 경우(중간), 사용한 경우(아래)

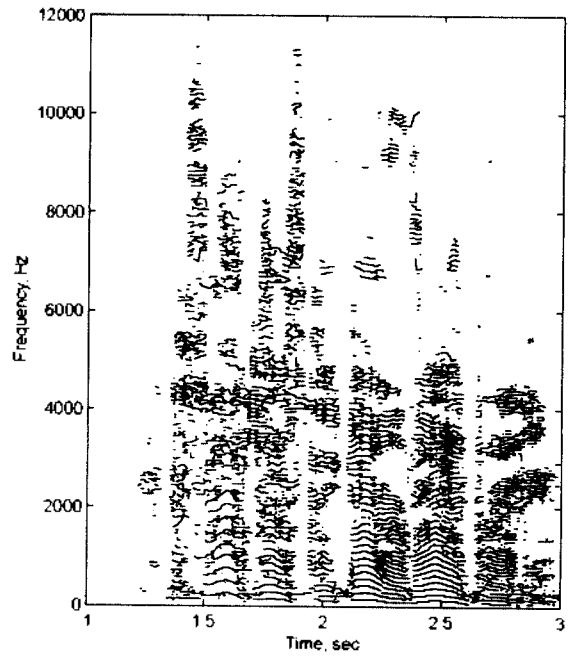

그림 5. 피치가 변하는 오디오 신호의 주파수 추적