1999 年度 大韓電子工學會 夏季綜合學術大會 論文集 第 22 卷 第 1 號 99/6

웨이브렛 패킷을 이용한 고음질 오디오 부호화

안광호, 정전대, 신재호
동국대학교 전자공학과
Tel (02)2260-3336 / Fax 2272-5667

High Quality Audio Coder Using a Wavelet Packet Decomposition

Kwangho Ahn, Cheondae Cheong, Jaeho Shin
Dept. of Electronic Eng., Dongguk Univ.
khahn@cakra.dongguk.ac.kr

Abstract

In this paper we propose high quality audio coding algorithm using psychoacoustic modelling and the adaptive wavelet packet decomposition. The bit allocation scheme exploits the remnants of temporal correlations that exist in the wavelet packet coefficients by SPIHT. The proposed algorithm achieve almost transparent coding of monophonic compact disk(CD) quality signals at about 44kbps.

1. 서 론

인터넷과 디지털 미디어의 증가로 고충실도를 갖는 비디오, 오디오 코딩에 대한 연구가 활발히 진행되고 있다. 오디오에 있어서는 표본화 주파수가 44.1 kHz인 컴팩트 디스크(CD) 신호의 음질이 고음질 디지털 오디오의 표준으로 인정되고 있지만, 채널당 705 kbps의 높은 비트율을 갖는다. 이렇게 높게 요구되는 비트율을 줄이는 것은 오디오 시스템의 디자인에 있어서 중요한 일이다. 전화와 관련된 애플리케이션은 음성 신호의 부호화에 중점을 두지만, 고음질 음악의 저장과 전송을 위한 애플리케이션은 보다 효율적인 압축에 대한 연구의 동기가 되었다.

오디오 압축의 표준으로서 대표적인 것으로 MPEG-1 (ISO/IEC 11172-3) 계층 III가 있으며 128kbps에서 CD 음질을 구현하고 있다. 오디오 압축에서의 핵심은 인간의 청각특성인 노이즈 마스킹 성질을 이용하는 것이다. MPEG-1 오디오에서는 심리음향모델 I,II를 사용하여 irrelevancy를 제거하고 있다. 하지만 MPEG-1 계층 III에서 쓰이는 32개의 동일한 서브밴드를 갖는 polyphase 필터뱅크는 critical band를 정확하게 반영하지 못하기 때문에 효율적이지 못하다. 이를 개선하기 위한 대안으로 웨이브렛 패킷을 이용한 서브밴드 분해에 대한 연구가 이루어지고 있다[1][2]. Srinivasan과 Jamieson은 [1]에서 영상 압축에서 점진적 전송과 효율적인 비트 할당을 위해 Shapiro[5]가 제안한 EZW(Embedded Zerotree Wavelet) 알고리즘을 1차원으로 수정하여 오디오 신호의 압축에 이용하였고, 채널당 45kbps에서 CD와 같은 음질의 압축이 가능하게 하였다.

본 논문에서는 MPEG 계층 III에서 사용하는 심리음향모델 II와 웨이브렛 패킷 분해를 이용한 오디오 신호의 압축을 설명하고, 영상 압축을 위해서 Said[6]가 제안한 SPIHT(Set Partitioning in Hierarchical Trees)를 적용하여 효율적인 비트할당을 통해서 비트율이 개선된 오디오 코딩 알고리즘을 제안한다.

2. 심리음향모델

심리음향모델은 오디오 압축에 있어서 중심적인 역할을 차지한다. 인간 청각 특성을 이용하여 톤의 주위에서 마스크되는 노이즈의 에너지를 구해내고 양자화기는 이를 이용하여 마스크되는 최대한의 노이즈를 허용하여 데이터를 줄이게 된다. 심리음향모델 I은 오디오 스펙트럼의 local peaks를 기반으로 tonal 콤포넌트를 결정한다. 심리음향모델 II는 두 개의 이전 윈도우의 데이터로부터 linear extrapolation을 통해서 현재 윈도우의 값을 예측하고 이것을 tonality의 지수로 삼는다. 본 논문에서는 MPEG-I 계층 III에서 사용하는 심리음향모델 II를 이용하였다[4].

심리음향 모델은 우선 1024-point FFT를 통해서 주파수 영역으로 신호를 표현한다. FFT 계수를 크기-위상으로 표현하고 두 개의 이전 프레임으로부터 현재 프레임의 예측도를 바탕으로 tonality index를 얻어낸다. 주파수 영역에서의 magnitude 값을 critical band 영역으로 바꾸어주고 spreading function과 convolution한다. Spreading function은 톤의 주파수 주위에서 노이즈가 마스크되는 인간의 청각 특성을 표현한다.

Spreading function과 tonality index를 통해서 "just masked" 노이즈 레벨이 계산되고, quiet 상태에서의 threshold와의 비교를 통해서 최종적으로 매스킹 threshold를 구한다. 같은 방법으로 각 서브밴드에서의 매스킹 threshold와 energy를 구한다.

Perceptual 서브밴드 코딩에서 서브밴드에서의 양자화 노이즈를 결정하기 위한 threshold는 서브밴드가 포함하는 주파수 line 중 최소의 threshold가 되어야 한다. 즉, 서브밴드가 critical band와 가까울수록 심리음향 모델이 효율적으로 사용될 수 있는 것이다. 따라서 일반적인 웨이브렛 구조가 아닌 웨이브렛 패킷을 사용하는 것이 바람직하다.

3. Wavelet Packet Decomposition

[1], [3]에서는 오디오 신호의 서브밴드 분해를 위하여 웨이브렛 패킷을 사용하였다. 웨이브렛 패킷 라이브러리의 생성 이론을 간단하게 살펴보면, quadrature mirror filter $h(n)$은 다음 식을 만족한다.

$$\sum_n h(n-2k)h(n-2l)=\delta_{n,l},$$

$$\sum_n h(n)=\sqrt{2}$$

$g_k=(-1)^k h_{1-k}$ 라 하고 $l^2(Z)$에서 $l^2(\in Z)$ 로의 operator F_i를 다음과 같이 정의한다.

$$F_0 s_k(2i) = \sum_k s_k h_{k-2i}$$

$$F_1 s_k(2i) = \sum_k s_k g_{k-2i}$$

$Z=F_0 \oplus F_1$에 의해서 정의되는 사상 $F: l^2(Z) \to l^2(2Z) \oplus l^2(2Z)$는 orthogonal하고

$$F_0F_0^* = F_1F_1^* = I$$

$$F_1F_0^* = F_0F_1^* = 0$$

$$F_0^*F_0 + F_1^*F_1 = I$$

이다.

Scaling 함수 ϕ인 $W_0(x)$와 웨이브렛 함수 ψ인 $W_1(x)$로부터 시작해서 다음과 같은 일련의 함수를 정의하고 이것은 기저를 이룬다.

$$W_{2n}(x) = \sqrt{2}\sum_k h_k W_n(2x-k)$$

$$W_{2n+1}(x) = \sqrt{2}\sum_k g_k W_n(2x-k)$$

다음을 정의한다.

$$m_0(\xi)=\frac{1}{\sqrt{2}}\sum_k e^{-ik\xi},$$

$$m_1(\xi)=-e^{i\xi}\overline{m_0}(\xi+\pi)=\frac{1}{\sqrt{2}}\sum g_k e^{ik\xi}$$

Self similarity relation에 의하여

$$\widehat{W_0}(\xi)=m_0(\xi/2)\widehat{W_0}(\xi/2)$$

또는

$$\widehat{W_0}(\xi)=\prod_{j=1}^{\infty} m_0(\xi/2^j)$$

이고,

$\widehat{W_1}(\xi)=m_1(\xi/2)\widehat{W_0}(\xi/2)=m_1(\xi/2)m_0(\xi/4)m_0(\xi/4)\ldots$

이며 일반적으로

$$\widehat{W_n}(\xi)=\prod_{j=1}^{\infty} m_{\varepsilon_j} 2^{j-1}$$

where

$$n=\sum_{j=1}^{\infty}\varepsilon_j 2^{j-1}$$

이고 $\varepsilon_j=0$ or 1 이다.

웨이브렛 기저 라이브러리는 $W_n(2^l x-k)$ (where $l,k,n\in Z$) 인 함수의 집합으로 이루어진다.

best basis approach에서는 모든 서브밴드를 마지막 레벨까지 분할하지 않는다. 더 이상 분할할 것인지는 애플리케이션의 특성에 따라 적절한 기준에 따라 결정된다. 본 논문에서는 두가지 조건에 따라서 best basis를 결정한다. 먼저 서브밴드를 분할함으로써 서브밴드가 갖는 최소의 threshold가 커지고 따라서 SMR(Signal to Mask ratio)가 작아짐으로써 요구되는 비트율이 작아진다면 서브분할을 수행한다. 예를 들어 서브밴드를 분할 했을 때 하나의 서브밴드가 tone을 가지고 있다면 threshold가 커지고 SMR이 작아지므로 분할을 수행하게 된다.

또 다른 기준은 주어진 computational complexity의 한도내의 계산량을 만족해야 한다. 부호화기와 복호화기에 따라서 허용할 수 있는 계산량에 따라서 이보다 많은 계산량이 필요하다면 대역 분할은 중지하게 된다. 수행되는 계산량은 첫 번째 level에서 c 라고 한다면 다음 서브밴드에서의 분할은 c+(1/2)c 가 된다. 레벨이 증가할수록 data point는 반으로 감소하기 때문이다. 이와 같은 방법으로 계산복잡도를 구해서 제한된 계산량까지의 대역 분할을 수행한다.

서브밴드의 분할은 주파수 해상도는 높이지만 시간 영역에 대한 해상도는 낮아지게 된다. 오디오 신호에

서 갑작스런 변화가 일어나는 구간에서는 pre-echo현상이 일어나므로 시간영역에 대한 해상도도 요구된다. 따라서 서브밴드 분할의 결정은 일반적으로 사용하는 top-down 방식을 이용하여 분할을 결정하기 전에 시간 영역의 해상도를 높인다.

4. SPIHT를 이용한 비트 할당

비트 할당을 위한 방법으로 Said[6]의 SPIHT(Set Partitioning in Hierarchical Tree)를 사용하였다. 이 알고리즘은 영상의 압축률을 높이기 위해서 Shapiro의 EZW(Embedded Zerotree Wavelet)[5]를 보완한 방법이다. 따라서 SPIHT도 계수의 대역간 상관관계를 이용하여 웨이브렛 계수를 부호화하는 embedded 부호화 알고리즘이다.

웨이브렛 패킷 계수들을 coarse scale에서 fine scale로 된 트리구조로 구성한다. SPIHT는 세 개의 리스트, 즉 LIS(List of Insignificant Sets), LIP(List of Insignificant Pixels), LSP(List of Significant Pixels)를 갖는다. Sorting pass에서 LIP의 계수를 지수적으로 감소하는 threshold와 비교하여 significant한 계수를 LSP로 이동시킨다. 마찬가지로 LIS의 set의 중요도를 테스트해서 significant한 set일 경우 LIS에서 제거되고 partition되어 LIS, LIP로 옮겨진다. LSP의 계수들은 refinement pass를 거친다. Threshold는 1/2로 줄어들고 다음 iteration 과정을 수행한다.

iteration 수는 심리음향모델과 서브밴드 분해를 통해 얻은 허용가능한 양자화 에러보다 작을 경우까지 계속된다. 또한 주어진 비트율보다 비트가 많아질 경우도 iteration은 중단된다.

헤더는 우선 필터 뱅크 구조를 트리구조로 인코딩한 값을 저장한다. 트리의 루트에서 시작해서 1은 서브밴드를 분해하는 것을 가리키고 0은 더 이상 분해하지 않는 것을 가리킨다. 다음은 SPIHT에서의 iteration 수를 저장한다. 헤더에 더해진 비트스트림은 lossless 부호화를 거쳐서 전송된다.

5. 전체 알고리즘 및 모의 실험 결과

본 논문에서 제안한 부호화기의 흐름도를 그림 2에서 나타내고 있다. 입력 frame은 웨이브렛 필터뱅크와 심리음향 모델로 들어가고, 심리음향 모델에서는 각 서브밴드에서의 perceptual rate를 계산한다. Threshold calculation partition을 t_i(i = 1... partition 수)라 하고, subband partition을 s_k (k = 1... subband partition 수)라고 하면, subband s_k에 대하여 noise masking threshold를 다음과 같이 구한다.

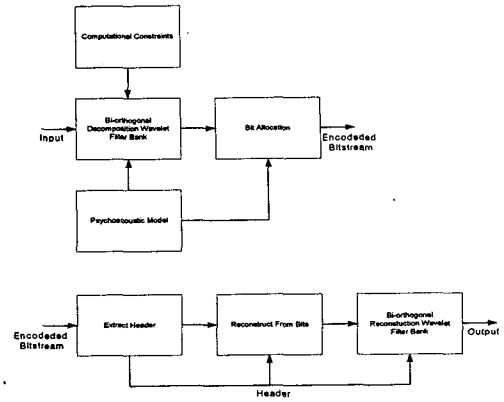

그림 1. 제안 알고리즘 부호화기/복호화기

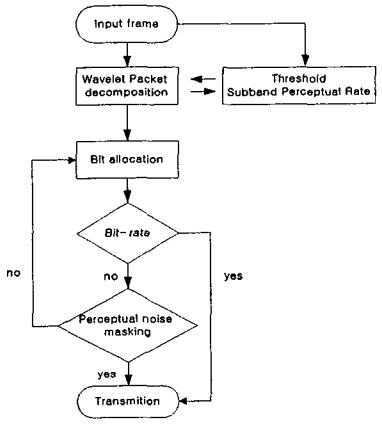

그림 2. 부호화기 흐름도

$$\bigcup_{j=i}^{i''} t_j \supseteq s_k$$

Subband k에서의 energy를 e_k라 하면, mask to signal ratio(SMR)은 t_{s_k}/e_k 로 주어지며, subband에서의 bit-rate $\sum_k b_k$를 최소화 하기 위하여, quantization error ε_{q_k}는 $\varepsilon_{q_k} < t_{s_k}$를 만족할 때까지, 그림 2의 반복을 계속하게 된다.

부호화 및 복호화기에서 사용한 필터 뱅크는 biorthogonal spline 웨이브렛 필터이다. Biorthogonal 필터는 FIR이며 perfect reconstruction을 제공하고 aliasing을 cancel하는 이점이 있다[1].

심리음향학에 의하면 5에서 80dB까지의 SNR (Signal to Noise Ratio) 범위에서 인간 인식에 의해 정확히 같게 들리는 신호와 그렇지 않은 신호를 만들어 낼 수 있다. 때문에 고전적인 방법인 SNR을 통한 perceptual 부호화의 음질 측정은 올바르지 않다.

모의 실험은 10초 가량의 오디오 데이터를 샘플로 사용하여 오디오 비전문가 5명을 대상으로 주관적 음질을 평가하였다. 사용한 데이터는 샘플링 주파수 44.1kHz, 모노 16bit PCM 샘플이며, "비발디의 사계"는 CD에서 데이터를 추출하였고, "cello"와 "quartet"은 EBU(European Broadcasting Union)의 SQAM(Sound Quality Assessment Material) 중에서 선택하였다. 음질 평가 방법은 double-blind test로 [1],[2]에서 사용한 방법을 따랐다. 한쌍의 샘플을 5초의 간격을 두고 들려주고 어떤 것이 음질이 좋은지를 답하는 것이고 "확실치 않음"도 인정하였다. 각 샘플에 대하여 원본 샘플이 더 좋은 음질을 갖는다고 답할 확률을 표 1에 나타내었다. 0.5에 가까울수록 CD와 구별할 수 없는 음질을 갖는다. 표에서 보이는 것과 같이 "quartet"에서 낮은 평가가 이루어졌지만 CD와 거의 같은 음질을 갖는 것으로 나타났다.

표 1 주관적 음질 평가 결과

Music type	원본 샘플을 좋게 평가할 확률
비발디의 사계	0.54
cello	0.52
quartet	0.57

6. 결 론

본 논문에서는 적응 웨이브렛 패킷 분해와 심리음향 모델을 이용한 오디오 부호화를 제안하였다. 보다 포괄적인 테스트는 이루어지지 않았지만, 제안한 부호화는 44kbps의 비트율에서 CD와 transparency한 음질을 가지므로 인터넷을 통한 전송과 저장에 적합하고 부호화기와 복호화기에 계산 복잡도의 제한과 가변 비트율의 전송이 가능하다. 이것은 SPIHT 알고리즘의 계수의 상관관계를 이용한 비트 할당이 있었기에 가능하다.

MPEG-2 AAC 등에는 stereo channel간의 상관관계를 제거하기 위하여 M/S (mid/sum) 부호화기와 intensity stereo 부호화기가 추가되어 있다. 제안한 부호화기에도 이와 유사한 스테레오 부호화 기법이 적용된다면 더 좋은 결과를 낼 수 있으리라 사료된다.

참고문헌

[1] P. Srinivasan and L. H. Jamieson. High quality audio compression using an adaptive wavelet packet decomposition and psychoacoustic modelling. IEEE Trans. on Signal Processing, Vol. 46, 1998.

[2] D. Sinha and A. Tewjik. Low bit rate transparent audio compression using adapted wavelets. IEEE Trans. on Signal Processing, Vol. 41, 1993.

[3] M. V. Wickerhauser. Acoustic signal compression with wave packets. Technical Report of Yale Universit, 1989.

[4] ISO/IEC IS11172-3. Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s - Part 3: Audio. 1992.

[5] J. K. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. IEEE Trans. on Signal Processing, Vol. 41, 1993.

[6] A. Said and W. A. Pearlman. A new fast and efficient image codec based on set partitioning in hierarchical trees. IEEE Trans. on Circuits and Systems for Video Technology, 6, 1996.