개선된 시간축 정보량 감축 기술 기반 오디오 부호화 기술

백승권, 임우택, 이태진
한국전자통신연구원
{skbeack, wtlim, tjlee}@etri.re.kr
Seungkwon Beack, Wootaek Lim, Taejin Lee
Electronics and Telecommunications Research Institute (ETRI)

요 약

본 논문에서는 시간축 정보량을 감축하여 오디오 부호화 효율을 개선하기 위한 기술을 제안한다. 시간축 정보량 감축 방법은 종전의 오디오 코덱에서도 활용되었던 대표적인 기술로 TNS(temporal noise shaping) 기술이 있다. 그러나 TNS 기술은 오디오 신호의 천이구간에서 선별적으로 유효하게 동작하며 그 효율성도 간헐적으로 나타나는데 이는 MDCT(modified discrete cosine transform)에서 예측 과정을 수행하는 구조적인 문제를 갖고 있기 때문이다. 본 논문에서는 종전의 TNS 기술의 취약점을 보완한 ITES(intensive temporal envelope shaping) 기술을 제안하였다. 제안 기술은 TNS 보다 유효한 오디오 시간영역 정보량을 예측하고 감축하였으며, 개선된 음질을 나타냄을 주관적 평가를 수행하여 검증하였다.

1. 서론

오디오 부호화 기술은 사람의 청각 인지 특성을 고려한 양자화 수행 방식을 적용하여 정보량을 감축하고 압축하는 기술로 발전해왔다[1][2]. 사람의 청각 인지 특성은 심리음향 모델로부터 도출하는데 심리음향 모델은 오디오 신호의 주파수 영역에서만 분석이 가능하다. 따라서 사람의 청각 인지 특성을 반영한 모델을 적용하여 양자화를 수행하기 위해서는 먼저 오디오 신호를 주파수 영역으로 변환하여야 한다. 오디오 신호를 주파수 영역으로 변환하기 위하여 기본적으로 사전에 결정해야 할 것은 주파수 분석을 위한 윈도우 형태와 주파수 변환을 위한 프레임 크기이다[2]. 여기서 주파수 변환을 위한 오디오 프레임 크기에 따라 시간/주파수 해상도를 달리하게 되는데, 프레임 크기가 클수록 주파수 영역의 양자화 효율을 높일 수 있으나 시간축 상에서 오디오 신호의 급격한 변화 구간인 천이구간 등에서는 양자화 잡음이 발생하여 음질 저하를 초래한다. 이를 프리에코(pre-echo) 잡음이라고 한다[2][3]. 프리에코 잡음을 처리하기 위해서 TNS(temporal noise shaping) 기술이 소개되었으며[3], TNS 기술은 현재 최신 오디오 표준 코덱인 MPEG-H 3D Audio 까지도 채택되어 활용되는 중요도가 매우 높은 부호화 기술이다[4].

본 논문에서 제안하는 시간축 정보량 감축 기술은 TNS 기술을 개선한 기술로, 전반적인 부호화 효율을 높일 수 있도록 보다 정확한 시간축 포락선(envelope) 정보를 분석하고 추출할 수 있도록 고안된 기술이다. 제안된 기술을 간략하게 ITES (intensive temporal envelope shaping)라 정의한다. 본 논문에서는 제안된 ITES 의 기술의 기본 원리와 특징을 설명하고, 그 효과를 주관적 청취평가를 통해 제시하였다.

2. TNS 기술

TNS 기술은 MPEG-2 AAC(advanced audio coding) 표준 코덱 기술에 처음 적용되었다[5]. MPEG 오디오 부호화 기술은 심리음향 모델을 적용하기 위하여 주파수 단위의 오디오 정보 처리를 요구한다.

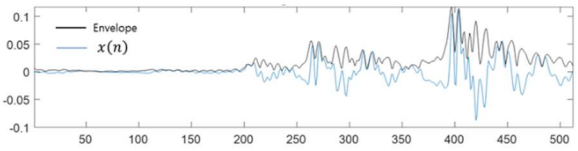

그림 2. 분석신호 $x_a(n)$로부터 추출한 오디오 포락선 정보

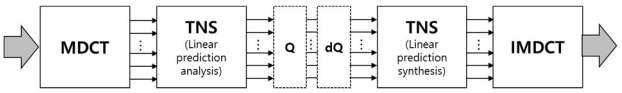

그림 1. TNS 기술을 적용한 MDCT 기반 오디오 부호화기 블록도

오디오 신호의 주파수 정보는 시간/주파수 변환 방식에 따라 다양한 형태로 얻어질 수 있다. AAC 압축 기술은 MDCT(modified discrete cosine transform)를 활용하여 오디오 신호를 주파수 성분으로 변환한다. MDCT 는 시간축 상에서 에일리어싱(aliasing)이 발생하는 불완전한 변환방식이나, 프레임간 오버랩-에드(overlap-add) 연산을 수행하여 완전한 원 오디오 신호로 복원이 가능한 변환 방식이다[1][5]. MDCT 의 활용은 오디오 신호의 주파수 단위의 프레임 데이터 처리시 반드시 수행하는 오버랩-에드 연산으로부터 발생되는 시간 영역 데이터 처리의 증가를 주파수 영역에서 추가적인 데이터 증가없이 양자화 할 수 있는 변환방식이다. TNS 는 MDCT 기반의 오디오 주파수 성분을 대상으로 선형 예측 필터(linear prediction filter)를 적용하여 주파수 영역에서 잔차 신호를 얻고, 이를 양자화 하는 기술이다. 주파수 영역에서의 잔차 신호는 시간축 정보량이 감축된 오디오 신호이다. 주파수 영역에서의 선형 예측 필터에 의해 생성된 주파수 잔차 신호가 시간축 정보량이 감축되는 이유는 주파수 영역에서 선형 예측이 갖는 이중성(duality) 특성 때문이다[3]. 그림 1 은 간략하게 도식화된 TNS 기반 오디오 부호화 과정을 나타낸 것이다. TNS 를 수행하고 얻은 주파수 영역 잔차 신호는 원 신호 대비 시간축 정보량이 감축된 오디오 신호로 TNS 이후에 양자화를 수행하면 실제 시간축 정보량 보다 낮은 정보량에 비트 할당을 수행할 수 있다.

TNS 기술이 시간축 정보량을 감축하는 효과도 있으나, 근본적으로 중요한 특성은 시간축 상의 오디오 신호를 평탄화(flatness) 시키는 효과를 갖는다는 것이다. 오디오 신호의 시간축 상의 평탄화는 오디오 신호가 급격히 변하는 천이구간 등이 발생할 경우 이를 평탄화 하여 천이구간 발생을 감추는 효과를 얻을 수 있다[3]. 따라서 천이구간에서 발생하는 시간축 양자화 잡음을 줄일 수 있고 오디오 신호를 보다 장구간 프레임으로 변환할 수 있으며 주파수 영역에서 상대적으로 많은 데이터를 분석할 수 있어 양자화 효율을 높일 수 있다. 따라서 이상적인 TNS 동작은 부호화 효율도 높임과 동시에, 천이구간에서 발생하는 프리에코 잡음을 억제할 수 있는 두 가지 효과를 기대할 수 있다. 그러나 AAC 등의 기존의 오디오 코덱에서는 TNS 가 프리에코 잡음을 억제하기 위한 수단으로 주로 활용되고 있는데, 이는 과도한 MDCT 영역에서의 선형 예측은 시간축 상에서 포락선이 원본 신호 대비 과도하게 이질적으로 변형되기 때문이다[1][6].

2. ITES 기술

ITES 기술은 기존의 TNS 기술의 성능을 개선하고 부호화 효율도 높이고자 제안된 기술이다. 이상적으로는 TNS 기술이 시간축 정보량을 평탄화 하여 오디오 천이구간에서 양자화 왜곡을 억제하여 부호화 효율을 얻음을 앞서 언급하였다. ITES 기술은 보다 완벽한 시간축 정보량의 평탄화를 수행하기 위한 기술로 TNS 기술의 이론적 근간인 Hilbert envelope 을 보다 충실히 예측하기 위해 제안된 기술이다.

Hilbert envelope 에 대해서는 간략하게 다음과 같이 수식 (1)로부터 설명할 수 있다.

$$x_a(n) = x(n) + jx_h(n) \quad (1)$$

일반적으로, 오디오 신호를 포함한 임의의 실수부만을 갖는 입력 신호 $x(n)$ 은 분석(analytic) 신호인 복소수 $x_a(n)$ 로 표현할 수 있다. 따라서 분석 신호를 얻기 위해서는 $x(n)$ 의 허수부인 $x_h(n)$ 을 구해야 한다. 이때 $x_h(n)$ 를 얻는 방법이 Hilbert transform 이다[3][7]. 분석 신호는 다루고자 하는 신호의 실수부와 허수부를 모두 갖는 복소수 형태의 신호로 그 복소수의 절대값이 원 신호가 나타내는 실제 크기(amplitude) 값에 해당한다. 따라서 오디오 신호 측면에서 분석 신호의 절대값의 형태는 오디오 신호의 시간축 상에서 포락선을 예측하는 정보로 해석할 수 있다.

그림 2 는 임의의 오디오 신호를 수식 1 에 근거하여 분석신호 $x_a(n)$ 로 변환 뒤 분석 신호의 절대값 분석으로부터 포락선 정보를 예측한 결과를 나타낸 것이다. 예측된 시간축 포락선 정보를 활용하여 원 오디오 신호에서 포락선 정보를 추출해 준다면 오디오 시간축 신호를 평탄화 할 수 있다. 시간축 상에서 평탄화 된 오디오 신호는 이상적으로는 천이구간을 감지할 수 없는 잔차 신호로 표현되며, 양자화 왜곡을 줄이기 위하여 천이구간을 따로 감지할 필요가 없어진다. 이는 천이구간 정보가 이미 추출되었다고 판단할 수 있겠다. 그러나 종전의 TNS 기술은 MDCT 를 활용한 예측 파라미터를 추출하기 때문에 에일리어싱 특성에 기인하여 완전한 포락선 정보 추출을 수행할 수 없으며 프리에코 잡음 억제 효율이 떨어진다[7].

본 논문에서 제안하는 ITES 는 보다 충실히 포락선 정보를 시간축 오디오 신호로부터 추출하고자 제안된 기술이다. 그림 3 은 ITES 인코딩 과정을 간략하게 도식화한 블록도이다. 그림에서 분석신호 $x_a(n)$ 에 대해서 주파수 단위 b 번째 입력 프레임 주파수 신호 $x_{a,f}(b)$ 에 대해서 선형예측을 수행한다. 주파수 변환 방식은 복소수 분석이 가능한 DFT(discrete fourier transform)를 적용한다. 따라서 선형예측 계수인 $lpc_a(b)$ 도 복소수 형태이다. 선형예측 계수의 차수는 80msec 기준으로 8 차 복소수 계수를 추출한다. 선형예측 계수를 $x(n)$ 과 동일한 프레임 사이즈(80msec)의 샘플 수를 갖도록 제로 패딩(zero-padding) 이후에 IDFT 를 수행한다. 이렇게 얻은 정보 $x_{env}(n)$ 는 $x(n)$ 의 시간축 포락선 정보이다. 시간축 포락선 정보를 $x(n)$ 에 적용하여 포락선 정보를 추출하면 $x_{res}(n)$ 잔차신호를 얻을 수 있다. 잔차신호 $x_{res}(n)$ 는 종전의 방식대로 양자화를 수행하기 위하여 MDCT 변환 후, 양자화를 수행한다.

그림 4 는 종전의 TNS 방식과 ITES 방식으로 추정한 시간축 포락선 정보를 나타낸 것이다. 종전의 TNS 는 MDCT 영역에서 선형 예측을 수행하여 충실한 포락선 정보를 추정할 수 없으며, 시간축 상에서 발생하는 에일리어싱 신호에 의해서 원 시간영역 신호의 포락선 정보 추정에 오류가 발생한다. 이를 해결하기 위하여 분석 윈도우 형태를 달리하여 TNS 를 적용하는 기술이 현재 AAC 표준에서 활용되고 있다[3][6]. 그러나 이는 완벽하지 않은 TNS 동작에 기인한 것으로 추가적인 구조적 복잡도를 야기시킬 뿐만 아니라 부호화 효율도 떨어트린다. 반면 ITES 는 다양한 형태의 입력 신호에 대해서 일관되게 그 포락선 정보를 추정함을 알 수 있다. 따라서 이를 시간축 오디오 신호로부터 성공적으로 추출할 경우 TNS 와 비교하여 시간영역 오디오 신호가 보다 평탄화 될 가능성이 높아진다. 이는 천이구간에 대한 코딩 효율을 높일뿐만아니라 프리에코 잡음을 좀더 억제할 수 있으며 장구간 프레임 분석이 일관되게 가능하므로 전반적인 부호화 효율을 높일 수 있다.

4. 실험 결과

본 논문에서는 제안된 방법의 성능을 검증하기 위하여 주관적 평가를 수행하였다. 평가 환경은 표 1 에 나타내었다. 주관적 음질 평가 방법은 MUSHRA(MUltiple Stimuli with Hidden Reference and Anchor) 방식을 따라 수행하였다[8]. 본 논문의 제안 방법의 성능 검증을 위하여 본 기술이 효과적으로 동작하는 음성과 음악 오디오 샘플을 선정하여 진행하였다. 테스트 아이템의 표본화 주파수는 12.8 kHz 로 6.4 kHz 의 대역폭을 갖도록 하였다. 이는 본 기술이 주로 오디오 신호의 코어대역 신호를 부호화 할 때 동작하므로 그 효과를 검증하기

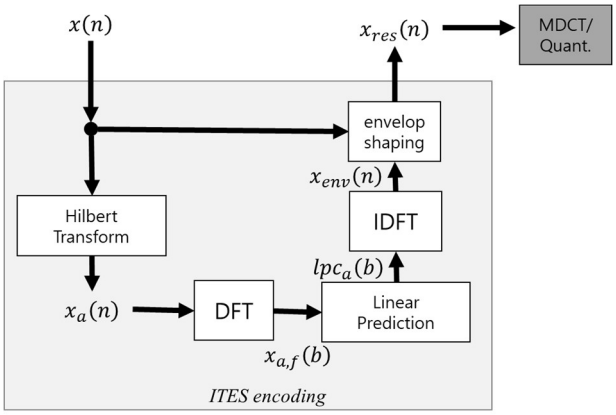

그림 3. ITES 기반 오디오 인코더 구조 블록도

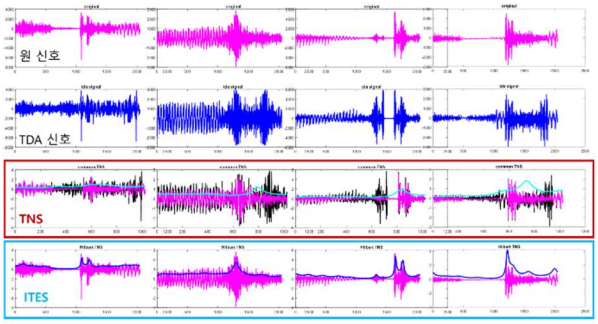

그림 4. TNS 와 ITES 적용시 원 오디오 신호 대비 추출된 포락선 정보

위하여 설정된 대역이다. 참조 시스템은 기존의 TNS 를 적용하고 동일한 양자화 과정을 거치도록 구성한 시스템을 구현하여 평가시스템으로 활용하였다. 제안 방식의 유효성을 검증하기 위하여 코딩 모드는 모두 80msec 의 프레임 사이즈로 고정하여 부/복호화를 수행하였다. 비트레이트는 13kbps 로 부호화 효율 차이를 극명하게 확인하기 위해 양자화 왜곡이 심한 저 비트율에서 음질 평가를 수행하였다. 평가 아이템은 5 개로 음성 아이템 3 개, 음악 아이템 1 개, 음악과 음성이 동시에 포함된 1 개의 아이템으로 구성하였으며, 이는 MPEG 오디오 표준화 단체에서 선정한 테스트 아이템이다[9]. 그림 5 와 6 은 청취 평가를 수행한 결과이다. 그림 5 는 청취 평가를 수행한 절대 점수에 대한 평균치를 기준으로 95% 신뢰구간을 나타낸 결과이며, 신뢰구간이 겹치지 않을 경우, 비교 시스템 간에 통계적으로 의미 있는 차이가 있다고 판단할 수 있다. 절대 점수에 대한 분석결과, 아이템 별로 신뢰 구간이 겹치고 있으나 평균값에서 다소 우위를 나타내고 있음을 관측할 수 있었다. 또한 아이템 별로 보다 분명한 차이를 보기 위하여 절대 평가 점수의 시스템 간의 차이를 통계적으로 분석하여 그림 6 과 같이 나타냈다. 차분 점수의 통계치가 평균값을 기준으로 95% 신뢰구간이 영점에 걸치지 않는다면 두 시스템 간의 차이가 통계적으로 있다고 판단할 수 있다. 제안 시스템을 기준으로 차이

표 1. 주관적 성능측정 평가 환경

평가환경	채택항목
평가 방법	MUSHRA
피험자	6 명
평가 아이템/비트율	5 개(10 초 이상)/13kbps
표본화 주파수	12.8 kHz
평가 시스템	org : Hidden reference
	System A : ITES 기반 부호화기
	System B: TNS 기반 부호화기

점수에 대한 통계치를 그림 6 에 나타내었다. 그림 6 을 살펴보면 제안 시스템이 4 개의 아이템에 대해서 비교 시스템보다 우세함을 관측할 수 있으며, 최종적으로 전체 차이 점수의 평균은 비교시스템보다 통계적으로 우세한 결과를 관측할 수 있다. 다만 하나의 아이템 'te15'에 대해서는 상대적으로 다소 낮은 주관적 평가결과를 얻었으며 이는 천이 구간이 아닌 특정구간에 발생한 양자화 왜곡에 청취자들이 더 민감하게 반응한 것으로 파악된다. 아직까지는 제안 기술을 기존 코덱 시스템에 완벽히 적용하지 않아서 예외적인 처리가 적절히 동작하지 않아 발생한 원인으로 파악된다.

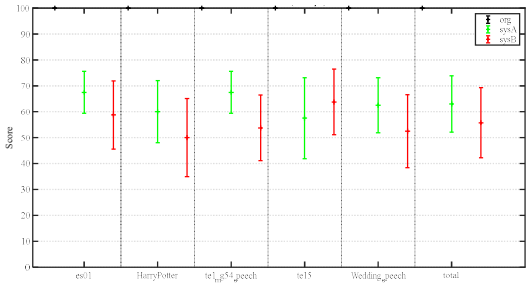

그림 5. 주관적 성능평가 절대점수 평균치 비교

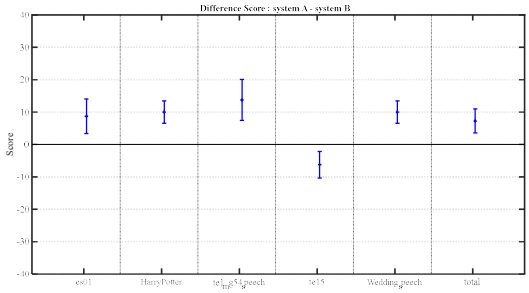

그림 6. 주관적 성능평가 절대점수 차 평균치 비교 (system A 기준)

5. 결론

본 논문에서는 시간축 정보량 감축 방식을 적용한 오디오 부호화 기술인 ITES 기술을 소개하였다. 제안 기술은 종전 기술인 TNS 의 성능을 향상시키고 보다 충실한 정보량 감축을 달성하여 양자화 효율을 개선시켰으며, 그 결과 개선된 음질 성능을 나타냄을 관측할 수 있었다. 특히 종전의 TNS 기술이 간헐적으로 발생하는 특정 천이구간에서만 높은 부호화 효율을 보이는 반면에, 제안 기술은 시간축 상에서 정보량 감축을 전반적으로 수행할 수 있도록 개선하고 천이구간에 일괄된 동작 및 성능을 제공하는 기술로써 신호의 특성에 상관없이 효과적으로 동작함을 확인할 수 있었다. 향후 본 기술의 활용도를 높이고자 기존 오디오 코덱에 실제로 탑재하여 그 성능을 검증하는 작업이 수행되어야 하겠다.

감사의 글

이 논문은 2021 년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2017-0-00072, 초실감 테라미디어를 위한 AV 부호화 및 LF 미디어 원천기술 개발)

참고문헌

[1] T. Painter and A. Spanias, "Perceptual coding of digital audio," Proc. IEEE, vol. 88, no. 4, pp. 451-515, 2000.

[2] M. Bosi and E. Goldberg, Introduction to Digital Audio Coding and Standards. Norwell, MA: Kluwer, 2002

[3] J. Herre and J.D. Johnston, "Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)," in Proc. 101st AES Conv., Nov 1996

[4] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H audio—the new standard for universal spatial/3D audio coding," in Proc. 137th AES Conv., Los Angeles, CA, USA, 2014.

[5] M. Bosi, et al. "ISO/IEC MPEG-2 advanced audio coding." Journal of the Audio engineering society 45.10 (1997): 789-814.

[6] Liu, Chi-Min, Han-Wen Hsu, and Wen-Chieh Lee. "Compression artifacts in perceptual audio coding." IEEE transactions on audio, speech, and language processing 16.4 (2008): 681-695V.

[7] Cizek, "Discrete Hilbert transform," IEEE Trans. Audio Electroacoust., vol. 18, no. 4, pp. 340-343, Dec. 1970. .

[8] International Telecommunication Union, "Method for the subjective assessment of intermediate sound quality (MUSHRA)," 2001, ITU-R, Recommendation BS, 1543-1, Geneva, Switzerland.

ISO/IEC SC29 WG11 N9638, "Evaluation guidelines for unified speech and audio proposals," MPEG, Jan. 2008.