엔트로피 모델을 활용한 심층 신경망 기반 오디오 압축 모델 최적화

*임형섭 *강홍구 **장인선
*연세대학교 **한국전자통신연구원
*hyungseob@dsp.yonsei.ac.kr *hgkang@yonsei.ac.kr **jinsn@etri.re.kr

DNN-based Audio Compression Model Optimization Utilizing Entropy Model

*Lim, Hyungseob *Kang, Hong-Goo **Jang, Inseon
*Yonsei University **Electronics and Telecommunications Research Institute (ETRI)

요약

본 논문에서는 심층 신경망 기반 점진적 다계층 오디오 코덱의 비트 전송률 효율 향상을 위한 엔트로피 모델 기반 양자화 방식을 제안한다. 최근 심층 신경망을 이용하여 전통적인 신호 처리 이론 기반의 상용 오디오 코덱들을 대체하기 위한 오디오 압축 및 복원 시스템에 관한 연구가 활발하게 이루어지고 있다. 그러나 아직은 기존 상용 코덱의 성능에 도달하지 못하고 있으며 특히 종단 간 오디오 압축 모델의 경우, 적은 정보량으로 높은 품질을 얻기 위해서는 부호화기의 양자화 구조를 개선하는 것이 필수적이다. 본 연구에서는 기존에 제안된 종단 간 오디오 압축 모델 중 하나인 점진적 다계층 오디오 코덱의 벡터 양자화기를 엔트로피 모델 기반 양자화기로 대체하고 전송률-왜곡 트레이드오프 관계를 활용하여 전송률을 다양한 형태로 조절할 수 있음을 보임으로써 엔트로피 모델 기반 양자화기 도입의 타당성을 검증한다.

1. 서론

오디오 압축(audio compression) 기술은 오디오 신호(audio signal)를 보다 적은 데이터양으로 전송 혹은 저장하기 위해 사용되는 부호화(encoding), 양자화(quantization), 복호화(decoding) 등의 제반 기술을 포괄하여 지칭하는 용어로, 오디오 데이터를 효율적으로 저장하거나 급증하는 인터넷 기반 멀티미디어 환경에서의 데이터 운송량을 절감하기 위해 일상생활에서 매우 광범위하고 빈번하게 활용되고 있다.

신호 처리(signal processing) 이론에 기반을 둔 전통적인 오디오 압축 기술 중 손실 압축(lossy compression) 방식은 양자화 과정에서 입력 신호를 일부 제거하거나 적은 비트 수를 할당하는 방식으로 데이터양의 감소를 도모하는데, 이는 필연적으로 복호화 이후의 복원 신호에 원본 신호와는 다른 왜곡(distortion)을 초래한다. 양자화 과정에서 발생하는 왜곡이 인지되어 복원 신호의 청취 시 음질의 저하되는 것을 막기 위해 일반적인 오디오 코덱(audio codec)들은 심리 음향 모델(psychoacoustic model)을 활용하여 입력 신호를 구성하는 주파수 성분을 분석한 다음, 일반적인 청취자가 양자화를 통해 발생하는 왜곡을 가능한 인지할 수 없도록 주파수 성분별로 최적의 비트 수를 결정한 뒤 그에 따라 비트를 할당한다. 이러한 인지 기반 압축 방식은 변형 이산 코사인 변환(modified discrete cosine transform; mDCT)과 결합되어 원본 대비 약 10분의 1의 데이터양만으로 원본 신호와 청취 시 거의 구분이 불가능한 복원 신호를 얻어낼 수 있다[1, 2].

기존의 오디오 코덱들이 전통적인 신호 처리 이론에 바탕을 둔 선형 변환(linear transform) 및 양자화 규칙들로 시스템을 구성하였다면, 최근에는 심층 신경망(Deep Neural Network; DNN)을 활용하여 다량의 데이터로부터 최적의 변환 및 양자화 과정을 학습하고자 하는 시도가 다양하게 이루어지고 있다. 종단 간(end-to-end) 오디오 압축 모델은 부호화기와 복호화기 양단을 합성곱 심층 신경망(convolutional neural network; CNN)과 같은 심층 신경망으로 구성하여 신호를 압축하고 복원하는 데 필요한 변환 함수를 최적화하고, 전송해야 할 잠재 영역(latent domain) 특징 변수 양자화기를 미분 가능한 임의의 함수로 대체 및 근사하여 모델을 구성한 후, 최종 복원 신호와 입력 신호 사이의 왜곡을 최소화하도록 전체 신경망 파라미터(parameter)를 최적화한다[3, 4, 5, 6]. 이러한 방식을 통해 원본 대비 약 4분의 1의 데이터양으로도 원본 신호와 거의 구분이 불가능한 복원 신호를 얻어낼 수 있음이 확인되었지만[6], 해당 비트 전송률(bitrate)은 상용화되어 있는 기존의 많은 오디오 코덱들과 견주어보았을 때 상당히 높은 수준이기 때문에, 향후 상용화를 위해서는 전송률을 더욱 줄여야 한다.

본 연구에서는 다양한 종단 간 오디오 압축 모델 중 점진적 다계층 구조를 가지는 모델[6]을 기준으로 기존 양자화 방식의 한계점을 제시하고, 이를 보완하기 위해 엔트로피 모델(entropy model)을 도입한 후, 다양한 실험을 통해 비트 전송률 절감 가능성을 확인한다.

2. 심층 신경망 기반 점진적 다계층 오디오 코덱

[6]은 종단 간 오디오 압축 모델을 통해 고품질 오디오를 복원하기 위해 복수의 코딩 계층을 직렬로 연결하고, 각각의 코딩 계층이 서로 다른 주파수 대역을 중점적으로 복원하도록 설계하였다. 이때, 각 코딩 계층은 이전 코딩 계층에서 발생한 복원 오차까지 함께 부호화하므로 이를

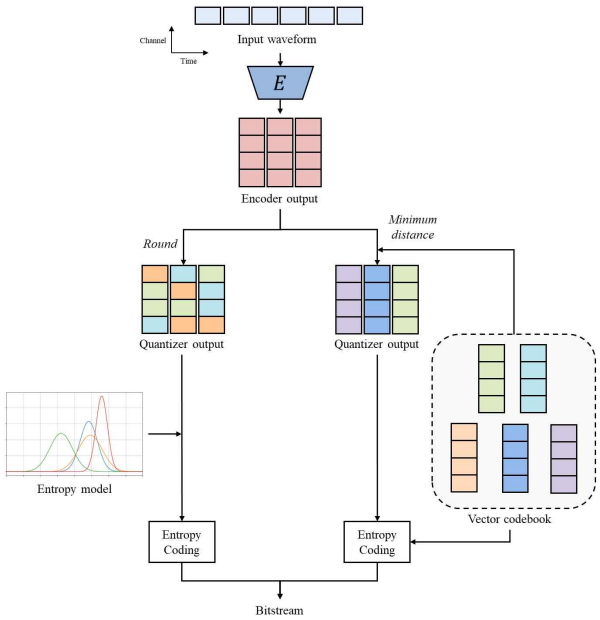

그림 1. 엔트로피 모델 양자화기(왼쪽)와 벡터 양자화기(오른쪽) 구조도

통해 전대역 신호에 대해서도 고품질의 음원을 복원할 수 있다. 본 연구는 상기 모델의 부호화기와 복호화기는 그대로 사용하되, 양자화기의 구성을 변경함으로써 비트 전송률의 절감을 도모한다.

기존 모델의 경우 [7]에서 제안한 벡터 양자화 방식을 준용하였는데, 이는 <그림 1>의 오른쪽 도식과 같이 부호화기를 거쳐 벡터들이 얻어지면 각 벡터와 벡터 코드북(vector codebook)에 포함된 대표 벡터들 사이의 거리를 측정하여 가장 거리가 가까운 대표 벡터로 대체하는 작업이다. 이후에는 비트열 생성을 위해 엔트로피 코딩(entropy coding) 과정을 거쳐 비트를 할당하는데, 이때 할당되는 비트열은 별도로 학습된 각 대표 벡터의 확률 질량 함수(Probability Mass Function; PMF)에 따라 결정된다.

상기한 벡터 양자화는 실험적으로 유의미한 데이터양 절감 효과를 보여주었으나 코드북의 크기가 사전에 결정되어야 하므로 최적화된 모델을 찾는 과정에서 광범위한 실험 및 탐색이 요구된다. 또한, 입력 신호의 특성 변화에 무관한 동일 코드북 및 대표 벡터들을 사용하므로 시간에 따라 매우 큰 폭으로 변하는 오디오 신호의 특징을 모두 반영하기 위해서는 많은 수의 대표 벡터를 준비하여야 한다. 이러한 문제점을 보완하기 위해 입력 신호의 특성 변화에 따라 코딩 방식을 적응적으로 정하는 것이 바람직하나 이러한 코딩 방식을 적용하기 위해 학습 가능한 적응형 코드북을 구현하기는 쉽지 않다. 따라서 본 연구는 기존 벡터 양자화 기반 양자화기의 이러한 한계점을 극복하기 위해 다음 절에서 자세히 설명할 엔트로피 모델을 제안한다.

3. 엔트로피 모델을 활용한 부호율-왜곡 최적화

엔트로피 모델은 임의의 미분 가능한 단일 변수(univariate) 확률 밀도 함수(Probability Density Function; PDF) 혹은 누적 분포 함수(Cumulative Distribution Function; CDF)를 지칭하는 용어로, [8]에서 학습 가능한 양자화기를 구성하는 과정에서 소개되었다. 엔트로피 모델은 미분 가능하므로 심층 신경망의 최적화 과정에서 일반적으로 사용되는 기울기 하강(gradient descent) 알고리즘을 통해 함께 학습될 수 있는 특징을 가진다.

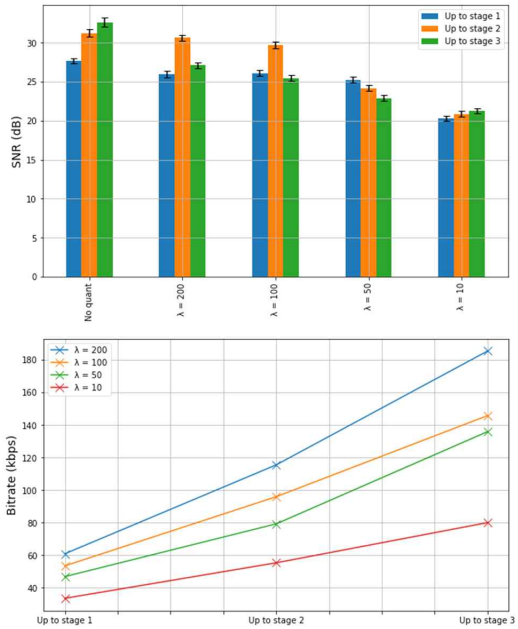

그림 2. λ의 값에 따른 스테이지별 SNR 및 비트 전송률

종단 간 신호 압축 모델은 일반적으로 다음 <수식 1>과 같은 손실 함수 L를 최소화하는 과정에서 달성 가능한 부호율-왜곡(rate-distortion) 균형점 가운데 최적의 값을 탐색한다.

$$L = R + \lambda D \quad (1)$$

<수식 1>에서 R은 신호를 부호화한 뒤의 비트열의 길이(혹은 부호율)를 의미하고, D는 복원 신호와 원본 신호 사이의 왜곡을 의미한다. 이때 각 항의 중요도는 임의의 양의 실수 λ에 의해 결정되는데, λ의 값이 크면 비트 전송률은 높더라도 보다 왜곡을 낮추도록 학습되며, λ의 값이 작으면 왜곡이 더 발생하더라도 비트 전송률을 낮추는 방향으로 학습된다.

심층 신경망으로 구성된 종단 간 신호 압축 모델에서 왜곡 D는 복원 신호와 원본 신호 사이의 평균 제곱 오차(Mean-Squared Error; MSE) 등의 수식을 통해 쉽게 계산될 수 있으나, 양자화 함수는 미분 불가능하므로 그 결과로 얻어지는 비트열의 길이를 직접 최적화하는 것은 불가능하다. 그러나 양자화 이후에 얻어지는 기호(symbol)의 확률 분포를 기반으로 한 엔트로피는 직접 계산 가능하며, 산술 코딩(arithmetic coding)과 같은 엔트로피 코딩(entropy coding) 방식을 통해 엔트로피와 근사한 부호율을 달성하는 것이 가능하다고 알려져 있으므로 종단 간 신호 압축 모델은 비트열의 길이 대신 부호화기 출력의 엔트로피를 최적화한다.

엔트로피 모델, 즉 미분 가능한 확률 밀도 함수(혹은 누적 분포 함수)는 이 과정에서 각 기호의 정보량을 측정하는 데 이용된다. 양자화기 출력 y에 대하여 실제 확률 분포를 P_y, 엔트로피 모델을 Q_y라 할 때, Q_y의 확률을 기반으로 한 엔트로피 코딩 시의 엔트로피는 다음 <수식 2>와 같이 계산된다.

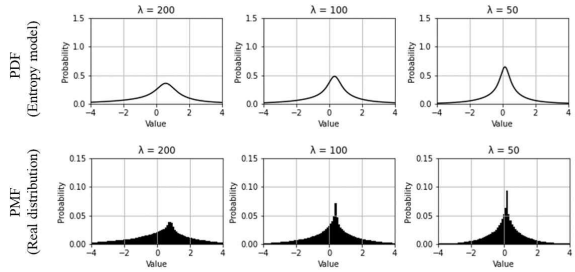

그림 3. 엔트로피 모델과 실제 분포 비교

$$R = E_{y \sim P_y}\left[-\log_2 Q_y(y)\right] \quad (2)$$

한 가지 주의할 점으로, Q_y는 오직 비트열의 형성에만 관여하므로 두 분포가 같지 않더라도 양자화 과정에서 별도의 왜곡이 발생하지는 않고, 따라서 P_y와 Q_y는 반드시 같은 분포일 필요가 없다. 단, 엔트로피 모델 Q_y가 실제 분포 P_y를 제대로 모방하지 못하는 경우 실제로 발생 확률이 적은 기호에 짧은 길이의 코드를 할당하는 등의 비효율성이 발생할 수 있으므로 엔트로피 모델은 실제 분포와 같아지는 것이 바람직하다. 이와 관련하여 엔트로피 모델이 실제 기호의 분포를 제대로 학습하는지와 엔트로피 모델을 통해 추정된 엔트로피가 실제로 학습을 통해 최적화될 수 있는지는 다음 절의 실험 결과를 통해 확인할 수 있다.

부가적으로, 실제 양자화는 임의의 양자화 레벨 Δ을 갖는 mid-rise 균일 양자화기를 통해 이루어지는데, 이 함수는 미분 불가능하므로 실제 학습 과정에서는 균일 확률 분포 $U(-\Delta/2, \Delta/2)$로부터 샘플된 잡음이 가산되어 양자화 과정을 근사화한다. [8]

4. 실험 결과

엔트로피 모델 도입의 효과를 확인하기 위해 점진적 다계층 오디오 코덱[6]의 벡터 양자화기를 엔트로피 모델 기반 양자화기로 대체하고, 다양한 λ값 조건에서 모델을 학습한 뒤 평가 데이터에 대하여 신호 대 잡음비(Signal to Noise Ratio; SNR)와 비트 전송률(bitrate)을 측정하였다.

부호화기와 복호화기에는 [6]의 구조가 그대로 사용되었는데, 부호화기 출력인 8차원 벡터에 대해서 각 채널 및 스테이지별로 독립된 별개의 엔트로피 모델이 할당되었다. 엔트로피 모델의 구현에는 [8]의 univariate non-parametric density model이 사용되어 기존 논문과 같은 방식대로 엔트로피가 계산되었다. 단, 본 논문의 목적은 엔트로피 모델의 도입 효과를 확인하는 것이므로 [6]과는 달리 적대적 생성 신경망(Generative Adversarial Network; GAN)의 적대적 손실 함수를 사용하는 대신 평균 제곱 오차만을 이용해 왜곡을 계산하였다.

모델의 학습에는 약 6시간 분량의 BBC Sound Effects [9] 오디오 샘플이 사용되었으며, 평가에는 USAC(Unified Speech and Audio Coding)[10] test set에 포함된 43개의 오디오 샘플이 사용되었다.

<그림 2>는 4가지 λ값(200, 100, 50, 10) 조건에서 학습된 모델들의 스테이지별 성능을 요약하여 보여준다. 최종 출력인 stage 3를 기준으로 보면 λ의 값이 작아짐에 따라 SNR이 점차 감소하는 한편 보다 적은 비트 전송률이 요구됨을 확인할 수 있다. 실험 결과 내에서 SNR은 λ의 값이 200일 때 27.11 dB로 가장 컸으나 이때의 비트 전송율은 단일 오디오 채널 기준 185.30 kbps로 가장 높았고, λ의 값이 가장 작을 때는 SNR이 21.28 dB로 가장 왜곡이 컸으나 비트 전송률을 80 kbps까지 낮출 수 있었다. 이를 통해 λ의 값을 적절하게 설정함에 따라 왜곡과 부호율 사이의 최적화 지점을 조절할 수 있음을 확인할 수 있었으며, 향후 네트워크 구조 및 학습 방법을 개선함으로써 비트 전송률을 더욱 낮출 수 있을 것으로 기대된다.

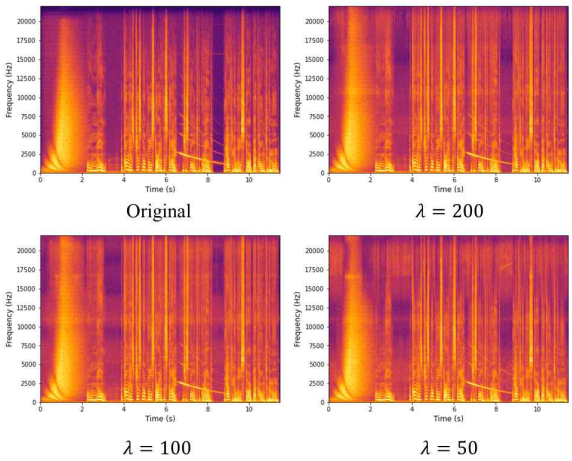

그림 4. 복원 신호의 스펙트로그램 비교

<그림 3>은 특정 스테이지의 특정 채널에 할당된 엔트로피 모델이 학습한 분포와 실제 양자화 후 기호의 히스토그램(histogram)을 비교하여 보여준다. 그림을 통해 엔트로피 모델이 실제 기호의 분포를 잘 학습함을 확인할 수 있으며, λ가 작아짐에 따라 확률 변수의 분포가 첨예해지면서 엔트로피가 감소하는 것 또한 확인할 수 있다.

단, <그림 4>의 스펙트로그램에서 볼 수 있는 것처럼 낮은 비트 전송률에서는 업샘플링(up-sampling)으로 인한 왜곡(11 kHz와 16.5 kHz 인근의 수평선으로 관측되는 aliasing 현상)이 두드러지게 나타났다. 이는 다계층 구조에서 이전 계층의 양자화 오류가 다음 계층으로 전파되면서 나타난 것으로 추정되며, <그림 2>의 SNR 그래프에서 계층이 올라감에 따라 SNR이 낮아지는 현상 역시 이로부터 기인하는 것으로 보인다. 따라서 향후 연구에서는 계층 간 왜곡의 전파를 보완할 방법이 고안되어야 할 것으로 생각된다.

5. 결론

본 논문에서는 심층 신경망 기반 점진적 다계층 오디오 코덱의 비트 전송률의 효율을 향상시키기 위하여 엔트로피 모델 기반의 양자화기 도입을 제안하였고, 실험을 통해 해당 방식의 유용성과 한계점을 확인하였다. 적절한 λ 값의 조절을 통해 비트 전송률을 기존 모델보다 낮출 수 있음을 확인하였으나 양자화 오류로 인한 왜곡이 커짐을 확인하였으며, 향후 연구에서는 이러한 왜곡 및 품질 저하 문제를 해결하고자 한다.

감사의 글

본 연구는 한국전자통신연구원 연구운영비지원사업의 일환으로 수행되었음. [22ZH1200, 초실감 입체공간 미디어 · 콘텐츠 원천기술 연구]

참고문헌

[1] International Organization for Standardization/Internation Electrotechnical Commission (ISO/IEC), “Coding of moving pictures and associated audio for digital storage media at up to about 1.5 mbit/s – Part 3: Audio,” ISO/IEC 11172, 1993.

[2] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs and M. Dietz, “ISO/IEC MPEG-2 advanced audio coding,” Journal of the Audio Engineering Society, vol. 45, no. 10, pp. 789-814, 1997.

[3] S. Kankanahalli, “End-to-end optimized speech coding with deep neural networks,” in Proc. IEEE Internation Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

[4] K. Zhen, J. Sung, M. S. Lee, S. Beack, and M. Kim, “Cascaded cross-module residual learning towards lightweight end-to-end speech coding,” in Proc. Interspeech, 2019.

[5] K. Zhen, M. S. Lee, J. Sung, S. Beack, and M. Kim, “Psychoacoustic calibration of loss functions for efficient end-to-end neural audio coding,” IEEE Signal Process Letters, vol. 27, pp. 2159-2163, 2020.

[6] C. Lee, H. Lim, J. Lee, I. Jang and H. Kang, “Progressive multi-stage neural audio coding with guided references,” in Proc. IEEE Internation Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.

[7] A. v. d. Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning,” in Proc. 30th Annual Conference on Neural Information Processing Systems (NIPS), 2017.

[8] J. Ballé, D. Minnen, S. Singh, S. J. Hwang and N. Johnston, “Variational image compression with a scale hyperprior,” in Proc. International Conference on Learning Representations (ICLR), 2018.

[9] BBC Sound Effects, https://sound-effects.bbcrewind.co.uk/

[10] International Organization for Standardization/Internation Electrotechnical Commission (ISO/IEC), “MPEG audio technologies – Part 3: Unified speech and audio coding,” ISO/IEC 23003, 2012.