심층신경망 기반 오디오 부호화기를 위한 Multi-time Scale 손실함수의 최적화

신승민[1] 변준[1] 박영철[1] 백승권[2] 성종모[2]

[1]연세대학교 지능형신호처리연구실

[2]한국전자통신연구원

nicesin97@yonsei.ac.kr bj9407@yonsei.ac.kr

young00@yonsei.ac.kr skbeack@etri.re.kr jmseong@etri.re.kr

Optimization of Multi-time Scale Loss Function Suitable for DNN-based Audio Coder

Shin, Seung-Min[1], Byun, Joon[1], Park, Young-Cheol[1], Beack, Seung-kwon[2], Sung, Jong-mo[2]

[1]Intelligent Signal Processing LAB, Yonsei University, Wonju, Korea.

[2]Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea

요약

최근, 심층신경망 기반 오디오 부호화기가 활발히 연구되고 있다. 심층신경망 기반 오디오 부호화기는 기존의 전통적인 오디오 부호화기보다 구조적으로 간단하지만, 네트워크의 복잡도를 증가시키지 않고 인지적 성능향상을 기대하는 것은 어렵다. 이 문제를 해결하기 위하여 인간의 청각적 특성을 활용한 심리음향모델 기반 손실함수를 사용한 기법들이 소개되었다. 심리음향 모델 기반 손실함수를 사용한 오디오 부호화기는 양자화 잡음을 잘 제어하였지만, 여전히 지각적인 향상이 필요하다.

본 논문에서는 심층신경망 기반 오디오 부호화기를 위한 Multi-time Scale 손실함수의 지역 손실함수 윈도우 크기의 최적화 제안한다. Multi-time Scale 손실함수의 지역 손실함수 계산을 위한 윈도우 크기를 조절하며, 이를 통하여 오디오 부호화에 적합한 윈도우 사이즈를 결정한다. 실험을 통해 얻은 최적의 Multi-time Scale 손실함수를 사용하여 네트워크를 훈련하였고, 주관적 평가를 통해 기존의 심리음향모델 기반 손실함수보다 좋은 음성 품질을 보여주는 것을 확인하였다.

1. 서론

멀티미디어의 발전이 이루어지면서 대용량 매체에 대한 효율적인저장과 통신을 위한 부호화 기술이 중요하게 대두되고 있다. 이중 최근에는 인공지능의 발전과 더불어 딥러닝 기반 오디오 부호화기 연구도함께 진행되고 있으며, 기존의 전통적인 방법들과 비교하여 데이터 기반의 접근방법이 더 우수한 성능과 잠재력을 보인다. 하지만, 여전히객관적인 지표와 더불어 주관적인 청취 평가에서도 좋은 성능을 내기위한 네트워크 및 훈련 방법들이 활발히 연구되고 있다.

이전 음성 처리 어플리케이션에 대한 연구들에서는 인간 청각 모델을 고려하여 네트워크 훈련 손실함수를 설계하는 방법들을 사용하였다. 음성신호의 경우에는 지각적인 성능 지표인 STOI(short-time objective intelligibility)와 PESQ(perceptual assessment of speech quality)를 차용하여 손실함수를 설계된 바 있다. 오디오 및 음성 부호화에서는 심리음향모델을 활용하여 인간이 인지할 수 있는 청각 커브인 GMT(Global Masking Threshold) 밑으로 양자화 잡음을 은닉하는 방법들 [1-3]이 사용되었다. 결과적으로 rate-distortion 컨트롤이 동반되는 오디오 및 음성 부호화기에서 심리음향모델을 활용하여 음질을 유지하면서 양자화 잡음을 은닉하는 것이 가능하지만, 여전히 투명성이 있는 수준까지 도달하기 위한 추가적인 연구가 필요하다.

본 논문은 심층신경망 기반 오디오 부호화기를 위한 Multi-time Scale 손실함수의 최적화를 제안한다. 한 프레임에 대하여 수행하는전역 심리음향모델 기반 손실함수와 더불어 해당 프레임에 대한 서브프레임의 지역 심리음향모델 기반 손실함수 적용을 통해 전역으로 처리했을 때는 고려되지 못했던 순간적으로 발생하는 음성 구간의 양자화 잡음에 대한 대응이 가능해진다. 이를 위해 지역 손실함수 계산에적합한 윈도우 크기의 결정이 필요하다. 전역 및 지역 심리음향모델 기반 손실함수를 결합한 Multi-time Scale 손실함수의 최적화를 통해,주관적인 평가에서 기존의 심리음향모델 기반 손실함수보다 좋은 청취 품질을 보여주는 것을 확인하였다.

2. 1D CNN 기반 오토인코더

오디오 부호화기를 위한 심층신경망은 1D Convolutional Neural Network (CNN) 기반 오토인코더를 사용하였다. 전체 과정을 나타낸그림은 그림. 1과 같다. 각 인코더와 디코더는 ResGLU (Residual Gated Linear Unit) 블록 기반 네트워크로 구성되어있다. 각 네트워크의 ResGLU 블록은 6개의 1D convolution layer로 구성되어 있으며각 convolution layer의 dilation factor는 1, 2, 4, 8, 16, 32로 구성되어있다. Quantizaer는 이전 연구 [3]에서 사용하던 규일 잔은 모델을 사용하였다.

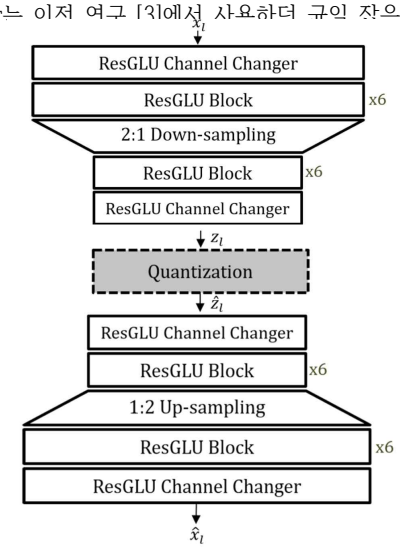

그림 1. 심층신경망 오디오 부호화기의 구조

3. 심층신경망 최적화를 위한 제안된 손실함수 최적화

심층신경망 기반 오디오 부호화기의 종단간 학습을 위하여, Rate-Distortion의 최적화가 필요하다. 심층신경망 모델의 R-D 최적화 문제는 다음과 같이 표현된다.

$$L = L_e + \lambda L_d \qquad (1)$$

L_e는 rate term이다. 즉, 인코더를 통해 만들어진 잠재 벡터로부터 계산된 엔트로피이다. L_d는 distortion term으로 네트워크의 입력 x_l과 출력 $\hat{x}_l$의 차이를 통해 계산한다. L_d는 다음과 같이 계산된다.

$$L_d = L_{mse} + \lambda_p L_p \qquad (2)$$

$L_{걍걸간}$는 mean square error (MSE)이며, $L_{걾}$는 Multi-time Scale Perceptual Loss Function[4]을 의미한다. Multi-time Scale Perceptual Loss Function은 심리음향모델을 사용하여 손실함수 계산에 각기 다른 time-scale 정보를 제공한다.

그림. 1에서의 입력 프레임 크기 T는 다음과 같다. T = 512. 지각적인 손실함수 $L_{걾}$는 한 입력 프레임 사이즈인 512-샘플에 대해 계산된다. 한 개의 입력 프레임을 사용하면 프레임 내부의 지역적인 양자화 잡음을 제어하기 어렵다. 따라서 한 프레임 사이즈 512에 대한 전역 정보와 지역 정보를 모두 사용하기 위해 Multi-time Scale 함수를 사용한다. 전역 정보는 512-샘플, 한 프레임의 크기이며, local 정보는 윈도우 사이즈에 따라 달라진다. 512-샘플로 구성된 하나의 프레임에서64-샘플의 윈도우 사이즈를 50% overlap으로 사용하면 총 15개의subframe을 지역 정보로 사용할 수 있다. 128-샘플 윈도우 사이즈를50% overlap하여 지역 정보로 사용하면 총 7개의 subframe을 지역 정보로 사용할 수 있다. 본 논문에서는 Multi-time Scale 함수 [4]를 오디오 부호화기에 적용하여 최적의 성능을 보이는 윈도우 크기를 확인하였다. 결과적으로 실험에 사용한 손실함수는 다음과 같이 구성된다.

$$L = L_e + \lambda_1 L_{mse} + \lambda_2 L_p^G + \lambda_3 L_p^L \qquad (3)$$

$\lambda_1, \lambda_2, \lambda_3$는 결합 가중치 인자이며 실험적으로 결정하였다.

4. 실험 결과 및 분석

실험을 위한 데이터로는 32kHz로 샘플링 되어있는 1,000개의 상업용 음악 클립을 사용하였다. 700개의 클립은 훈련에 사용하고 200개는 검증, 100개는 테스트용으로 사용하였다. 미니 배치 사이즈는 128이며, 학습률은 0.0002에서 0.0001로 하강하는 cosine annealing을 사용했다. Adam optimizer [5]를 사용했으며, 손실함수 계산에 필요한 가중치 값은 실험적으로 결정하였다. 실험은 Multi-time Scale 손실함수를 사용하지 않은 전역 심리음향 손실함수를 사용한 것과 Multi-time Scale 손실함수를 사용하여 전역과 지역 정보 (64, 128)를 결합하여 훈련한 결과를 비교하였다.

먼저 훈련된 결과에 따른 선호도 실험을 진행하였다. 총 5명의 경험 있는 청취자가 참여하였으며 15개의 테스트 파일을 실험에 사용하였다. 실험 결과에 대한 선호도 실험 결과는 표. 1과 같다.

표 1. 선호도 실험 결과

Case-1	전역	선호 없음	전역+지역(64)
	20%	13.33%	66.67%
Case-2	전역	선호 없음	전역+지역(128)
	40%	13.33%	46.67
Case-3	전역+지역(64)	선호 없음	전역+지역(128)
	40%	33.33%	26.67%

선호도 실험은 56kbps에서 진행하였다. Case-1에서는 전역 정보만을사용한 것보다 전역 정보와 64 윈도우 크기의 지역 정보를 결합하여사용한 것이 3배 이상의 높은 선호도를 기록하였다. Case-2에서는 전역 정보와 128 윈도우 크기를 갖는 지역 정보를 결합하여 비교하였고지역 정보를 결합하여 사용한 것이 6.67% 좋은 성능을 기록하였지만큰 차이를 보이지 않았다. 마지막으로 Multi-time Scale 손실함수끼리의 비교를 진행하였고, 64 윈도우 사이즈를 갖는 손실함수가 가장 높은 선호도 결과를 보임을 확인하였다.

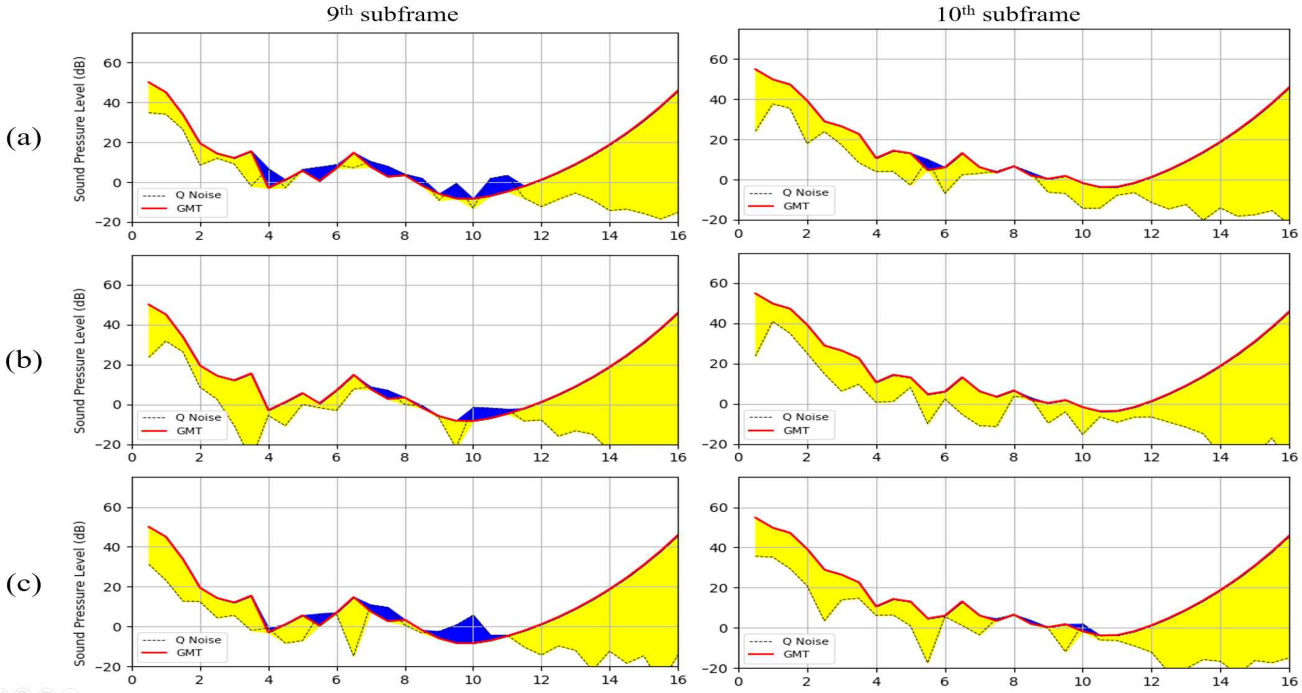

그림 2. 9번째, 10번째 subframe의 양자화 잡음(점선)과 GMT(실선): (a) 전역 (b) 전역+지역(64) (c) 전역+지역(128)

선호도 실험을 논리적으로 입증하기 위하여 양자화 잡음 분포와 GMT를 그려 함께 비교하였다. 512-샘플인 1개 프레임을 가장 작은 지역 윈도우 64 크기로 프레이밍하여 총 15개의 서브프레임을 구한다. 그 중9번째와 10번째의 프레임을 그려 각 손실함수가 프레임 내부에 발생하는 지역적인 변화에 대응하는 능력을 확인하였다. 9번째 서브프레임에서는 (b) 전역+지역(64) 손실함수를 사용했을 때 가장 좋은 결과를 보여주었다. (a) 와 (c) 에 비해 중간 대역 주파수 부분의 양자화 잡음을잘 컨트롤 하는 것을 확인할 수 있다. 양자화 잡음이 크지 않은 10번째프레임과 같은 상황에서도 마찬가지로 (b)가 가장 좋은 성능을 보여주었다.

4. 결론

본 논문은 심층신경망 기반 오디오 부호화기의 최적화를 위한Multi-time Scale 손실함수의 지역 윈도우 사이즈를 제안하였다. 선호도 평가를 통해 전역정보와 64 서브프레임 크기를 갖는 지역정보를 결합하여 사용하는 것이 가장 좋은 성능을 나타내는 것을 확인하였다.

5. 감사의 글

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

참고문헌

[1] Kai Zhen, Mi Suk Lee, Jongmo Sung, Seungkwon Beack, and Minje Kim, “Psychoacoustic calibration of loss functions for efficient end-to-end neural audio coding,” in IEEE Signal Processing Letters, 2020, vol. 27, pp. 2159-2163

[2] Joon Byun, Seungmin Shin, Youngcheol Park, Jongmo Sung, and Seungkwon Beack, “Development of a psychoacoustic loss function for the deep neural network (DNN)-based speech coder,” in Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), 2021, pp. 1694-1698.

[3] Seungmin Shin, Joon Byun, Youngcheol Park, Jongmo Sung, and Seungkwon Beack, “Deep neural network (DNN) audio coder using a perceptually improved training method,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.

[4] Joon Byun, Seungmin Shin, Jongmo Sung, Seungkwon Beack, Youngcheol Park, “Optimization of Deep Neural Network (DNN) Speech Coder Using a Multi Time Scale Perceptual Loss Function”, in Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), 2022.

[5] Diederik P. Kingma and Jimmy Lei Ba, “Adam: A method for stochastic optimization,” in arXiv preprint arXiv:1412.6980, 2014, vol. 19, pp. 2046-2057.