Wave-U-Net을 이용한 오디오 부호화의 성능 향상 기법

*안순호 김재원 박호종
광운대학교
*soonhoan@kakao.com

Audio Coding Enhancement Using Wave-U-Net

*An, Soonho Kim, Jaewon Park, Hochong
Kwangwoon University

요약

본 논문에서는 Wave-U-Net 기반의 오디오 부호화 성능 향상 기법을 제안한다. 기존의 인공지능 기반 오디오 부호화 기술은 오디오의 주파수 정보를 복원하는 방식이기 때문에 완전한 복원을 위해서 주파수의 위상 정보를 별도로 부호화하여 전송해야 한다는 문제점이 있다. 따라서 본 논문에서는 오디오 부호화의 성능 향상을 위해 음원의 주파수 분석을 필요로 하지 않은 end-to-end 모델인 Wave-U-Net을 사용할 것을 제안한다. Wave-U-Net을 사용한 음원이 사용 전의 음원보다 객관적, 주관적 평가 지표에서 우수한 성능을 보이는 것을 확인하였다.

1. 서론

최근 오디오 부호화의 성능 향상을 위해 인공지능 기술을 접목하는 연구가 진행되고 있다[1, 2]. [1, 2]는 오디오의 일부 주파수 정보를 사용하여 이를 학습된 신경망으로 복원하는 방법을 제안한다. 이때 오디오의 주파수 정보는 크기 정보와 위상 정보로 나눠지는데 신경망을 이용한 주파수 복원은 주파수의 크기 정보에 한정된다. 이로 인해 기존 방법은 완전한 복원을 위해서 주파수의 위상 정보를 별도로 부호화하여 전송해야 한다는 문제점이 있다.

본 논문에서는 최근 다양한 음성 및 오디오 신호처리 연구 분야에서 좋은 성능을 보이는 Wave-U-Net을 사용하여 오디오 부호화의 성능을 높이는 방법을 제안한다[3, 4]. Wave-U-Net은 음원 분리, 음성 향상에서 높은 성능을 보인 end-to-end 모델이다.

성능 평가에는 객관적 평가 지표인 log-spectral distortion (LSD)[5]과 주관적 평가 지표인 MUSHRA 청취 평가[6] 결과를 사용하였다. State-of-the-art 오디오 부호화기인 unified speech and audio codec (USAC)로 복원된 출력 신호에 제안한 Wave-U-Net 기반의 성능 향상 방법을 적용하여 오디오 품질 향상을 시도하였고, USAC 복원 신호에 비해 낮은 LSD와 높은 MUSHRA 점수를 가지는 오디오 신호가 출력되는 것을 확인하였다.

2. 제안하는 방법

본 논문에서는 부호화에 의하여 왜곡된 오디오 신호의 품질을 향상시키는 방법으로 Wave-U-Net을 사용할 것을 제안한다. 그림 1은 제안하는 방법의 모델 구조이다.

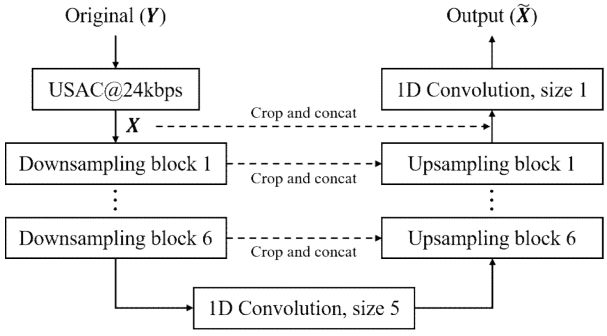

그림 1. 제안하는 방법의 모델 구조
Fig. 1. Model architecture of proposed method

먼저, USAC으로 음원을 손실 압축한 후, 압축된 음원을 Wave-U-Net을 이용하여 복원한다. Wave-U-Net은 6개의 down-sampling (DS) block과 6개의 upsampling (US) block으로 이루어져 있으며 그림 1에서 DS block과 US block 옆의 숫자는 깊이 (depth)를 의미한다.

그림 2는 DS block과 US block의 구조이다.

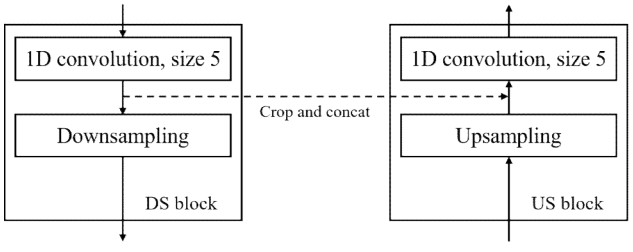

그림 2. Downsampling block과 upsampling block의 구조
Fig. 2. Structure of downsampling block and upsampling block

그림 2에서와 같이 n번째 DS block의 중간 출력은 n번째 US block의 중간 출력의 크기에 맞춰 자르고 합쳐 (concatenate)진다. 마지막으로, DS block 1을 통과하기 전의 음원을 US block 1을 통과한 음원의 크기에 맞춰 자르고 합쳐서 kernel size 1의 1-D convolution 연산을 진행한다. 최종 출력 $\widetilde{X}$와 원본 음원 Y의 mean squared error (MSE) loss를 계산하여 네트워크를 학습한다.

3. 성능 평가

본 논문에서는 제안하는 모델의 학습을 위해 약 57시간 길이의 Beethoven piano sonata, VCTK speech dataset, RWC music dataset을 9 : 1 비율로 나누어 각각 훈련 데이터와 검증 데이터로 사용한다. 제안하는 방법의 성능 평가를 위해 MPEG 오디오 그룹에서 제공하는 약 165초 길이의 12개 음원을 시험 데이터로 사용한다. 시험 데이터는 음성 (speech), 음악 (music), 혼합 (speech over music, SoM) 3가지 카테고리로 나누어지며, 각 카테고리마다 4개의 음원으로 구성한다. 모든 음원은 단일 채널 음원이며 샘플링 주파수는 32 kHz이다.

표 1은 여러 가지 부호화기를 통과한 음원과 원본 음원 사이의 LSD이다. LSD는 원본과의 스펙트럼 데시벨 차이로, 낮을수록 높은 성능을 의미한다[5]. 표 1에서 제안하는 방법의 LSD가 24 kbps의 USAC의 LSD보다 낮고 32 kbps의 USAC의 LSD보다 높은 것을 확인할 수 있다.

표 1. USAC과 제안하는 방법의 LSD 비교
Table 1. LSD of USAC and proposed method

Methods	Speech	Music	SoM	Average
USAC@24kbps	2.23 dB	2.17 dB	2.36 dB	2.25 dB
USAC@32kbps	1.92 dB	1.90 dB	2.12 dB	1.98 dB
Wave-U-Net	**2.10 dB**	**2.13 dB**	**2.26 dB**	**2.16 dB**

그림 3은 주관적 청취평가인 MUSHRA 청취평가 결과이다[6]. 24 kbps의 USAC 음원을 Wave-U-Net으로 향상시켰을 때, 얼마나 더 많은 비트 정보를 사용한 것과 동일한 효과가 나타나는지 알아보기 위해 더 높은 비트율의 USAC 음원들을 포함하여 비교하였다.

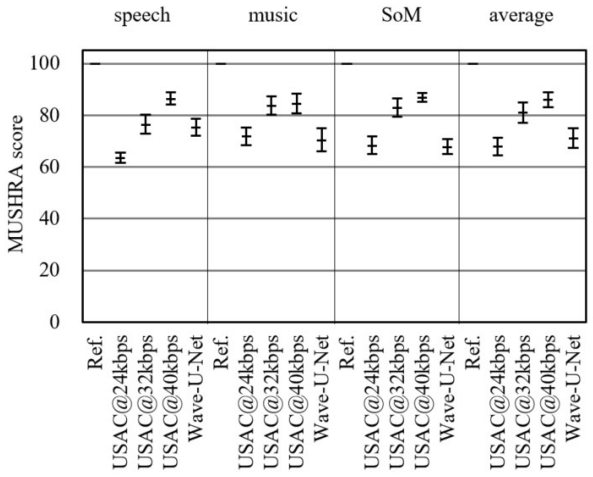

그림 3. 95% 신뢰구간을 갖는 MUSHRA 청취평가 결과
Fig. 3. MUSHRA scores with 95% confidence interval

그림 3을 통해 제안하는 방법이 24 kbps의 USAC과 평균적으로 비슷하거나 더 높은 성능을 보이는 것을 확인할 수 있다. 제안하는 방법은 음악, 혼합 카테고리에서 24 kbps의 USAC과 동일 성능으로 적용 전후의 성능 차이가 없었지만, 음성 카테고리에 대하여 32 kbps의 USAC과 동일 성능을 갖는 것을 확인할 수 있다.

4. 결론

본 논문에서는 Wave-U-Net 기반의 오디오 부호화의 성능 향상 기법을 제안한다. USAC으로 원본 음원을 손실 압축하고 Wave-U-Net으로 복원한다. 제안하는 방법이 동일 비트율 대비 기존의 USAC보다 같거나 높은 성능을 보이는 것을 확인할 수 있으며, 특히 음성 카테고리에 대하여 더 높은 성능을 보이는 것을 확인할 수 있다.

감사의 글

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2017-0-00072).

참고문헌

[1] S.-H. Shin, S.K. Baeck, T. Lee, and H. Park, “Audio coding based on spectral recovery by convolutional neural network," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* pp. 725-729, 2019.

[2] S.-H. Shin, S.K. Baeck, W. Lim, and H. Park, “Enhanced method of audio coding using CNN-based spectral recovery with adaptive structure,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* pp. 351-355, 2020.

[3] D. Stoller, S. Ewert, and S. Dixon, “Wave-u-net: a multi-scale neural network for end-to-end audio source separation,” in *Proc. Int. Soc. Music Inf. Retrieval,* pp. 334-340, 2018.

[4] C. Macartney and T. Weyde, “Improved speech enhancement with the wave-u-net,” *arXiv preprint arXiv:1811.11307,* 2018.

[5] A. Prodeus and I. Kotvytskyi, "On reliability of log-spectral distortion measure in speech quality estimation," in *2017 IEEE 4th International Conference Actual Problems of Unmanned Aerial Vehicles Developments (APUAVD)*, pp. 121-124, 2017.

[6] ITU-R, *Method for the subjective assessment of intermediate quality level of audio systems*, ITU-R BS.1534-3, 2015.