CNN을 이용한 TCX 모드 기반의 주파수 정보 복원 기술

*김재원 신성현 한석현 최현국 김상민 박호종
광운대학교
*meep7174@naver.com

Spectral recovery method based on TCX mode using CNN

*Kim, Jaewon Shin, Seong-Hyeon Han, Seokhyeon Choi, Hyunkook Kim, Sangmin Park, Hochong
Kwangwoon University

요약

본 논문에서는 CNN을 이용한 TCX 모드 기반의 주파수 정보 복원 기술을 제안한다. TCX 모드는 USAC에서 지원하는 음성을 위한 양자화 기술로 부호화 과정에서 포락선을 평탄화한 후 양자화한다. 이러한 평탄화 동작은 주파수 정보 간의 상관도를 높여 네트워크의 학습을 쉽게 만들고 예측 성능을 높인다. 제안하는 방법은 청각 심리 모델 기반으로 구현된 주파수 정보 복원 방법에 TCX 모드 기반의 양자화 방법을 적용하여 일부 주파수 정보만을 사용해 손실된 주파수 정보를 복원한다. 제안하는 방법을 사용해 기존 방법보다 낮은 학습 오차를 얻었고 최적화 되지 않은 조건에서 동등한 음질을 얻었다.

1. 서론

인공지능의 발달로 오디오 부호화 기술에 대한 새로운 연구가 진행되고 있다[1-3]. 인공지능 기반의 오디오 부호화 기술은 손실된 음원 정보를 학습된 네트워크를 사용해 복원함으로써 낮은 비트레이트에서 우수한 음질을 얻을 수 있도록 만들었다.

기존 인공지능 기반의 오디오 부호화 기술은 각 도메인에서 학습된 네트워크를 사용해 손실된 음원 정보를 복원한다. 네트워크는 제공된 음원 정보를 사용하여 손실된 음원 정보를 만들도록 학습되며 제공된 정보와 손실된 정보 간의 상관도가 높을수록 네트워크의 성능은 올라간다. 하지만 기존 방법은 정보 간 상관도를 높여주는 동작 없이 오디오 정보를 그대로 사용하여 음원에 따라 성능 차이를 보인다.

제안하는 방법은 unified speech and audio coding (USAC)에서 지원하는 TCX 모드를 기반으로 양자화된 주파수 정보를 사용하여 손실된 주파수 정보를 복원한다. TCX 모드는 음성을 위한 양자화 기술로 음원의 포락선을 계산하고 음원의 포락선을 평탄화한 후 양자화한다. 이러한 동작은 주파수 정보 간의 상관도를 높이며, 네트워크가 복원해야 할 정보 중 하나인 포락선 정보를 제거하여 네트워크의 성능을 높여준다. TCX 모드를 사용하여 양자화된 주파수 정보를 [3]에서 보고된 주파수 복원 방법에 적용하여 손실된 주파수 정보를 복원한다.

제안하는 방법을 통해 [3]에서 보고된 청각 심리 모델 기반 방법보다 학습 오차가 감소하였으며, 주관적 평가에서 동등한 성능을 얻었다. 주파수 복원 방법의 구조가 청각 심리 모델 기반으로 구현된 것을 고려할 때 TCX 모드의 가능성을 확인할 수 있다.

2. 제안하는 방법

2.1 TCX 모드 기반의 주파수 평탄화 방법

USAC의 TCX 모드는 시간 도메인 (time domain, TD)에서 포락선을 계산하고 주파수 도메인 (frequency domain, FD)으로 표현된 MDCT 계수의 포락선을 제거한다. 포락선 계산을 위해 16차 LPC 계수를 사용하며, DFT를 통해 얻은 64개의 포락선 크기 정보를 사용하여 평탄화한다. TCX 모드의 주파수 평탄화 수식은 식 (1)과 같다.

$$y[k] = \alpha\, x[k] + \beta x[k-1] \tag{1}$$

식 (1)에서 α와 β는 포락선을 통해 얻은 스케일 정보이며, 밴드 단위로 업데이트된다. 그림 1은 TCX 모드로 평탄화시킨 MDCT의 예를 보여준다.

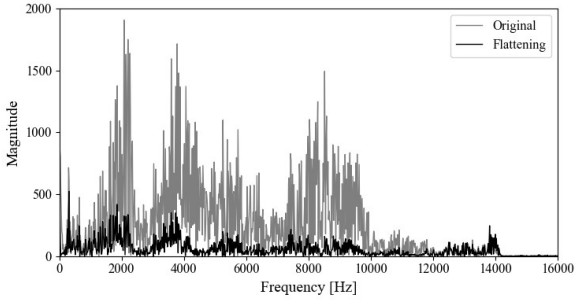

그림 1. TCX 모드 기반의 MDCT 평탄화 예시
Fig. 1. Example of MDCT flattening based on TCX mode.

그림 1을 통해 입력의 포락선이 평탄화된 것을 확인할 수 있다.

2.2 CNN을 이용한 주파수 정보 복원

제안하는 방법은 CNN을 사용하여 손실된 주파수 정보를 복원한다. 주파수 정보 복원은 중대역에 적용하며, 저대역은 USAC에서 지원하는

변환 기반 부호화 기술로 처리한다. 추가로 고대역은 오토인코더(autoencoder, AE)를 사용하여 부호화한다. 각 대역의 정의와 정보 전송 방법은 [3]에서 보고된 방법을 사용한다. 그림 2는 중대역 주파수 정보 복원 방법의 예시를 보여준다.

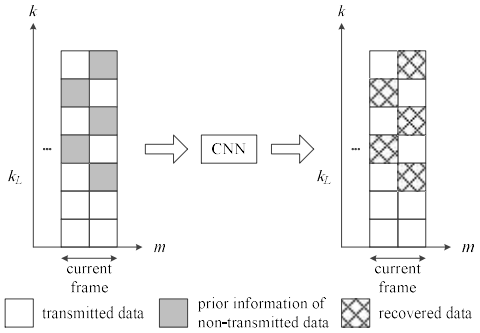

그림 2. CNN을 이용한 주파수 정보 복원 방법
Fig. 2. Spectral recovery using CNN.

그림 2에서 k는 주파수 인덱스이며 m은 서브 프레임 인덱스를 나타낸다. k_L은 저대역 경계이며 4 kHz로 설정한다.

인코더는 TCX 모드로 양자화된 중대역 정보 중 2-D 체크 패턴으로 주파수 정보를 선택하고 선택된 주파수 정보만을 디코더로 전송한다. 추가로 손실된 주파수 정보의 부호 및 크기 정보를 AE에 입력하여 잠재 벡터로 표현하고 벡터 양자화를 통해 전송한다. 이 정보를 사전 정보(prior information)라 정의한다. 디코더는 전송된 주파수 정보와 손실된 주파수 정보의 사전 정보를 복원한 후 주파수 크기 정보를 CNN에 입력하여 손실된 주파수 크기 정보를 복원한다. CNN을 통해 복원된 주파수 크기 정보에 전송된 사전 정보로 복원한 중대역 부호를 적용한다. 이후 포락선을 복원하여 MDCT 계수를 최종 복원한다.

3. 성능 평가

본 논문에서는 학습을 위해 총 57시간의 길이를 갖는 Beethoven piano sonata, VCTK speech dataset와 RWC music dataset을 학습 데이터로 사용한다. 학습 데이터와 확인 데이터는 9 : 1 비율로 나누어진다. 실험 데이터로는 MPEG 오디오 그룹에서 지원하는 총 165초 길이의 12개 음원을 사용한다. 실험 데이터는 음성, 음악, 혼합 데이터로 구성된 3개의 카테고리로 나누어 평가하며 각 카테고리마다 4개의 음원을 사용한다. 모든 음원은 단일 채널 음원이며, 32 kHz 샘플링 레이트를 갖는다. 부호화기로 사용되는 USAC의 비트레이트는 48 kbps를 사용하며 TCX 모드는 TCX1024를 고정으로 사용한다.

그림 3은 제안하는 방법과 기존 방법의 학습 오차 그래프를 보여준다. 보이는 것과 같이 동일한 학습 데이터에서 제안하는 방법이 기존 방법보다 낮은 학습 오차를 갖는 것을 확인할 수 있다.

그림 4는 주관적 청취평가 결과를 보여준다. 청취평가를 위해 4명이 참여하였으며, MUSHRA를 적용하였다[4]. 그림 4를 통해 제안하는 방법과 기존 방법이 모든 카테고리에서 동등한 성능을 갖는 것을 확인할 수 있다. 주파수 복원 방법이 청각 심리 모델을 기반으로 구현되었다는 것을 고려하였을 때 TCX 모드의 발전 가능성을 보여준다.

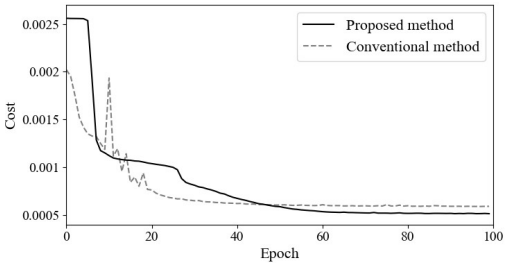

그림 3. 방법에 따른 네트워크 학습 오차
Fig. 3. Training error curve by method.

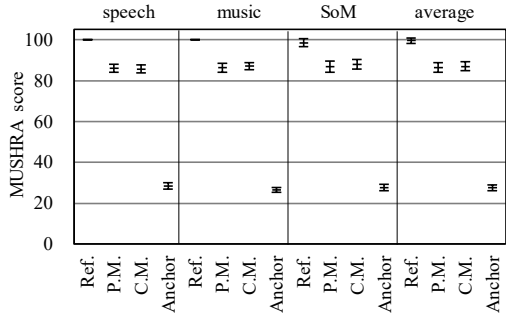

그림 4. 95% 신뢰 구간의 MUSHRA 점수. (P.M.) 제안하는 방법, (C.M.) 기존 방법
Fig. 4. The MUSHRA scores with 95% confidence interval. (P.M.) proposed method, (C.M.) conventional method.

4. 결론

본 논문에서는 CNN을 이용한 TCX 모드 기반의 주파수 정보 복원 기술을 제안한다. TCX 모드를 사용해 주파수 정보의 포락선을 평탄화시켜 네트워크의 학습 성능을 높였다. 제안하는 방법은 청각 심리 모델 기반으로 구현된 기존 주파수 정보 복원 구조에서도 우수한 성능을 얻었다.

감사의 글

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No.2017-0-00072).

참고문헌

[1] W.B. Kleijn, *et al.*, “WaveNet based low rate speech coding,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 676-680, 2018.

[2] 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.

[3] 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.

[4] ITU-R, *Method for the subjective assessment of*

intermediate quality level of audio systems, ITU-R BS.1534-3, 2015.