신경망 기반 오디오 초 해상도 기술 성능 분석

임우택, 백승권, 성종모, 이태진
한국전자통신연구원 미디어부호화연구실
wtlim@etri.re.kr

Performance analysis of audio super-resolution based on neural networks

Wootaek Lim, Seungkwon Beack, Jongmo Sung, Taejin Lee

Media Coding Research Section
Electronics and Telecommunications Research Institute (ETRI)

요 약

오디오 초 해상도 기술은 저 해상도의 오디오 신호를 이용하여 고 해상도의 오디오를 복원 또는 생성해 내는 기술이다. 본 기술 분야는 기존에 주파수 대역 확장, 인공 대역 확장 기술 등으로 연구되었으나, 최근 딥러닝 기술의 발전, 이미지 초 해상도 기술 연구 등에 힘입어 오디오 초 해상도 기술 이라는 이름으로 주로 연구되고 있다. 본 논문에서는 이러한 오디오 초 해상도 기술에 연구 동향에 대하여 설명하고, 기존의 논문 들에서 주로 다루고 있는 음성 데이터 베이스가 아닌 MedleyDB 음악 데이터 베이스를 활용하여 실험을 수행하였다. 실험은 4-폴드 교차 검증을 통해 수행되었으며, 실험 결과 제안하는 컨벌루션 신경망 구조 기반 오디오 초 해상도 기술은 입력 저해상도 오디오 대비 SNR 이 3.41 dB 향상됨을 확인하였다.

1. 서론

최근 딥러닝(Deep learning) 기술의 급격한 발전에 힘입어 오디오 신호처리(Audio signal processing), 음성 인식(Speech recognition), 음성 합성(Speech synthesis), 음악 추천 시스템(Music recommendation system) 등 다양한 오디오 프로세싱 분야에서 기술 발전이 이뤄지고 있다 [1]. 이와 더불어 방송 콘텐츠 시장에서도 고해상도 디스플레이나 고품질 오디오를 위한 재생 장치가 점차 보급되고 있다. 이에 따라 고 해상도의 콘텐츠를 소비하고자 하는 요구가 증가하고 있으며, 이를 위해 낮은 해상도의 콘텐츠를 높은 해상도의 콘텐츠로 변환하고자 하는 연구가 진행되고 있다 [2].

그 중 초 해상도(Super-resolution) 기술은 주로 컴퓨터 비전 분야에서 많은 연구가 이루어지고 있는 분야로, 저해상도 이미지나 동영상을 이용하여 고해상도 이미지 또는 영상을 복원하는 기술이다 [2]. 이미지 초 해상도 기술과 마찬가지로, 오디오 초 해상도 기술은 저 해상도의 오디오 신호를 이용하여 고 해상도의 오디오를 복원 또는 생성해 내는 기술 분야로, 인공 대역 확장(Artificial bandwidth extension), 주파수 대역 확장(Bandwidth extension) 등의 기술 분야로 연구되기도 한다. 이와 관련하여 블라인드(Blind) 방식이 아닌 추가 정보를 활용하여 주파수 대역 확장을 수행하는 스펙트럼 대역 복제(SBR: Spectral Band Replication) 연구가 있었다 [3]. 이 방법은 고주파 대역 신호를 실제 신호를 양자화 하지 않고, 저 대역 신호와 인코더에서 추출되어 전송 된 고주파 신호에 대한 파라미터를 이용하여 고대역 신호를 합성하는 코딩 기술이다. 또한 기계학습 기술의 발전과 함께 블라인드 방식의 대역 확장 연구도 많이 이루어 졌다. 특히 최근에는 딥러닝 기술의 발전과 함께 신경망을 이용한 블라인드 방식의 오디오 초 해상도 기술 연구가 많이 진행되고 있으며, 대표적으로 병목 구조 (Bottleneck architecture) 및 잔차연결(Residual connection) 을 특징으로 하는 오토 인코더(Auto-encoder) 기반의 오디오 초 해상도 방법이 제안되었다 [1]. 그러나 기존의 오디오 초 해상도 연구들은 대부분 음성 데이터에 대한 실험이 주를 이루며, 음악이나 오디오 데이터에 대한 실험을 수행한 연구 결과는 많지

않다. 또한 음성 데이터를 주로 다루고 있기 때문에 낮은 주파수 대역 신호만을 이용하여 초 해상도 기술을 적용하는 등 제한적인 범위 내에서 실험이 이루어진 경우가 대부분이다. 따라서 본 논문에서는 음악 데이터 베이스인 MedleyDB 를 활용하여 제안하는 신경망을 학습하고, 4-폴드 교차 검증(4-fold cross validation)을 통해 실험을 수행하여 신경망 기반 오디오 초 해상도 기술의 성능을 분석한다.

2. 제안하는 방법

본 논문에서는 신경망 기반의 주파수 대역 확장 실험을 위한 오디오 초 해상도 기술을 제안한다. 그림 1 은 제안하는 신경망 모델의 구조이다.

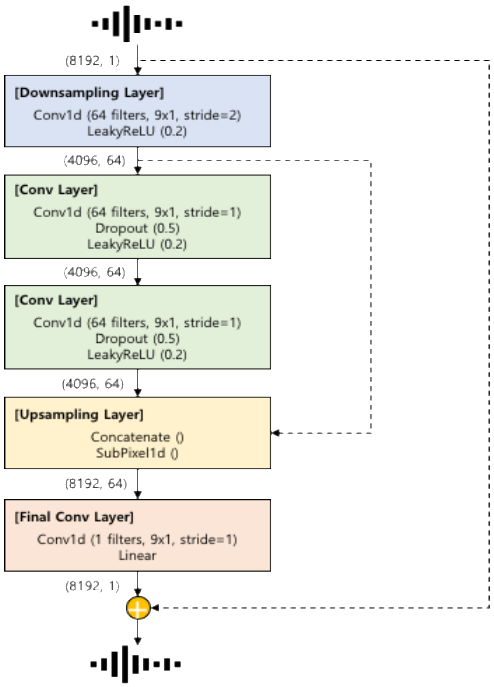

그림 1. 제안하는 신경망 구조
Fig. 1. Proposed neural network structure

제안 모델은 총 5 개의 레이어로 구성되어 있으며 각 레이어는 컨벌루션(Convolution), 활성화(Activation), 드롭아웃(Dropout), 병합(Concatenate), 서브픽셀(SubPixel) 함수로 구성된다. 특징 점은 다운샘플링(Downsampling)내의 컨벌루션 필터 적용 폭(Stride)이 2 인 것과, 업샘플링(Upsampling) 내의 병합 레이어를 통해 채널 수를 2 배로 만들어준 후에, 서브픽셀 레이어를 통해 다시 업샘플링을 수행한다는 점이다. 입력 신호로는 한 프레임당 8192 개 샘플의 오디오가 사용되었으며, 50% 오버랩을 통해 배치를 구성하였다. 학습을 위한 손실 함수로는 식 1 과 같은 평균 제곱 오차(MSE: Mean Squared Error)가 사용되었다.

$$Loss(x,y) = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\|y_i - f(x_i)\|_2^2} \quad (1)$$

3. 성능 평가

본 절에서는 2 절에서 서술한 제안하는 시스템의 성능을 분석하기 위해 공개된 음악 데이터 베이스인 MedleyDB 를 사용하여 실험을 진행하였다. MedleyDB 는 음악정보검색(MIR: Music Information Retrieval) 분야에서 주로 사용되는 음악 데이터 베이스로, 멀티채널 녹음 오디오(Multitrack recordings)와 함께 다양한 메타데이터(Metadata) 및 주석(Annotations) 정보가 포함된 데이터 베이스이다 [4]. 그 중 본 실험에서는 모노 채널로 믹싱 된 오디오를 사용하였다. 제공되는 오디오 클립은 총 122 개이며, 콘텐츠 별 길이는 13 초부터 약 17 분까지 다양하다. 원본 오디오의 샘플링율은 44.1kHz 로 주어지기 때문에 출력 오디오(y)는 원본 오디오를, 입력 오디오(x)는 2 배 다운샘플링을 적용한 뒤 사용했다.

성능 평가는 4-폴드 교차 검증 방법을 사용하였다. 따라서 총 122 개 오디오 클립 중에서 각 폴드당 91-92 개 클립이 학습 데이터(Training data)로 사용되었으며, 나머지 30-31 개 클립은 평가 데이터(Test data)로 사용되었다. 각 폴드당 학습은 100 epochs 동안 훈련되었다. 그림 2 는 오디오 초 해상도 실험 결과의 예제 스펙트로그램(Spectrogram)을 보여준다. 그림 2(a)는 입력 저 해상도(Low-resolution) 오디오, 2(b)는 출력 초 해상도(Super-resolution) 오디오, 마지막 2(c)는 원본 고 해상도(High-resolution) 오디오의 스펙트로그램이다.

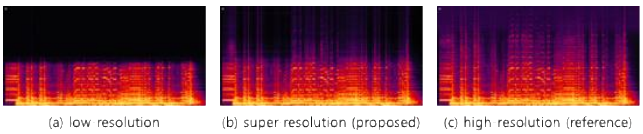

그림 2. 오디오 초 해상도 실험 결과 스펙트로그램 예시
Fig. 2. Example spectrogram of the audio super-resolution result.

그림에서 볼 수 있듯이 입력 저 해상도 오디오로부터 고주파 대역 신호가 생성되어 초 해상도 신호가 출력 된 것을 확인할 수 있으나, 원본 고 해상도 오디오와는 차이를 보여 기술적 한계가 있음을 확인할 수 있었다. 또한 실험 결과의 성능 측정은 식 2 와 같은 SNR 지표를 사용하여 측정하였다.

$$SNR(x,y) = 10log\frac{\|y\|_2^2}{\|x-y\|_2^2} \quad (2)$$

최종 실험 성능 결과는 표 1 과 같으며, 제안하는 오디오 초 해상도 기술을 적용한 결과는 32.48dB 의 SNR 을 보였다. 이는 입력 저 해상도의 오디오 대비 평균 3.41dB 향상된 결과이며, 실험에서 사용된 122 개 오디오 클립에 대해서 콘텐츠 별 편차는 존재하지만 전반적으로 모두 향상된 결과를 보였다.

표 1. 제안하는 오디오 초 해상도 방법 성능 평가 결과
Table 1. The performance evaluation of proposed audio super-resolution method

	SNR
Low-resolution	29.07 dB
Super-resolution (proposed)	32.48 dB

4. 결론

본 논문에서는 음악 데이터 베이스인 MedleyDB 를 활용하여 제안하는 신경망을 학습하고, 4-폴드 교차 검증을 통해 실험을 수행하여 신경망 기반 오디오 초 해상도 기술의 성능을 분석하였다. 실험 결과 입력 저 해상도 오디오 대비 향상된 SNR 을 갖는 것을 확인하였으며, 스펙트로그램 상에서도 성능 향상을 확인할 수 있었다. 그러나 고품질 오디오 서비스를 위해서는 여전히 원본 오디오와 많은 차이를 보여 블라인드 및 부가 파라미터를 이용한 오디오 초 해상도 기술 등에 대한 지속적인 연구가 필요해 보인다.

감사의 글

본 연구는 한국전자통신연구원 연구 운영 지원 사업의 일환으로 수행 되었음. [20ZH1200, 초실감 입체 공간 미디어. 콘텐츠 원천기술 연구]

참고문헌

[1] V. Kuleshov, S. Z. Enam and S. Ermon, "Audio Super Resolution using Neural Nets", International Conference on Learning Representation (ICLR), 2017.

[2] 서유림, 강석주, “딥러닝 기반 Super Resolution 기술의 현황 및 최신 동향”, 방송과 미디어, 25(2), 2020.

[3] M. Dietz, L. Liljeryd, K. Kjorling and O. Kunz, “Spectral band replication a novel approach in audio coding”, Proc. 112th AES Convention, 2002.

[4] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam and J. Bello, “MedleyDB: A Multitrack dataset for annotation-intensive MIR research”, 15th International Society for Music Information Retrieval Conference (ISMIR), 2014.