https://doi.org/10.7840/kics.2022.47.9.1306

심층신경망 기반 오디오 디클리핑 방법

최 승 운◆, 최 승 호°

An Audio Declipping Method Based on Deep Neural Networks

Seung Un Choi◆, Seung Ho Choi°

요 약

본 논문은 클리핑 된 오디오 신호에서 원음을 복원해주는 디클리핑(declipping)에 관한 것이며, 이를 위해 심층신경망을 기반으로 한 새로운 방법을 제안한다. 본 기법은 우선 클리핑 된 오디오 샘플 개수를 기반으로 클리핑 프레임을 검출한다. 이후 클리핑 프레임과 원음 프레임의 크기 스펙트럼을 각각 심층신경망의 입출력으로 사용하여 네트워크를 훈련한다. 음성 데이터베이스의 원음과 복원된 신호 간의 RMSE와 LSD 비교 실험을 통해, 제안된 방법이 기존 방법에 비해 성능이 향상됨을 보였다.

Key Words : Audio declipping, Clipping detection, Deep neural network, Speech communication, Broadcasting audio

ABSTRACT

This paper is about declipping that restores the original sound from a clipped audio signal, and for this purpose, we propose a new method based on a deep neural network. This technique first detects clipping frames based on the number of clipped audio samples. Then, the network is trained using the magnitude spectra of the clipping frame and the original sound frame as input and output of the deep neural network. Through the experiment comparing the RMSE and LSD between the original sound and the reconstructed signal in the speech database, the proposed method showed that the performance was improved compared to the existing method.

Ⅰ. 서 론

음성통신, 방송 오디오, 음성인식 등의 음성 및 오디오 관련 서비스 시스템에서 마이크로폰 등을 이용한 오디오 신호 취득 시 입력 신호가 최대 한계를 초과하여 증폭될 때, 클리핑(clipping) 혹은 포화(saturation) 현상이 나타나게 된다. 이 현상은 음성통신, 방송 오디오 시스템에서 청취자에게 불쾌한 소음을 유발하며, 음성인식 서비스의 성능 저하를 초래한다. 클리핑 발생을 방지하기 위해 마이크의 이득(gain)을 줄일 수 있으나, 이는 취득한 오디오 신호의 해상도를 낮추는 결과를 초래한다. 오디오 디클리핑(audio declipping) 기술은 클리핑 된 구간 주변의 신호를 이용하여 원음을 복원하는 것이다.

기존 디클리핑 기술에는 autoregressive modeling[1], bandwidth limited model[2], Bayesian estimation[3] 등이 있었다. 최근에는 시간-주파수 희소성 (sparsity) 기반 디클리핑 기술[4]이 연구되었는데, 그 예로는 사전 학습(dictionary learning)을 기반으로 하는 consistent dictionary learning[5] 기법 등이 있다. 사전 학습 기법은 비지도학습(unsupervised learning) 기법의 하나로서, 각종 신호를 기저벡터(basis vector)의 선형결합으로 표현할 수 있다는 점에 착안하여 신호에서 잘리지 않은 샘플을 사용하여 원신호의 기저벡터를 추정하여 신호를 복원하는 방식이다. Consistent dictionary learning은 사전 학습 방법과 유사하나, 고정된 사전을 이용하는 일관된 희소 코딩 방식과 다르게, 클리핑 된 신호에서 사전을 학습하고, 클리핑 발생 시의 임계값도 고려하여 더 나은 복원 성능을 보여주었다. 그러나 기저벡터의 선형결합으로 원신호를 복원하는 이 기법은 클리핑 시 발생하는 비선형 현상을 모델링하는 데 한계가 있다. 본 논문에서는 이러한 비선형 현상을 잘 모델링 할 수 있는 심층신경

※ 이 연구는 서울과학기술대학교 교내연구비의 지원으로 수행되었습니다.
◆ First Author : Department of Electronic Engineering, Seoul National University of Science and Technology, seungun9275@gmail.com, 석사과정, 학생회원
° Corresponding Author : Department of Electronic and IT Media Engineering, Seoul National University of Science and Technology, shchoi@seoultech.ac.kr, 교수
논문번호 : 202208-168-A-LU, Received August 4, 2022; Revised August 9, 2022; Accepted August 9, 2022

망(deep neural network, DNN) 기반 디클리핑 기법을 제안한다.

Ⅱ. 심층신경망 기반 디클리핑 알고리즘

제안한 방법의 전체 흐름은 그림 1과 같다. 클리핑 된 신호가 입력되면, 클리핑 검출 알고리즘을 통과한다. 이 알고리즘은 한 프레임 안에서 신호의 크기가 같은 값으로 일정 시간 지속되면 클리핑 프레임으로 검출한다. 이때, 한 프레임 내의 신호가 전부 클리핑 되었을 때를 고려하여, 이 경우 훈련에 사용하지 않고, 오버랩을 적용하여 이전 프레임과 다음 프레임을 사용해 원 신호를 추정할 수 있도록 설계하였다.

심층신경망의 훈련을 위해, 클리핑 프레임으로 검출된 프레임들과 원음 프레임들을 FFT을 통해 크기 스펙트럼(magnitude spectrum)을 구하고 이들을 각각 DNN의 입출력으로 사용한다. 테스트 시에는 DNN 출력인 크기 스펙트럼 $|\hat{X}_t(\omega)|$과 위상 스펙트럼을 함께 IFFT(inverse FFT)하여 추정된 신호 $\hat{x}_t(n)$를 얻는다. 이때, 위상 스펙트럼은 클리핑 프레임의 위상을 사용한다.

그림 2는 원음과 클리핑 된 신호의 스펙트로그램을 나타낸 것이다. 그림 2에서 오디오 신호에서 클리핑이 발생하면 원음에는 없는 고주파 성분이 발생하는 것을 확인할 수 있다. 그림 3(a)와 3(b)는 각각 모델 훈련에 사용한 프레임과 사용하지 않은 프레임의 예시를 나타낸다. 그림 3(b)의 예시에서 보듯이, 일부 오디오 샘플들이 클리핑이 되었으나 조건을 충족하지 못하여 해당 프레임은 훈련에 사용되지 않는다. 그림 4는 클리핑 검출 결과의 예시이다.

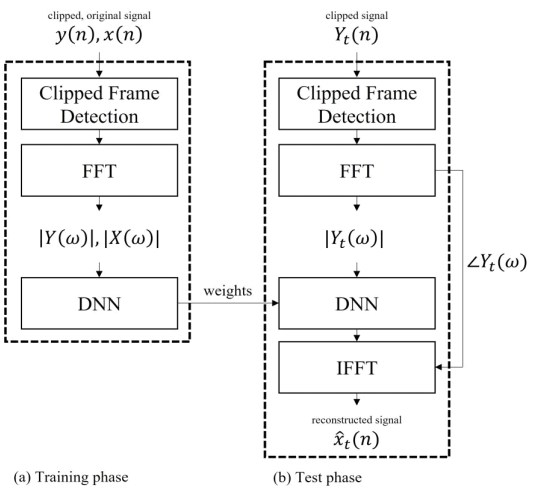

그림 1. 제안한 방법의 흐름도 (a) 학습 시 (b) 테스트 시
Fig. 1. Flow chart of the proposed method (a) Training phase (b) Test phase

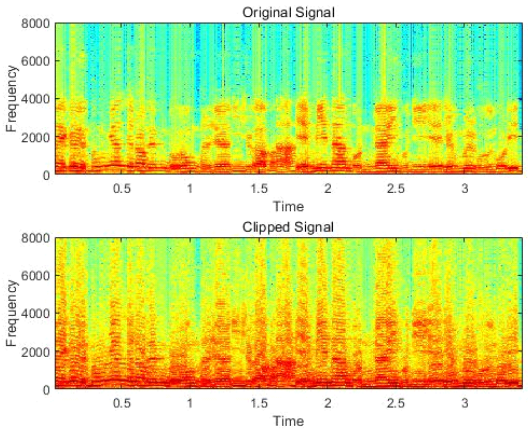

그림 2. 원음과 클리핑 된 신호의 스펙트로그램
Fig. 2. Spectrograms of original signal and clipped signal

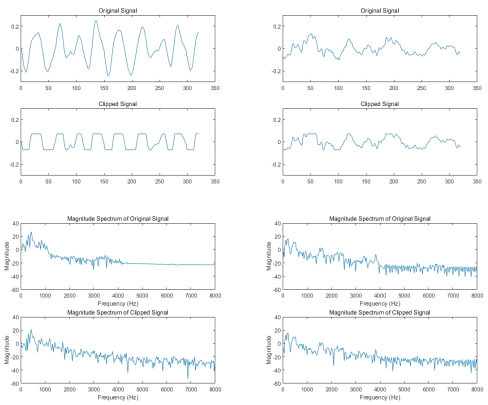

그림 3. (a) 훈련에 사용한 프레임 예시 (b) 훈련에 사용하지 않은 프레임 예시
Fig. 3. (a) Example of frame used for training (b) Example of frame not used for training

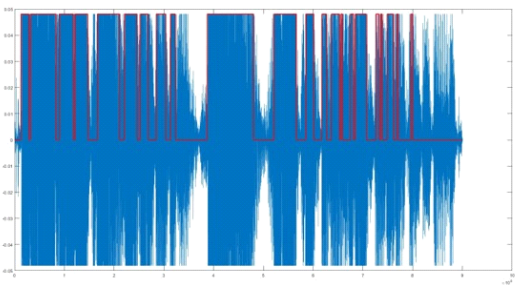

그림 4. 클리핑 검출 예시
Fig 4. Example of clipping detection

Ⅲ. 실험 및 결과

실험에 사용한 음성 데이터베이스는 NTT 한국어 음성[6]을 사용하였으며, 각 발성음은 5~15초 정도의 길이이다. 이중 훈련에는 약 300개 문장을, 테스트에는 100개 문장을 사용하였다. 프레임 길이는 20 ms로 50% 중첩을 하였다. 샘플링 주파수가 16 kHz이며, 한 프레임의 샘플 수는 320개 이다. 512 포인트 FFT를 통해 구한 257차의 크기 스펙트럼 벡터를 DNN의 훈련과 테스트에 사용하였다.

DNN 모델링을 위해 Python 환경에서 Keras, Tensorflow 라이브러리를 활용하였다. 활성화 함수는 마지막 층은 'ReLU' 함수를, 나머지 층은 'tanh'함수를 사용하였다. 손실 함수와 최적화 함수는 평균 제곱근 편차(mean-squared error)와 'Adam'[7]을 사용하였다. 그림 5는 제안한 DNN 구조와 입출력을 나타낸다.

본 논문에서는 디클리핑 성능을 평가하기 위해 모델이 추정한 신호와 실제 원 신호의 차이로서 아래 수식 (1)과 (2)와 같은 RMSE(Root Mean Square Error)와 LSD(Log Spectral Distance)를 사용하였다. 식 (1)에서 L은 총 샘플 개수이고, 식 (2)에서 K와 N은 각각 스펙트럼 차수와 총 프레임 개수이다.

$$RMSE = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(x_i - y_i)^2} \quad (1)$$

$$LSD = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{K}\sum_{k=1}^{K}(10\log_{10}(\frac{|Y(i,k)|^2}{|X(i,k)|^2}))^2} \quad (2)$$

그림 6은 평가에 사용한 데이터 중 예시로서, 원음, 클리핑 되었을 때 및 디클리핑 되었을 때 각각의 파형 및 스펙트로그램을 나타낸다. 그림에서와 같이, 클리핑 된 신호에 비해서 복원된 신호가 원음과 더욱 유사하며, 또한 클리핑 된 신호의 고주파 성분(원으로 표시한 부분)들이 복원된 신호에서는 제거되었음을 확인 할 수 있다.

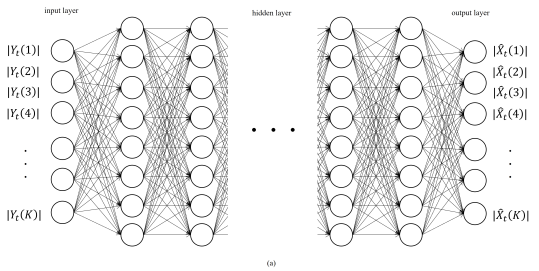

그림 5. 제안한 DNN 구조
Fig. 5. Structure of the proposed DNN

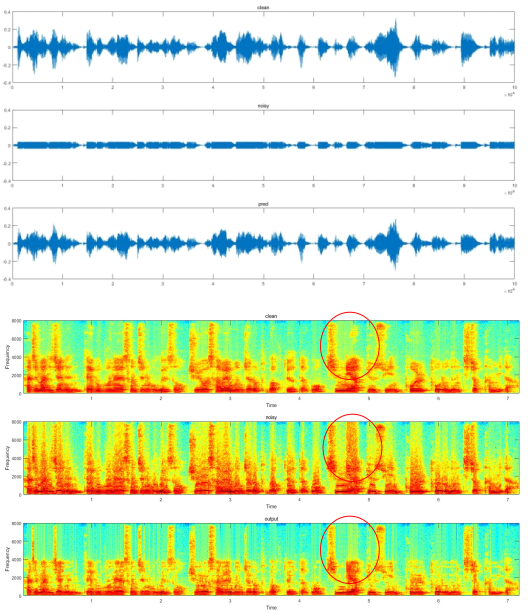

그림 6. 원음, 클리핑 된 신호 및 디클립 된 신호의 파형 및 스펙트로그램
Fig. 6. Waveforms and spectrograms of original, clipped, and declipped signal

표 1과 2는 기존의 consistent dictionary learning 기법과 본 논문에서 제안한 기법의 RMSE 및 LSD 비교 결과이며, 제안한 기법이 커다란 성능 개선을 보였다.

표 1. 처리 전과 처리 후 RMSE 비교
Table 1. RMSE comparison before and after processing

	Consistent Dictionary Learning	DNN
처리 전	0.033	
처리 후	0.022	0.016
개선율 (%)	33.30	51.52

표 2. 처리 전과 처리 후 LSD 비교
Table 2. LSD comparison before and after processing

	Consistent Dictionary Learning	DNN
처리 전	4.33	
처리 후	3.43	2.99
개선율 (%)	20.79	30.95

Ⅳ. 결론 및 향후 연구방향

본 논문에서는 클리핑 된 오디오 신호로부터 원 신호를 복원하기 위해 클리핑 프레임 검출 알고리즘과 심층신경망 기반의 디클리핑 기법을 제안했다. 우선 파형과 스펙트로그램을 통해 클리핑 현상에 대한 분석을 수행하였다. 그리고 기존의 consistent dictionary learning 기법과 제안한 기법의 RMSE 및 LSD 비교 실험 결과를 통해, 제안한 기법이 커다란 성능 개선을 보임을 확인하였다. 향후 원 신호를 복원할 때 스펙트럼의 크기뿐만 아니라 위상도 함께 추정하는 방법에 대한 연구를 수행할 계획이다.

References

[1] A. J. E. M. Janssen, R. Veldhuis, and L. Vries, “Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes,” *IEEE Trans. Acoustics, Speech, and Sign. Process.*, vol. 34, no. 2, pp. 317-330, 1986. (https://doi.org/10.1109/tassp.1986.1164824)

[2] J. S. Abel, “Restoring a clipped signal,” Acoustics, Speech, and Signal Processing, *IEEE Int. Conf. IEEE Comput. Soc.*, 1991. (https://doi.org/10.1109/icassp.1991.150655)

[3] S. J. Godsill, P. J. Wolfe, and W. N. W. Fong, “Statistical model-based approaches to audio restoration and analysis,” *J. New Music Res.*, vol. 30, no. 4, pp. 323-338, 2001. (https://doi.org/10.1076/jnmr.30.4.323.7489)

[4] C. Gaultier, et al., “Sparsity-based audio declipping methods: selected overview, new algorithms, and large-scale evaluation,” *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 29, pp. 1174-1187, 2021. (https://doi.org/10.1109/taslp.2021.3059264)

[5] L. Rencker, et al., “Consistent dictionary learning for signal declipping,” *Int. Conf. Latent Variable Anal. and Sign. Separation*, pp. 446-455, Springer, Cham, 2018. (https://doi.org/10.1007/978-3-319-93764-9_41)

[6] NTT-AT, *Multi-lingual speech database for telephonometry*, 1994.

[7] D. P. Kingma and J. Ba. “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014. (https://doi.org/10.48550/ARXIV.1412.6980)