일반논문 (Regular Paper)
방송공학회논문지 제28권 제1호, 2023년 1월 (JBE Vol.28, No.1, January 2023)
https://doi.org/10.5909/JBE.2023.28.1.132
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

클리핑 감지기를 이용한 음성 신호 클리핑 제거의 성능 향상

서 은 미[a)], 유 정 찬[a)], 임 유 진[a)], 박 호 종[a)‡]

Performance Enhancement of Speech Declipping using Clipping Detector

Eunmi Seo[a)], Jeongchan Yu[a)], Yujin Lim[a)], and Hochong Park[a)‡]

요 약

본 논문에서는 클리핑 감지기를 이용하여 음성 신호의 클리핑 제거 성능을 향상시키는 방법을 제안한다. 클리핑은 입력 음성 신호의 크기가 마이크의 동적 범위를 넘을 때 발생하며, 음성 품질을 저하시키는 요인이 된다. 최근 머신러닝을 이용한 많은 클리핑 제거 기술이 개발되었고 우수한 성능을 제공하고 있다. 그러나 머신러닝 기반의 클리핑 제거 방법은 신호 복원 과정의 왜곡으로 인해 클리핑이 심하지 않을 때 출력 신호의 품질이 저하되는 문제를 가진다. 이를 해결하기 위해 클리핑 제거기를 클리핑 감지기와 연동시켜 클리핑 수준에 따라 클리핑 제거 동작을 선택적으로 적용하는 방법을 제안하고, 이를 통해 모든 클리핑 수준에서 우수한 품질의 신호를 출력하도록 한다. 다양한 평가 지표로 클리핑 제거 성능을 측정하였고, 제안 방법이 기존 방법에 비해 모든 클리핑 수준에 대한 평균 성능을 향상시키고, 특히 클리핑 왜곡이 작을 때 성능을 크게 향상시키는 것을 확인하였다.

Abstract

In this paper, we propose a method for performance enhancement of speech declipping using clipping detector. Clipping occurs when the input speech level exceeds the dynamic range of microphone, and it significantly degrades the speech quality. Recently, many methods for high-performance speech declipping based on machine learning have been developed. However, they often deteriorate the speech signal because of degradation in signal reconstruction process when the degree of clipping is not high. To solve this problem, we propose a new approach that combines the declipping network and clipping detector, which enables a selective declipping operation depending on the clipping level and provides high-quality speech in all clipping levels. We measured the declipping performance using various metrics and confirmed that the proposed method improves the average performance over all clipping levels, compared with the conventional methods, and greatly improves the performance when the clipping distortion is small.

Keyword : speech declipping, clipping detector, U-Net, deep neural network

Ⅰ. 서 론

클리핑 (clipping)은 마이크와 증폭기 등의 장치에서 입력 신호의 크기가 장치의 동적 범위 (dynamic range)를 넘을 때 신호 값들이 포화 (saturation)되어 출력되는 왜곡 현상이다. 이러한 클리핑은 음성 통화, 영상 통화, VoIP 등의 음성 전송 과정에서 음성 품질을 저하시키고, 음성 인식에서 음성 신호를 왜곡시켜 인식률을 저하시킨다[1,2,3]. 따라서 클리핑이 발생한 음성 신호에서 클리핑을 제거하여 음성의 품질을 향상시키는 기술은 고품질 음성 통신과 음성 인식 등을 위해 필요한 연구 분야이다[3,4].

클리핑은 비선형 왜곡이므로 선형 동작으로는 제거가 어렵고 왜곡 모델링의 복잡도가 높아 전통적인 분석적 신호처리 (signal processing, SP) 기법으로는 해결하기 어려운 문제이다. 특히, 클리핑이 심하게 발생할 때 SP 방법의 성능은 크게 저하되고 출력 음성의 품질을 향상시키는데 한계를 가진다[5]. 최근 머신러닝 (machine learning, ML) 기술의 발전으로 ML을 이용한 클리핑 제거 방법이 많이 연구되고 있으며, ML 기법은 성능이 우수하여 기존의 SP 방법을 대체하고 있다[3]. 그러나 ML 방법은 일반적으로 신호 분석 후 새로운 신호를 복원하는 과정으로 진행되므로 클리핑 왜곡이 심하지 않을 때 오히려 음성 품질을 저하시키는 경향을 보이며, 이 경우에는 SP 방법보다 낮은 클리핑 제거 성능을 가진다. 이와 같이 SP 방법과 ML 방법은 각각 고유의 문제점을 가지며, SP 방법은 클리핑 왜곡이 심할 때 성능이 저하되고, ML 방법은 클리핑 왜곡이 심하지 않을 때 성능이 저하된다. 따라서 클리핑 왜곡 수준에 관계없이 우수한 성능을 가지는 새로운 클리핑 제거 방법이 필요하다.

본 논문에서는 기존 ML 기반의 클리핑 제거기의 문제점을 해결하기 위하여 클리핑 수준을 판단하는 클리핑 감지기를 활용하는 방법을 제안하고, 이를 통해 선택적 클리핑 제거를 구현하여 모든 클리핑 수준에서 우수한 성능을 가지도록 한다. 즉, 본 논문의 목표는 새로운 신경망을 설계하여 클리핑 제거 성능을 향상시키는 것이 아니라, 기존 ML 기반의 클리핑 제거기를 클리핑 감지 동작과 연동시키는 아이디어를 통해 클리핑 제거 성능을 향상시키는 것이다. 따라서 제안 방법은 신경망을 사용하는 간단한 클리핑 제거기를 클리핑 감지기와 결합시키고, 클리핑 감지기를 통해 얻는 성능 향상을 확인하여 제안 방법의 우수성을 검증한다.

본 논문에서 제안하는 방법은 클리핑 제거기와 클리핑 감지기가 순차적으로 연결된 구조를 가진다. 클리핑 제거기는 U-Net를 사용하여 입력 신호에 대한 클리핑 제거 동작을 수행한다[7,8,9,10]. 특히 클리핑은 신호 영역에서 발생하는 왜곡이고 클리핑된 신호는 주파수 영역보다 시간 영역에서 더 뚜렷한 특성 차이를 가지므로 U-Net가 시간 축 파형을 입력하여 모든 동작을 시간 영역에서 수행하도록 한다[6]. 클리핑 감지기는 입력 신호와 클리핑 제거기의 출력 신호를 ML 기반으로 분석하여 입력 신호의 클리핑 여부를 결정한다. 만일 클리핑이 존재하면 U-Net에서 클리핑을 제거시킨 신호를 최종 출력하고, 클리핑이 존재하지 않으면 입력 신호를 그대로 출력하여 클리핑 제거 과정에서 발생하는 음질 왜곡을 방지한다[11].

클리핑 제거 성능은 클리핑이 제거된 음성 신호의 품질로 측정하였고, 평가지표로 PESQ (perceptual evaluation of speech quality), SSNR (segmental signal-to-noise ratio), CSIG, CBAK, COVL을 사용하였다[12,13]. 모든 클리핑 수준에 대한 평균 성능에서 제안 방법이 기존 SP 및 ML 방법보다 우수한 것을 확인하였다. 또한, 클리핑 왜곡이 심하지 않을 때 제안 방법이 기존 ML 방법의 문제점을 해결하여 많은 성능 향상을 제공하고, 클리핑 왜곡이 심할 때는 제안 방법이 기존 ML 방법 대비 동등 이상의 성능을 가지는 것을 확인하였다. 이를 통해 클리핑 감지기를 적용한 제안 방법이 특정 클리핑 수준에서 성능이 저하되는 기존 방법의 문제를 개선시키고 모든 클리핑 수준에서 우수한 클리핑 제거 동작을 하는 것을 검증하였다.

a) 광운대학교 전자공학과(Dept. of Electronics Engineering, Kwangwoon Univ.)
‡ Corresponding Author : 박호종(Hochong Park)
E-mail: hcpark@kw.ac.kr
Tel: +82-2-940-5104
ORCID: https://orcid.org/0000-0003-1600-6610
※ 이 논문은 2022년도 광운대학교 교내학술연구비 지원과 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(NRF-2021R1F1A1059233)과 2022년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원(P0017124, 산업혁신인재성장지원사업)을 받아 수행된 연구임.
· Manuscript December 8, 2022; Revised January 25, 2023; Accepted January 25, 2023.

Ⅱ. 제안하는 방법

1. 클리핑된 신호 생성

클리핑 동작은 하드 (hard)-클리핑과 소프트 (soft)-클리핑으로 분류되며, 본 논문에서는 하드-클리핑만 다룬다[2]. 식 (1)이 하드-클리핑 동작을 보여주고, x_n은 시간 인덱스 n에서의 원 샘플값, $\widetilde{x_n}$는 클리핑된 샘플값, θ는 임계값, $sgn(\cdot)$은 부호함수이다.

$$\widetilde{x_n} = \begin{cases} x_n & \text{if } |x_n| < \theta \\ \theta\, sgn(x_n) & \text{if } |x_n| \geq \theta \end{cases} \tag{1}$$

본 논문에서 θ는 식 (2)와 같이 한 파일의 전체 신호 x에 대한 최댓값에 임의의 α를 곱하여 정의한다. α는 0.1에서 0.9까지 0.1 간격으로 주어지고, 각 파일마다 무작위로 결정된 α을 적용하여 클리핑을 구현한다.

$$\begin{gathered} \theta = \alpha \times \max(x) \\ \alpha \in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9] \end{gathered} \tag{2}$$

그림 1은 α가 0.2일 때 클리핑 전과 후의 시간 축 파형과 스펙트로그램을 보여주며, 그림 1(a)가 원 신호이고 그림 1(b)가 클리핑된 신호이다. 시간 축 파형은 신호의 최댓값을 기준으로 정규화한 결과이다. 스펙트로그램에서 볼 수 있듯이, 클리핑에 의하여 저대역 하모닉의 에너지가 줄어들고 고대역에 추가적인 하모닉 성분이 생겨 전체적으로 하모닉 구조에 변형이 발생한다.

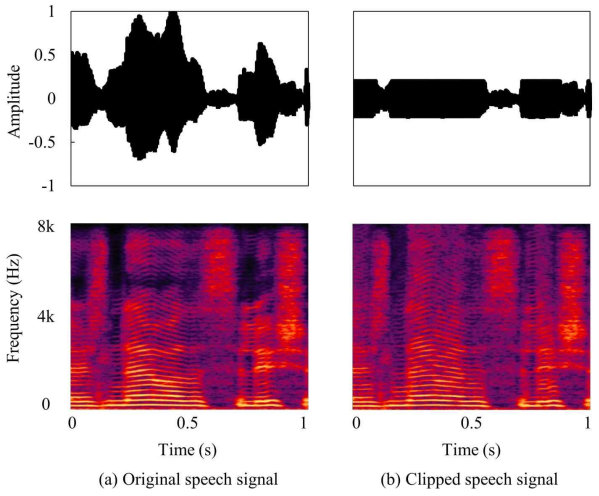

그림 1. 클리핑이 발생한 음성 신호의 예 (α = 0.2). (a) 원본 음성 신호, (b) 클리핑된 음성 신호
Fig. 1. Clipping example of speech signal (α = 0.2). (a) Original speech signal, (b) Clipped speech signal

2. 전체 동작 구조

제안 방법은 클리핑 제거기와 클리핑 감지기를 결합한 구조를 가지며, 클리핑 감지기의 역할은 입력 신호의 클리핑 여부를 판단하여 선택적 동작을 하는 것이다. 기존 클리핑 제거기에 클리핑 감지 기능을 추가하는 일반적인 구조는 클리핑 제거기의 전처리 과정에서 클리핑 감지 기능을 수행하고 그 결과에 따라 클리핑 동작을 선택하는 것이다. 이 구조에서는 입력 신호로부터 클리핑 여부를 판단해야 하며, 이 경우 클리핑 감지 정확도가 저하되고 그에 따라 최종 출력 신호의 품질이 저하되는 것을 실험을 통해 확인하였다.

이 문제를 해결하기 위하여 제안하는 방법은 그림 2와 같이 클리핑 제거기와 클리핑 감지기가 순차적으로 연결된 구조를 가진다. 클리핑 제거 신경망이 입력 신호 $\tilde{x}$에서 클리핑을 제거하여 $\hat{x}$를 출력하고, 클리핑 감지기는 $\tilde{x}$와 $\hat{x}$를 모두 입력하여 $\tilde{x}$의 클리핑 발생 확률을 구한다. 클리핑 수준에 따라 $\tilde{x}$와 클리핑 제거된 신호 $\hat{x}$ 사이의 차별 정도가 다르므로 $\tilde{x}$와 $\hat{x}$를 동시에 사용하면 $\tilde{x}$의 클리핑 발생 확률의 정확도를 높일 수 있다. 제안 방법과 같이 $\tilde{x}$와 $\hat{x}$를 사용하면 $\tilde{x}$만 사용할 때에 비하여 클리핑 감지 성능이 높아지고 최종 출력 신호 y의 품질이 향상되는 것을 실험을 통해

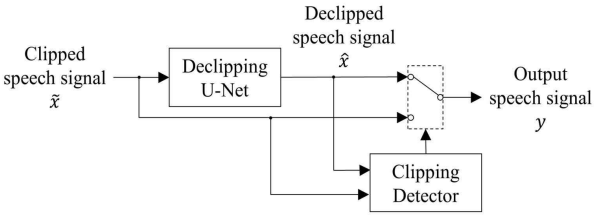

그림 2. 제안하는 클리핑 제거 방법의 전체 구조
Fig. 2. Overall structure of proposed declipping method

확인하였다.

본 논문에서 클리핑 감지기는 프레임 단위로 클리핑 여부를 판단하고, 프레임의 한 샘플이라도 클리핑 되면 클리핑 발생 프레임으로 정의한다. 클리핑 감지기는 각 프레임별로 $\tilde{x}$의 클리핑 확률 p를 계산하고 식 (3)와 같이 확률 0.5를 기준으로 클리핑 여부 d를 결정한다. 마지막으로 최종 출력 샘플 y_n는 식 (4)과 같이 결정되어 최종 출력은 $\hat{x}$와 $\tilde{x}$ 중 하나가 된다. 이와 같은 선택적 동작을 통해 클리핑 왜곡이 작은 신호가 입력될 때 클리핑 제거기에서 발생하는 음질 저하를 방지할 수 있다.

$$d = round(p) = \begin{cases} 1 \Rightarrow clipped \\ 0 \Rightarrow non-clipped \end{cases} \tag{3}$$

$$y_n = d \cdot \hat{x_n} + (1-d) \cdot \tilde{x_n} \tag{4}$$

3. 신경망 구조

앞에서 언급하였듯이 본 논문의 목표가 클리핑 제거를 위한 신경망 개발이 아니라 클리핑 감지기를 통한 성능 향상이므로, 클리핑 제거를 위한 신경망은 기존에 널리 이용되는 간단한 시간 영역 U-Net 구조를 사용한다[6]. U-Net 상세 구조는 그림 3과 같고, c는 채널 수, k는 커널 크기, s는 스트라이드 (stride), a는 활성함수 (activation function)를 나타낸다. 채널 변환을 위한 입력 층, 인코딩과 디코딩을 위한 각각 4개의 합성곱 (convolution) 층과 전치 합성곱 (transposed convolution) 층, 출력 층을 포함하여 총 10개 층으로 구성된다.

시간 영역 U-Net는 프레임 단위로 동작하며, 프레임 길이는 0.1 초이고 샘플링 주파수는 16 kHz이며, 한 프레임은 1600 샘플을 가진다. 한 프레임의 신호 $\tilde{x}$가 입력되면 1차원 합성곱 동작을 수행하여 (16 채널 × 100 차원)의 latent 벡터로 인코딩 되고, 1차원 전치 합성곱과 스킵 연결 (skip connection)을 수행하는 디코딩 과정을 거쳐 최종 1600 샘플의 $\hat{x}$이 출력된다. 신경망 입력과 출력에서 모두 sine 윈도우와 50% 중첩을 적용한다.

그림 4가 클리핑 감지기의 신경망 구조를 보여주고, 클리핑 제거기와 동일한 프레임 단위로 동작하고, 3단의 1차원 합성곱 층과 1단의 완전-연결 층 (fully-connected layer)으로 구성되고, u는 유닛 (unit) 개수이다. 클리핑 감지기는 각각 sine 윈도우가 적용된 1600 샘플의 $\tilde{x}$와 $\hat{x}$를 2 채널로 결합하여 입력하고, $\tilde{x}$에서의 클리핑 확률 p를 출력한다.

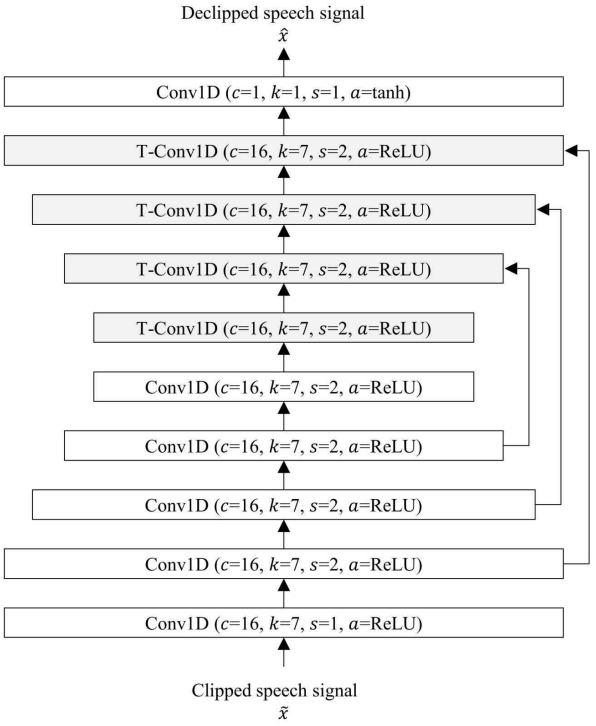

그림 3. 클리핑 제거를 위한 시간 영역 U-Net 구조
Fig. 3. Architecture of time-domain U-Net for declipping

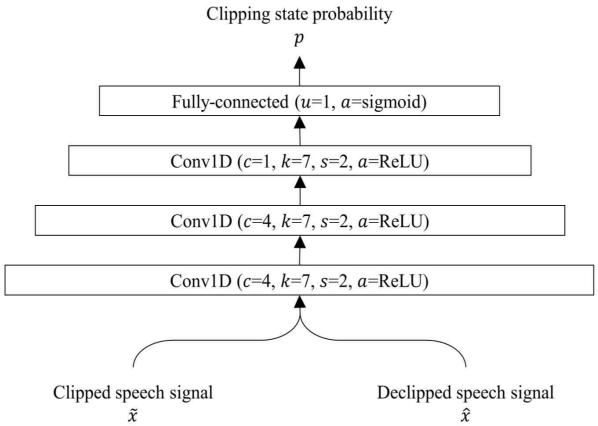

그림 4. 클리핑 감지기의 신경망 구조
Fig. 4. Network architecture of clipping detector

Ⅲ. 성능 평가

1. 데이터셋과 학습 방법

신경망 학습과 성능 평가에는 CSTR voice cloning toolkit (VCTK) 데이터셋을 사용하였다[14]. VCTK 데이터셋은 109명의 남녀 화자에 대한 약 44시간의 음성 신호로 구성되며, 데이터 분할은 표 1과 같이 하였다. 총 107명 화자에 대한 약 40시간의 음성 신호를 신경망 학습과 검증(validation)에 사용하였고, 나머지 약 4시간을 성능 평가에 사용하였다.

표 1. 학습과 성능 평가에 사용된 파일 개수와 시간
Table 1. The number of files and time length used in training and testing

	Train	Validation	Test
No. of files	35,922	4,376	3,945
Time length	35h 42m	4h 23m	3h 56m

식 (1)과 (2)의 방법에 따라 클리핑된 신호를 생성하였고, α는 균등 분포를 가지도록 파일마다 무작위로 선정하였다. 각 α를 사용하는 파일 수와 클리핑 발생 비율은 그림 5에 정리되어 있고, α가 클수록 클리핑 발생 프레임 비율이 낮은 것을 알 수 있다.

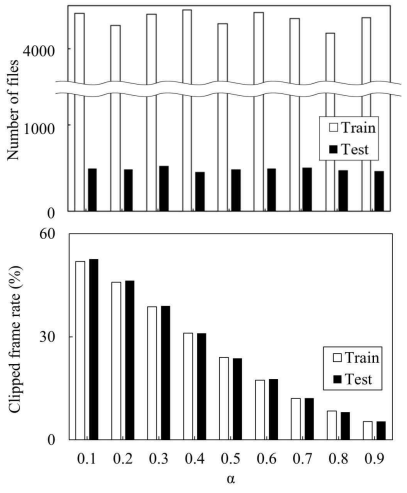

그림 5. 각 α를 사용하는 파일 수(위)와 각 α에 대한 클리핑 발생 프레임의 비율(아래)
Fig. 5. The number of files (upper) and the rate for clipped frame (lower) for each α

신경망 학습은 두 단계로 진행하였다. 먼저, 클리핑 제거를 위한 U-Net를 단독으로 학습하고, 이 때 클리핑 되지 않은 원 신호를 목표 신호로 사용한다. 다음, 클리핑 감지기와 학습된 U-Net를 결합하여 클리핑 감지기를 단독으로 학습하며, 미리 주어지는 클리핑 여부에 대한 정답을 목표 출력으로 사용한다. 모든 신경망 학습에서 Adam 최적화기, mini-batch 크기 64, learning rate 0.001을 사용하였고, 200 epoch로 학습 진행하였고, learning rate는 2 epoch 마다 0.99배 감쇠시켰다[15]. U-Net 학습에서의 손실함수는 시간 영역 MSE (mean squared error)이고, 클리핑 감지기 학습에서의 손실함수는 BCE (binary cross-entropy)이다.

2. 클리핑 제거 성능

제안한 클리핑 제거 방법의 성능 평가를 위해 SP 방법, 시간 영역 U-Net를 사용하는 ML-baseline 방법, 제안하는 방법의 성능을 비교하였다. 여기서 SP 방법으로 Adobe Audition 3.0에서 지원하는 Clip restoration 기능을 사용하였다[5]. 또한, ML-baseline 방법의 U-Net는 그림 3의 U-Net와 동일하며, 이를 통해 클리핑 감지기에 의한 성능 향상을 확인하고자 한다. 성능 평가를 위한 지표로는 PESQ, SSNR, CSIG, CBAK, COVL를 사용하였다[12,13].

표 2는 모든 α 값에 대한 평균 성능을 보여주며, 모든 평가 지표에서 제안 방법 (Prop.)이 가장 우수한 성능을 가진다. Prop.* 항목은 해당 절의 마지막 부분에서 설명한다. 그림 6은 α가 0.2일 때 각 방법으로 클리핑을 제거한 신호의 예시이다. 파형의 시간 축 포락선을 비교하면, Audition 3.0에서 포락선이 크게 왜곡되고 ML-baseline과 제안 방법은 모두 원 신호와 비슷한 포락선을 제공하는 것을 알 수

표 2. 클리핑 제거의 평균 성능
Table 2. Average declipping performance

	Clipped signal	Audition 3.0	ML-Baseline	Prop.	Prop.*
PESQ	3.12	3.20	3.51	3.62	3.63
SSNR	31.85	32.31	30.50	32.84	32.76
CSIG	4.65	4.64	4.80	4.89	4.88
CBAK	4.76	4.75	4.91	4.94	4.94
COVL	4.04	4.10	4.25	4.46	4.46

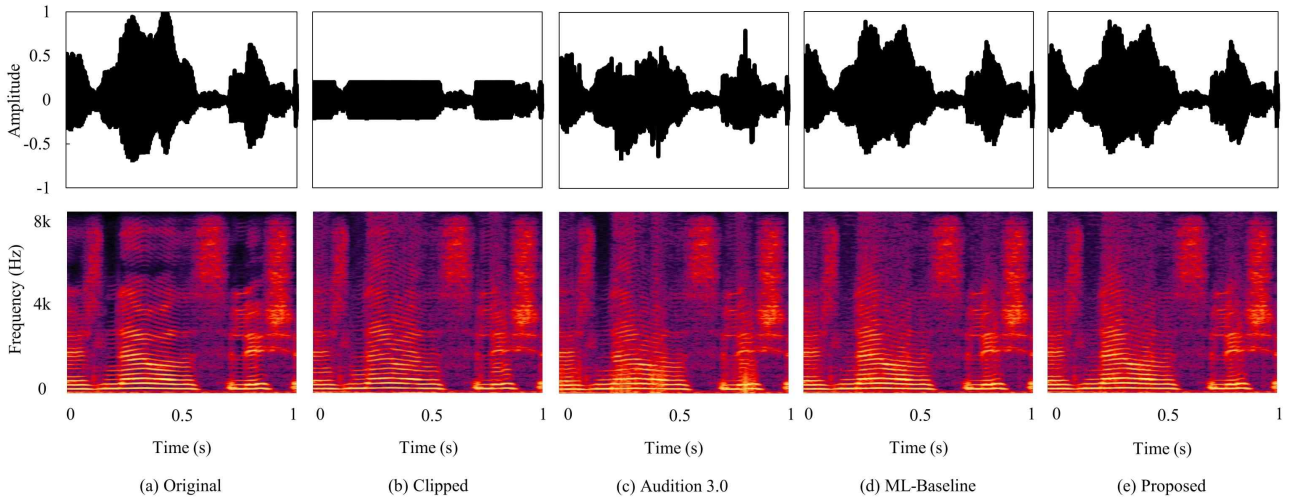

그림 6. 각 클리핑 제거 방법에 대한 시간 축 파형과 스펙트로그램 예시 (α = 0.2)
Fig. 6. Examples of waveform and spectrogram for each declipping method (α = 0.2)

있다. 이를 통해 클리핑 왜곡이 심할 때 SP 방법 대비 ML 방법의 우수성을 확인할 수 있다.

표 3은 각 α에 대한 성능을 보여준다. 여기서 α가 클 때 CSIG와 CBAK 값이 5.0이 되는 것은 해당 지표가 정확한 음성 품질을 측정 못하는 한계를 가지기 때문이다. 여러 방법의 클리핑 제거 성능을 α가 클 때와 작을 때로 구분하여 비교하면 다음과 같다. $\alpha < 0.7$일 때, 즉 클리핑 왜곡이 심할 때 전반적으로 Audition 3.0의 성능이 ML-baseline 및 제안 방법 성능 보다 낮다. 즉, SP 기법으로는 비선형 특성으로 심하게 클리핑된 신호를 복원하는데 한계가 있으며, 제안 방법을 사용하면 SP 기법의 문제점인 심한 클리핑에서의 성능 저하 문제를 해결할 수 있다. $\alpha > 0.7$이면, 즉 클리핑 왜곡이 작으면 ML-baseline이 Audition 3.0보다 낮은 성능을 가진다. 즉, ML-baseline은 클리핑 제거가 필요 없는 경우에서의 bypass 성능이 낮다. 그 이유는 클리핑 왜곡이 작을 때 신경망의 신호 복원 과정에서 오히려 신호 왜곡이 발생하여 클리핑 제거의 효과를 얻지 못하기 때문이다. 반면, 클리핑 감지기를 적용하면 클리핑 왜곡이 작을 때 클리핑 제거 동작을 선택적으로 중지하고 입력 신호를 그대로 출력하여 신호 왜곡을 방지하고 기존 ML 방법의 문제점을 해결할 수 있다. 만일, $\alpha = 1$ 이면 클리핑 감지기가 모든 프레임을 클리핑 미발생으로 판단하여 출력이 입력과 동일하게 되는 것을 확인하였고, 따라서 제안한 방법은 완벽한 bypass 동작을 수행한다.

이상의 성능 분석을 통해, SP 방법과 ML 방법은 특정 α에서 성능이 크게 저하되는 문제점을 가지지만, 제안 방법은 모든 α에서 기존 방법의 문제점을 해결하고 우수한 성능을 가지는 것을 알 수 있다. 즉, 제안 방법은 신경망과 클리핑 감지기의 결합을 통해 ML 방법의 장점을 유지하면서 클리핑 감지 동작을 통해 불필요한 클리핑 제거 동작을 배제하여 출력 신호의 품질을 향상시키는 것을 확인할 수 있다.

표 4가 클리핑 감지기의 성능을 보여준다. Positive와 negative는 각각 클리핑 발생과 클리핑 미발생을 의미하고, 숫자는 해당 프레임의 수이다. 이 결과로부터 민감도 (sensitivity, true positive rate) 0.79와 특이도 (specificity, true negative rate) 0.99를 얻는다. 대부분의 오류는 클리핑 발생 프레임을 미발생 프레임으로 판단하는 경우 (false negative)이고, 프레임 내에 클리핑된 샘플 수가 적거나 윈도우 영향으로 프레임 앞과 뒤 구간에서 클리핑 효과가 거의 나타나지 않을 때 발생하는 것으로 확인하였다. 그림 7은 false negative의 예시를 보여주며, 첫 번째는 화살표로 표시된 두 위치에서 클리핑이 발생하여 클리핑된 샘플 수가 매우 적은 경우이고, 두 번째는 화살표 위치에서 클리핑이 발생하였으나 윈도우 영향으로 샘플값이 크게 감쇄 되어 클리핑 현상이 뚜렷하게 나타나지 않은 경우이다.

표 3. 각 α에 대한 클리핑 제거 성능
Table 3. Declipping performance for each α

PESQ	Clipped signal	Audition 3.0	ML-Baseline	Prop.	Prop.*
0.1	1.42	1.08	2.66	2.67	2.67
0.2	2.10	1.97	3.03	3.05	3.05
0.3	2.63	2.66	3.29	3.33	3.33
0.4	3.04	3.16	3.51	3.58	3.57
α 0.5	3.30	3.52	3.64	3.73	3.74
0.6	3.58	3.82	3.76	3.90	3.91
0.7	3.81	4.06	3.86	4.04	4.05
0.8	4.01	4.25	3.90	4.14	4.16
0.9	4.24	4.37	3.93	4.22	4.25

SSNR	Clipped signal	Audition 3.0	ML-Baseline	Prop.	Prop.*
0.1	23.98	24.72	24.98	26.91	26.70
0.2	27.95	28.95	27.13	30.12	30.06
0.3	30.48	31.32	29.87	32.14	32.00
0.4	32.43	33.04	31.04	33.43	33.33
α 0.5	33.49	33.94	31.63	34.10	34.04
0.6	34.46	34.44	32.10	34.50	34.46
0.7	34.61	34.76	32.26	34.74	34.73
0.8	34.86	34.92	32.34	34.85	34.85
0.9	34.98	34.98	32.36	34.90	34.91

CSIG	Clipped signal	Audition 3.0	ML-Baseline	Prop.	Prop.*
0.1	3.51	3.33	4.17	4.37	4.36
0.2	4.10	4.09	4.51	4.72	4.71
0.3	4.52	4.58	4.73	4.91	4.70
0.4	4.82	4.87	4.88	4.98	4.98
α 0.5	4.94	4.97	4.94	5.00	4.99
0.6	5.00	4.99	4.97	5.00	5.00
0.7	5.00	5.00	4.99	5.00	5.00
0.8	5.00	5.00	4.99	5.00	5.00
0.9	5.00	5.00	4.99	5.00	5.00

CBAK	Clipped signal	Audition 3.0	ML-Baseline	Prop.	Prop.*
0.1	3.77	3.64	4.44	4.57	4.55
0.2	4.37	4.37	4.81	4.91	4.90
0.3	4.77	4.81	4.96	5.00	4.99
0.4	4.96	4.97	5.00	5.00	5.00
α 0.5	5.00	5.00	5.00	5.00	5.00
0.6	5.00	5.00	5.00	5.00	5.00
0.7	5.00	5.00	5.00	5.00	5.00
0.8	5.00	5.00	5.00	5.00	5.00
0.9	5.00	5.00	5.00	5.00	5.00

COVL	Clipped signal	Audition 3.0	ML-Baseline	Prop.	Prop.*
0.1	2.50	2.23	3.46	3.57	3.56
0.2	3.15	3.08	4.82	3.95	3.94
0.3	3.63	3.68	4.07	4.21	4.21
0.4	4.00	4.11	4.27	4.44	4.43
α 0.5	4.24	4.42	4.38	4.58	4.58
0.6	4.47	4.65	4.48	4.72	4.72
0.7	4.66	4.83	4.57	4.83	4.84
0.8	4.81	4.95	4.60	4.90	4.71
0.9	4.95	4.99	4.63	4.95	4.96

표 4. 클리핑 감지기의 성능
Table 4. Performance of clipping detector

True \ Predicted	Positive	Negative
Positive	66,218	17,268
Negative	1,706	246,003

표 2, 3의 Prop.* 항목은 클리핑 감지기 오류가 발생한 프레임을 정답으로 수정하여 클리핑 감지 오류 없이 동작시킬 때의 성능이고, 클리핑 감지기의 오류가 출력 음성 품질에 영향을 미치지 않는 것을 보여준다. 이를 통해 클리핑 감지기가 제안하는 방법에서 요구하는 충분한 성능을 제공하는 것을 확인할 수 있다.

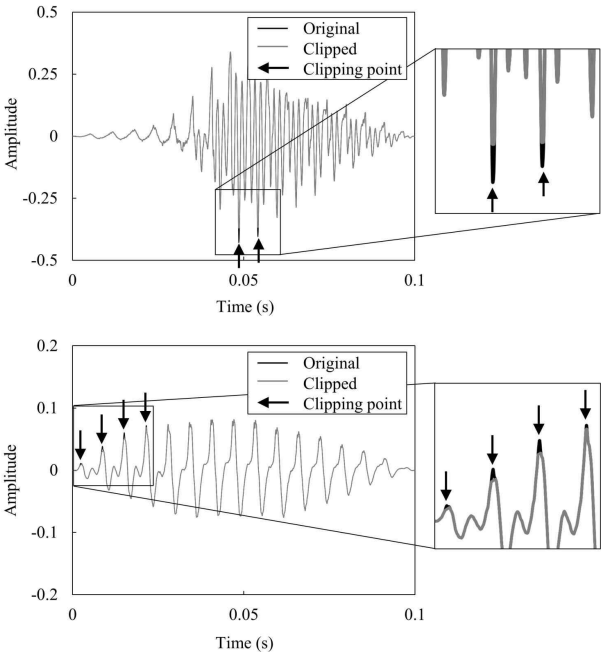

그림 7. 클리핑 감지기에서 오류가 발생한 프레임 예시
Fig. 7. Examples of error frame for clipping detector

Ⅳ. 결 론

본 논문에서는 ML 기반의 클리핑 제거기와 클리핑 감지기를 결합하여 클리핑 제거 성능을 향상시키는 방법을 제안하였다. 기존 ML 방법은 클리핑 왜곡이 심하지 않을 때

신경망의 신호 복원 과정에서의 왜곡으로 인하여 출력 신호의 품질을 저하시킨다. 반면, 제안 방법은 클리핑 여부를 판단하여 모든 클리핑 왜곡 수준에서 출력 신호의 품질이 유지되도록 동작을 조정한다. PESQ, SSNR, CSIG, CBAK, COVL를 이용하여 성능을 측정하였고, 제안 방법이 클리핑 왜곡이 작을 때 기존 ML 방법보다 우수한 성능을 가지고, 클리핑 왜곡이 심할 때 기존 ML 방법의 성능을 유지하는 것을 검증하였다.

본 논문에서는 간단한 클리핑 제거기에 대하여 클리핑 감지기와의 결합을 통한 성능 향상을 검증하였고, 향후 프레임 길이와 클리핑 제거 및 감지 성능의 관계를 분석하여 동작을 최적화할 예정이다. 또한, 신경망 구조의 최적화를 통해 고성능 클리핑 제거기를 개발하고 이를 클리핑 감지기와 결합시킬 때 클리핑 감지기 역할과 성능 향상에 대한 연구를 진행할 예정이다.

참 고 문 헌 (References)

[1] S. Maymon, E. Marcheret, and V. Goel, “Restoration of clipped signals with application to speech recognition,” *Proc. Interspeech*, pp. 3294 – 3297, Aug. 2013.
doi: https://doi.org/10.21437/Interspeech.2013-729

[2] W. Mack and E. A. P. Habets, “Declipping speech using deep filtering,” *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 200 – 204, Dec. 2019.
doi: https://doi.org/10.1109/WASPAA.2019.8937287

[3] P. Zaviska, P. Rajmic, A. Ozerov, and L. Rencker, “A survey and an extensive evaluation of popular audio declipping methods,” *IEEE J. Selected Topics in Signal Processing*, Vol. 15, No. 1, pp. 5 – 24, Jan. 2021.
doi: https://doi.org/10.1109/JSTSP.2020.3042071

[4] J. Y. Jung and G. B. Kim “Adaptation of classification model for improving speech intelligibility in noise,” *J. of Broadcast Engineering*, Vol. 23, No. 4, pp. 511 – 518, July 2018.
doi: https://doi.org/10.5909/JBE.2018.23.4.511

[5] Adobe Audition 3 USER GUIDE, https://help.adobe.com/archive/en_US/audition/3/audition_3_help.pdf (accessed Nov. 1, 2022).

[6] A. A. Nair and K. Koishida, “Cascaded time + time-frequency Unet for speech enhancement: Jointly addressing clipping, codec distortions, and gaps,” *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*. pp. 7153 – 7157, 2021.
doi: https://doi.org/10.1109/ICASSP39728.2021.9414721

[7] C. Macartney and T. Weyde, “Improved speech enhancement with the Wave-U-net,” arXiv preprint arXiv:1811.11307, Nov. 2018.
doi: https://doi.org/10.48550/arXiv.1811.11307

[8] S. Pascual, A. Bonafonte, and J. Serra, “SEGAN: Speech enhancement generative adversarial network,” *Proc. Interspeech*, pp. 3642 – 3646, Mar. 2017.
doi: https://doi.org/10.21437/Interspeech.2017-1428

[9] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, “U2-net: Going deeper with nested U-structure for salient object detection,” *Pattern Recognition*, Vol. 106, No. 107404, Oct. 2020.
doi: https://doi.org/10.1016/j.patcog.2020.107404

[10] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, “Phase-aware speech enhancement with deep complex u-net,” arXiv preprint arXiv:1903.03107, Mar. 2019.
doi: https://doi.org/10.48550/arXiv.1903.03107

[11] G. B. Kim “Binary mask estimation using training-based SNR estimation for improving speech intelligibility,” *J. of Broadcast Engineering*, Vol. 1, No. 6, pp. 1061 – 1068, Nov. 2012.
doi: http://dx.doi.org/10.5909/JBE.2012.17.6.1061

[12] Y. Hu and P. C. Loizou, “Evaluation of objective measures for speech enhancement.” *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 16, No. 1, pp. 229 – 238, Jan. 2008.
doi: https://doi.org/10.1109/TASL.2007.911054

[13] Evaluation measures open source, https://www.crcpress.com/downloads/K14513/K14513_CD_Files.zip (accessed Nov. 1, 2022).

[14] J. Yamagishi, C. Veaux, and K. MacDonald, “CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92),” *University of Edinburgh. The Centre for Speech Technology Research*, 2019.
doi: https://doi.org/10.7488/ds/2645

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

저 자 소 개

서 은 미

- 2021년 2월 : 광운대학교 전자공학과 학사
- 2021년 3월 ~ 현재 : 광운대학교 전자공학과 석사과정
- ORCID : https://orcid.org/0000-0002-5523-1522
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

유 정 찬

- 2021년 2월 : 광운대학교 전자공학과 학사
- 2021년 3월 ~ 현재 : 광운대학교 전자공학과 석사과정
- ORCID : https://orcid.org/0000-0003-0441-1280
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

임 유 진

- 2021년 2월 : 광운대학교 전자공학과 학사
- 2021년 3월 ~ 현재 : 광운대학교 전자공학과 석사과정
- ORCID : https://orcid.org/0000-0002-5233-0728
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

박 호 종

- 1986년 2월 : 서울대학교 전자공학과 공학사
- 1987년 12월 : Univ. of Wisconsin-Madison 공학석사
- 1993년 5월 : Univ. of Wisconsin-Madison 공학박사
- 1993년 9월 ~ 1997년 8월 : 삼성전자 선임연구원
- 1997년 9월 ~ 현재 : 광운대학교 전자공학과 교수
- ORCID : https://orcid.org/0000-0003-1600-6610
- 주관심분야 : 오디오/음성 신호처리, 3D 오디오, 음악정보처리