JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 23, No. 9, pp. 1847-1855, Sep. 2022

Check for updates

음질 개선을 위한 새로운 활성함수와 데이터 전처리를 가진 4단계 U-Net 신경망 제안

김 세 하[1] · 김 동 회[2*]
[1]강원대학교 IT대학 전기전자공학과 학사과정
[2*]강원대학교 IT대학 전기전자공학과 교수

The proposal of 4 Steps U-Net Neual Network with New Active Function and Data Preprocessing for Sound Quality Improvement

Se-Ha Kim[1] · Dong-Hoi Kim[2*]
[1]Undergraduate, Electrical and Electronic Engineering, IT College, Kangwon National University, Chuncheon, Korea
[2*]Professor, Electrical and Electronic Engineering, IT College, Kangwon National University, Chuncheon, Korea

[요 약]

본 논문에서는 기존 U-Net 신경망의 예측 과정에서 고주파수 영역의 음성의 데이터를 잡음으로 잘못 인식하여 원본 데이터가 잡음과 같이 제거되는 문제점을 개선하고자 한다. 제안하는 U-Net 신경망에서는 음성에 비해 잡음의 주파수 분포가 위쪽 영역에 존재하는 것을 확인하여 음성과 잡음을 차단할 특정 주파수을 찾아냈고 그 주파수 이상의 영역을 차단하는 전처리를 우선적으로 거침으로써 기존 U-Net 신경망 알고리즘과의 음질 향상 성능의 개선을 달성할 수 있었다. 또한 가중치가 업데이트되는 과정에서 최적의 파라미터를 찾아가는 새로운 활성함수를 사용함으로써 과적합을 방지하고 검증 손실값을 낮출 수 있었다. 평가지표 SNR, RMSE를 통해 음질의 개선을 정량적으로 평가하였다. 실험을 통해서 SNR은 50%, RMSE는 30% 이상 성능이 개선되었음을 확인하였다. 새로운 활성함수 PReLU를 사용함으로써 검증 손실값이 30%가량 낮아지는 결과를 확인하였다.

[Abstract]

In this paper, we aim to improve the problem that the original data are removed like noise by mistakenly recognizing the data in the high frequency domain as noise in the prediction process of the existing U-Net neural network. The proposed U-Net neural network identified a specific frequency to block the voice and noise by confirming that the frequency distribution of the noise exists in the upper region compared to the voice, and the improvement in sound quality with the existing U-Net neural network algorithm was achieved by prioritizing preprocessing. In addition, by using a new active function that finds the optimal parameters in the process of updating the weights, it is possible to prevent overfitting and lower the verification loss value. Improvement in sound quality was quantitatively evaluated through evaluation indicators SNR and RMSE. Through the experiment, it was confirmed that the performance of SNR was improved by 50% and RMSE by 30% or more and the verification loss value decreased by 30% by using the new activation function PReLU.

색인어 : 파이썬, 딥러닝, 인공지능, 음질향상, 활성함수, 신경망 알고리즘
Keyword : Python, Deep Learning, Artificial Intelligence, Speech Enhancement, Active Function, Neural Network

http://dx.doi.org/10.9728/dcs.2022.23.9.1847

Received 23 July 2022; **Revised** 16 August 2022
Accepted 18 August 2022

***Corresponding Author; Dong-Hoi Kim**
Tel: +82-33-250-6349
E-mail: donghk@kangwon.ac.kr

Ⅰ. 서 론

음향 기술 및 오디오 신호처리는 중요한 분야로 이미 자리매김하고 있다. 우리의 스마트폰에서 많이 사용하는 음성인식을 통해 우리의 삶의 질을 증가시키거나, 블루투스 이어폰을 통한 통화품질 및 음질 향상 또한 오디오 신호처리이다. 우리가 영화를 볼 때 영화관에서 듣는 3D 음향 기술 또한 오디오 기술이다. 이러한 음향 및 오디오 신호처리의 대부분은 딥러닝을 이용한 기술로써, 우리는 딥러닝의 기술 향상에 따른 삶의 질의 증가는 필수불가결하다. [1]

신호처리는 자기지도 학습이다. 기존에 학습법은 비지도학습과 지도 학습이었지만 둘의 차이는 레이블된 데이터의 유무이다. 비지도 학습이 레이블된 데이터가 없기 때문에 데이터의 특징에 따라 범주로 묶은 Clustering을 수행하고 지도학습은 분류나 예측의 목적으로 활용되었다. 하지만 최근에 구글에서 발표된 자연어처리 관련 BERT[2]를 살펴보면 전체 문장에서 하나의 단어를 지운 후 그 단어를 추측하는 방법과 그 다음에 올 문장에 대해서 추측하는 방법으로 기학습한다. 이를 자기 스스로 학습한다고 하여 자기지도학습이라고 한다.

딥러닝 기술의 초기 단계에서는 CNN(Convolutional Neural Network)을 주로 이용했지만, 오디오 신호처리는 시간 영역에서 분석을 할 여지가 남아 있고 그에 따른 DNN(Deep Neural Network), RNN(Recurrent Neural Network), FCN(Fully Convolutional Network), LSTM(Long Short Time Memory) 등, 다양한 발전된 분야에서 널리 쓰여지고 있다[3][4][5][6]. 오디오 신호처리의 초기에는 시간 영역에서 푸리에 변환을 통한 분석을 했지만 [7], 최근의 기술에서는 Mel Spectrogram을 이용해 시간 영역 과 주파수 영역를 동시에 분석함으로써 더욱 정밀하고 정확하게 분석하는 것이 가능해졌다[8]. 이에 따라 현재 우리가 사용하고, 또 그 정확성에 놀라는 많은 기술들이 발전되어져 왔다.

기존 U-net 알고리즘에서 prediction 부분에서 소음을 제거하는 과정에서 잡음이 덜 제거되거나 타겟 목소리가 같이 제거되어 마지막에 추출된 오디오에서 목소리가 잘 들리지 않는 문제점이 발생하였다. 따라서 본 논문에서는 현재 많이 사용되는 Mel Spectrogram을 이용한 U-Net 알고리즘으로 잡음을 제거하기 위한 목적에 있다. 그 과정에서 잡음과 음성의 데이터를 분석하고 그에 따라 저주파 통과 필터를 이용해서 차단하는 주파수를 파라미터값으로 둔 뒤 최적의 SNR을 찾고, 해당 주파수에 기반하여 활성함수로 PReLU를 사용한다.

결과적으로 기존의 U-Net알고리즘과 전처리 과정을 통한 U-Net과 PReLU(Parametric ReLU) 알고리즘의 결합한 새로운 제안 방법에 대해 SNR(Signal-Noise Ratio) 및 RMSE(Root Mean Square Error) 비교를 통해 성능 개선의 결과를 보여주고자 한다.

본 논문에서의 제안 방법은 새로운 활성함수 PReLU를 사용함으로써 학습 과정에서 epoch 수가 점점 증가함에 따라 과적합이 발생할 수 있는 가능성을 낮추고 검증 손실(validation loss)값이 떨어짐을 확인하였다. 또한 오디오 데이터를 멜 스펙트로그램과 페이즈 스펙트로그램으로 나누기 전 단계에서 일정 주파수 위 구간의 영역을 일차적으로 제거한 뒤 학습 과정을 진행함으로써 두 가지 평가지표인 SNR 및 RMSE에서 성능이 향상된 결과를 얻을 수 있었다.

본 논문의 Ⅱ장에서는 기존의 음질 향상을 위한 연구를 설명하며 Ⅲ장에서는 제안하는 4단계 알고리즘의 순서도를 설명하고 음성 데이터를 분석한다. Ⅳ장에서 실험 환경을 설명하고 실험에 대한 결과값을 보여준다. Ⅴ장에서는 제안하는 연구에 대한 실험을 결과값에 따라 정리한다.

Ⅱ. 기존 연구

2-1 Speech Enhancement

Speech Enhancement는 딥러닝을 사용한 분야에서 오디오에 섞인 잡음을 제거하기 위한 개념으로 통화품질, Noise Canceling등 다양한 분야에서 여러 방면으로 사용되고 있다. CNN, DNN, FCN등 음질 향상을 위한 연구는 계속되고 있다.[3][4][5][6]

Speech Enhancement 분야는 좁게는 통화품질 개선으로 볼 수 있지만, 넓게는 다양한 방면으로 사용된다. Speech Enhancement는 사용자가 원하는 데이터 신호를 추출하여 원하는 신호 외에의 잡음은 제거하여 잡음이 없는 환경에서의 데이터에 최대한 다가가고자 하는 개념이다. 이는 우리가 공연을 녹화하거나, 운전 중에 네비게이션이 음성를 인식할 때 원하는 가수의 음성이나 밴드의 소리, 원하는 목적지로 가고자 할 때의 자신의 음성이 사람들의 환호성이나 외부의 잡음, 자동차의 배기음 등을 제거해야만 사용자가 원하는 음성 데이터를 얻을 수 있다. 이처럼 음성 향상을 사용한 알고리즘은 음성에 국한된 알고리즘이 아니라 사용자가 원하는 데이터를 추출하고자 할 때에도 널리 쓰이게 된다.

2-2 Mel Spectrogram

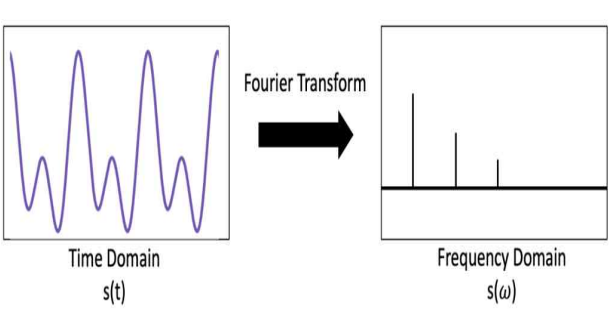

그림 1. 푸리에변환
Fig. 1. Fourier Transform

원본의 오디오 신호는 여러 주파수들이 겹쳐져 있는 상태이므로 이를 푸리에 변환을 통해 각각의 주파수들을 추출해야 한다. 푸리에 변환은 하나의 파형을 다양한 주파수를 가진 성분들로 분해를 해주지만, 그림 1을 참고하면 시간 영역의 파형을 주파수 영역으로 변환하여 각 주파수의 크기 정보밖에 알 수 없기 때문에 해석에 어려움이 생길 수밖에 없다. 이를 위해서 STFT(Short-Time-Fourier-Transform)을 이용해 시간에 대해서 자른 프레임마다 옆으로 쌓아서 시간 영역과 같이 해석할 수 있는 여지를 만들어 준다[9]. 이를 완료하면 전체 시간 영역 신호에 대한 주파수에 대한 정보를 얻을 수 있다. Mel Spectrogram은 현재 오디오 및 음성 처리 분야에서 가장 많이 활용되고 있는 신호처리 방법이다. 여기에서 Mel은 달팽이관을 모티브로 따온 것으로 사람의 귀는 저주파 대역에서는 주파수나 주파수의 변화를 잘 감지하는 반면, 고주파 대역에서는 잘 감지하지 못한다. 따라서 이러한 특성에 맞춰서 특징을 고려해야 한다. 이러한 특징을 고려한 값을 Mel-scale이라고 한다. 이를 이용하면 Mel filter bank가 나오는데 이를 STFT한 결과에 곱해주고 dB Magnitude로 변경하면 Mel Spectrogram이 추출된다.

$$Mel(f) = 2595\log(1+f/700) \quad (1)$$

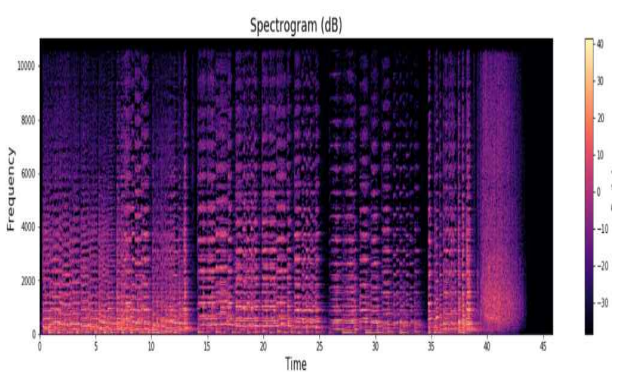

그림 2. 스팩트로그램
Fig. 2. Spectrogram

그림 2는 한 음성과 잡음이 섞인 데이터에 대해서 스펙트로그램으로 변환한 그림이다. 그림을 보았을 때 x축은 축이고 각 시간마다 주파수의 크기를 보여준다. 이를 통해서 각 잡음에 대한 스팩트로그램을 학습하고 분류가 가능하게된다.

2-3 기존의 U-Net

U-Net은 FCN을 기반으로 개선된 알고리즘으로써, 본래 의료영상 분야에 활용하기 위해 만들어진 신경망이다. U-Net은 그림 3과 같이 알고리즘 구조의 형태가 U자 형태를 띄고 있어서 U-Net이라고 명명한다.[10] U-Net은 맨아래 꼭지점을 기준으로 왼쪽을 축소영역, 오른쪽을 확장 영역이라고 한다. 또한 알고리즘 중간 공백의 화살표는 FCN의 장점중 하나인 Skip-Connection 구조를 나타낸다.

그림 3은 U-Net의 알고리즘 구조를 보여주고 있는데, 축소 영역은 CNN을 따르며, downsampling을 위한 Stride2, 2x2 max pooling 연산 및 ReLU를 두번 반복하여 3x3 unpadded convolutions 연산을 거친다. 그림 3과 같이 downsampling 과정에서 feature map의 채널 수는 연산을 거칠 때마다 2배로 증가시킨다.

확장 영역은 2x2 convolutions를 통해 upsampled 된 feature map과 축소영역의 cropped feature map과 결합 후, ReLU연산을 포함하는 두 번의 3x3 convolutions 연산을 거친다. 이때, 결합과정은 위치 정보가 날아가는 것을 방지하기 위해서 실행하는데, 이를 skip-connection이라고 부른다. 각 연산마다 채널의 수를 반으로 줄이는데, 이는 축소영역의 최종 feature map보다 높은 해상도를 얻기 위함이다.

U-Net의 구조에서 가장 중요한 것은 skip-connection이다. 앞서 말한 바와 같이 skip-connection은 이미지 분류를 위해 가장 중요한 위치 정보를 잃지 않고, 최대한 원본의 데이터와 유사하게 분류하고자 사용되는 구조이다. 이 구조 덕분에 원본 데이터와 가장 유사한 데이터를 얻을수 있게 된 것이다. 이러한 구조는 기존 FCN의 구조보다 확장된 개념으로써, 시간이 단축될 뿐만 아니라 아주 적은 양의 학습 데이터만으로도 우수한 성능을 보여준다.

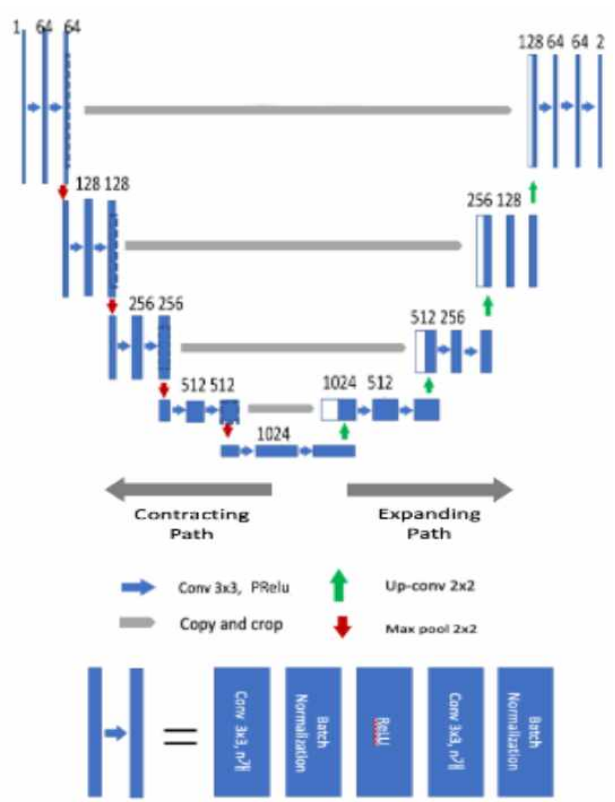

그림 3. U-Net 알고리즘 구조
Fig. 3. Structure of U-Net algorithm

2-4 기존의 3단계 알고리즘

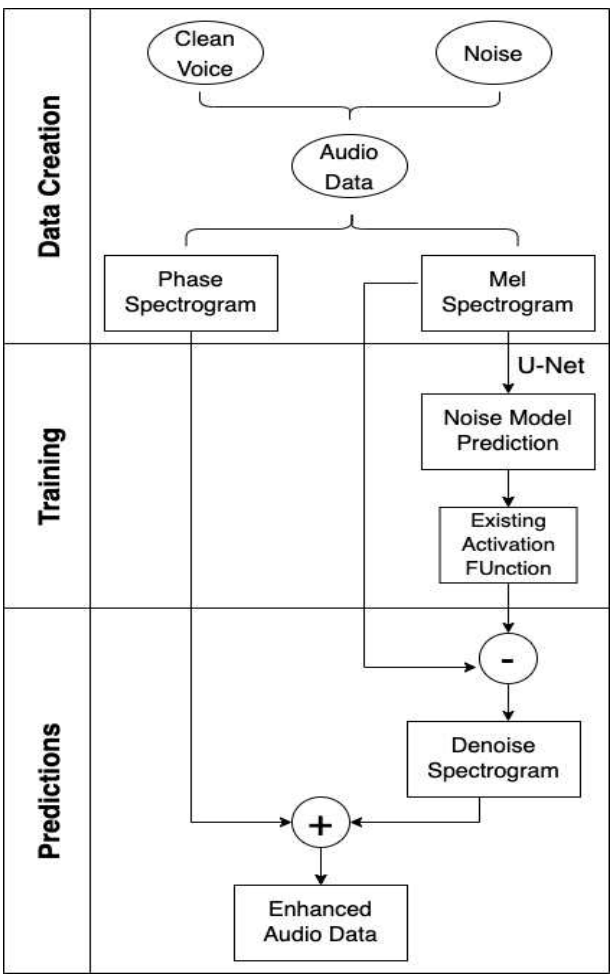

그림 4. 기존 U-Net 알고리즘 순서도
Fig. 4. Original U-Net Algorithm flowchart

기존의 U-Net의 과정은 크게 3단계로 거친다. 음성과 잡음을 합치는 데이터 합성, 합친 데이터를 중심으로 학습시키는 학습, 학습된 데이터를 이용해 잡음과 타겟 음성를 분리시키는 예측이그 단계이다.

그림 4는 기존 U-Net 알고리즘 순서도의 구조를 보여주고 있다. 데이터 합성 단계에서 음성과 잡음을 결합하는 동시에, 음성, 잡음, 혼합 데이터를 STFT을 거쳐서 각각의 데이터에 대한 크기 스펙트로그램과 위상 스펙트로그램으로 분리시킨다. 분리된 데이터중 스펙트로그램의 데이터를 U-Net 신경망을 거쳐서 Noise Model Prediction을 추출한다. 마지막인 예측 단계에서는 혼합 데이터의 크기 스펙트로그램에서 학습 단계에서 추출된 데이터를 제거한다. 그 후의 남은 데이터와 위상 스펙트로그램을 ISTFT(Inverse Short Time Fourier Transform)을 통해 소리 데이터로 복원시키는 것으로 U-Net 신경망의 잡음 제거 알고리즘이 종료되고 알고리즘의 출력값은 Softmax로 예측된다.

2-5 기존의 활성함수

기존의 3단계 알고리즘에서 사용된 활성함수는 Leaky ReLU 함수이다. 이는 ReLU 함수를 개선시킨 함수로써 ReLU는 그림 5과 같이 y가 음수의 구간에서의 기울기가 0으로 일정하다. 하지만 이는 입력값이 음수일 경우에 기울기가 0이 되어 가중치가 업데이트되지 않는다는 문제점이 발생한다. 이를 죽은 뉴런이라고 불린다.[11]

따라서 이를 개선한 함수가 Leaky ReLU이다[12]. 그림 6과 같이 y가 음수인 구간에서 기울기를 파라미터값으로 주어진다면 입력값이 음수이더라도 출력값이 0이 되지 않는, 죽은 뉴런을 방지할 수 있다. 파라미터값은 사용자가 지정할 수 있으며 0보다 크고 1보다 작은 값이면 어떠한 값이라도 정할 수 있다.

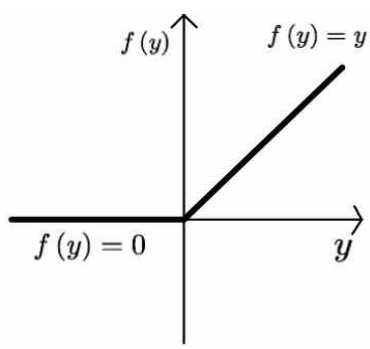

그림 5. ReLU 그래프
Fig. 5. ReLU graph

$$f(y_i) = \begin{cases} y_i & y_i > 0 \\ 0 & y_i < 0 \end{cases}$$
$$f(y_i) = \max(0, y_i) \quad (2)$$

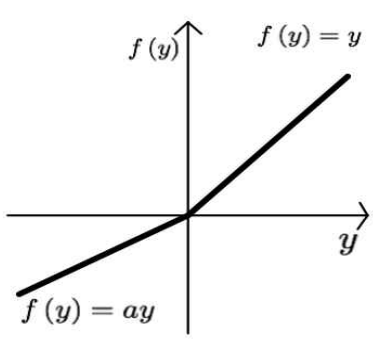

그림 6. Leaky ReLU 그래프
Fig. 6. Leaky ReLU graph

$$f(y_i) = \begin{cases} y_i & y_i > 0 \\ ay_i & y_i < 0 \end{cases}$$
$$f(y_i) = \max(ay_i, y_i) \quad (3)$$

Ⅲ. 제안하는 4-Step U-Net Network

3-1 기존의 문제점

기존의 연구에서는 잡음과 음성의 주파수가 겹치는 구간에서 잡음이 잘 제거되지 않거나 제거되었다고 하더라도 잡음이 제거되는 과정에서 타겟 음성과 같이 제거되는 문제점이 발생하였다. 이는 예측과정에서 추출된 오디오 데이터를 들었을 때, 타겟 음성의 소리가 명확하게 들리지 않는 문제점이 발생하였다.

따라서 이번 연구에서는 잡음 제거와 타겟 음성 추출의 정확도를 높이기 위해서 전처리 과정으로 일정 주파수를 먼저 차단한 뒤에 알고리즘을 실행함으로써 기존 U-Net의 문제점을 개선 시키고자 한다.

3-2 오디오 데이터 분석

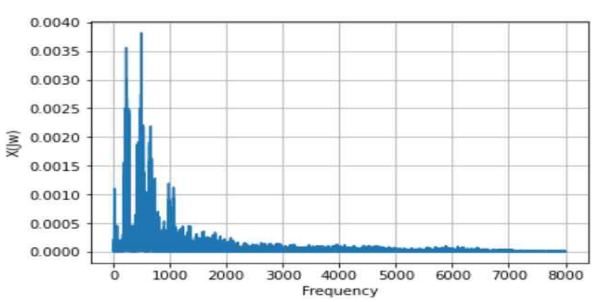

그림 7. 음성의 주파수 분포
Fig. 7. Voice Frequency Spectrum

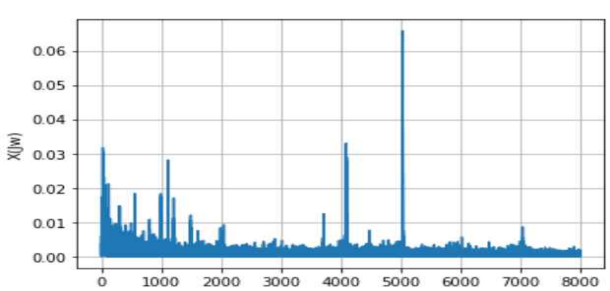

그림 8. 잡음의 주파수 분포
Fig. 8. Noise Frequency Spectrum

그림 7과 8은 음성과 잡음의 주파수 분포를 나타낸다. 그림 7과 8을 통해 알 수 있듯이, 음성에 비해 잡음의 주파수 분포가 위쪽 영역에 존재하는 것을 확인할 수 있다. 따라서 본 실험에서는 저주파 통과 필터를 이용해서 통과할 주파수의 영역을 2000Hz부터 8000Hz까지 500Hz씩 올려가며 4단계 알고리즘을 실험한다. 이를 통해 차단할 주파수를 5500Hz로 선정한 후 기존의 3단계 알고리즘과 SNR 값을 비교해가며 가장 좋은 SNR값의 차단 주파수를 이용하여 실험을 진행한다.

3-3 제안하는 활성함수

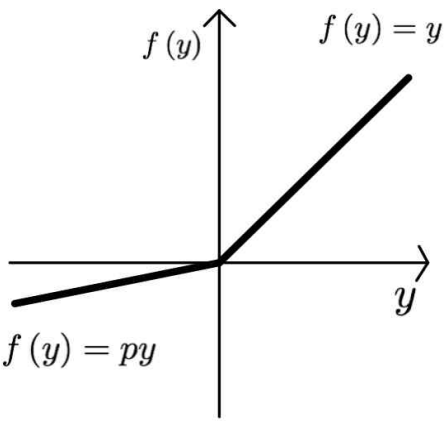

그림 9. PReLU 그래프
Fig. 9. PReLU graph

$$f(y_i) = \begin{cases} y_i & y_i > 0 \\ p_i y_i & y_i < 0 \end{cases}$$

$$f(y_i) = \max(0, y_i) + p_i \min(0, y_i)$$

$$\frac{\partial f(y_i)}{\partial p_i} = \begin{cases} 0 & y_i > 0 \\ y_i & y_i \leq 0 \end{cases} \quad (4)$$

그림 6의 Leaky ReLU 는 음수의 구간에서의 기울기가 고정되어 있는 상태에서 학습을 진행한다. 고정되어 있는 상태에서 학습하면 설정한 기울기의 값이 최적의 파라미터가 아니라고 하더라도 학습을 계속 진행하게 된다. 하지만 그림 9의 PReLU는 이름 그대로 음의 구간에서의 기울기를 파라미터값으로 두고 학습을 통해 계속적으로 업데이트를 통해서 최적의 값을 찾아가는 것이 PReLU의 개념이다.[13] 학습이 될 때마다 음의 구간의 파라미터가 업데이트되는 과정에서 최적의 기울기를 찾아감으로써 검증 손실의 값을 최대한 낮출 수 있다는 장점이 있다. 검증 손실값을 낮추면 낮출수록 과적합을 방지할 수 있다는 장점 또한 있다.

3-4 제안하는 4단계 알고리즘

그림 10는 제안하는 4단계 U-Net 순서도를 나타내고 있다. 제안하는 4단계 알고리즘이 기존의 3단계와 다른 점은 Data Creation 단계 전에 제안하는 전처리를 우선적으로 한다는 것이다. 전처리의 과정은 차단할 주파수의 값을 파라미터 값으로 정한 뒤에 정한 주파수 이상의 영역의 데이터를 삭제한 뒤에 기존의 알고리즘을 진행하였다. 실험을 진행할 때 전처리를 제외한 다른 하이퍼 파라미터값들은 동일한 상태에서 진행하였다.

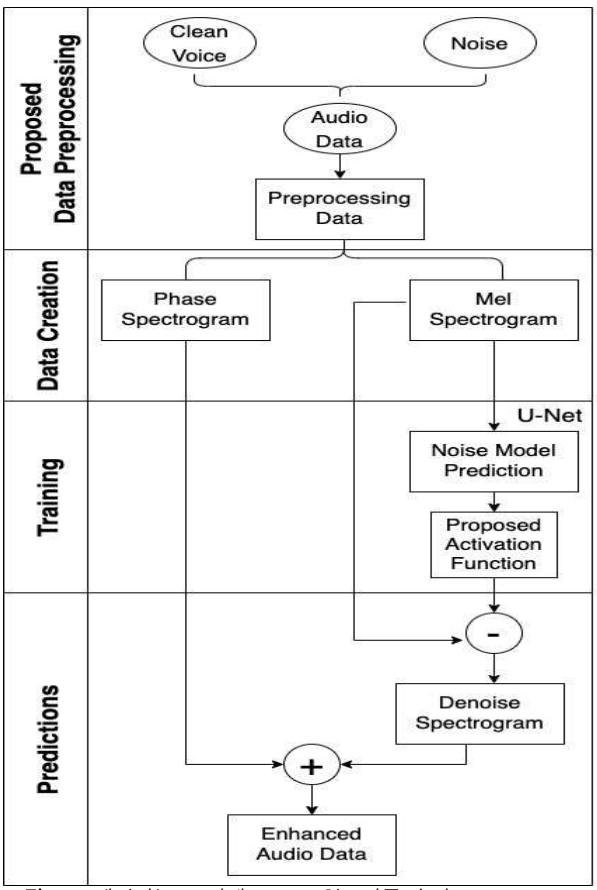

그림 10. 제안하는 4-단계 U-Net 알고리즘 순서도
Fig. 10. Flowchart of proposed 4-steps U-Net Algorithm flowchart

Ⅳ. 실험 결과 및 성능 평가

4-1 실험 환경

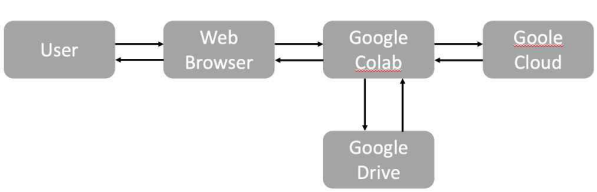

그림 11. 구글 코랩 구조
Fig. 11. Google colab structure

그림 11은 본 논문에서 사용한 실험 환경으로 구글 코랩 구조를 보여주고 있다. 구글 코랩은 구글에서 제공하는 Jupyter Notebook을 기반으로 한 클라우드 기반의 가상의 개발 환경이다. 코랩은 파이썬을 주로 사용하며 구글 클라우드의 GPU를 무료로 사용 할 수 있다. 본 논문에서는 구글 코랩을 사용하였으며 그림11은 구글 코랩의 구조를 나타낸다.[14]

4-2 실험 방법 및 평가지표

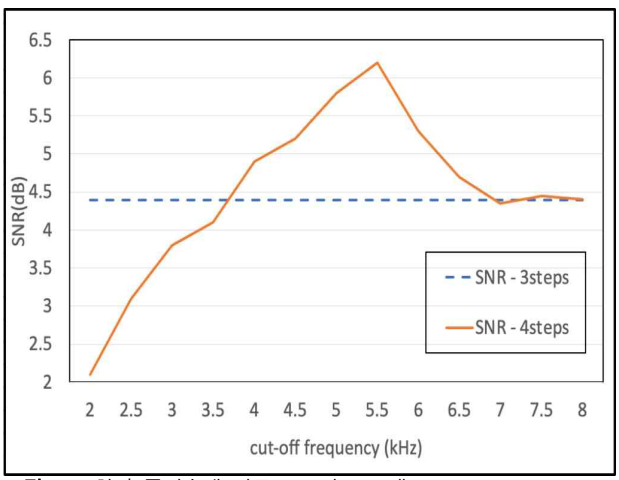

그림 12. 차단 주파수에 따른 snr 비교 그래프
Fig. 12. SNR comparion graph according to cut-off frequency

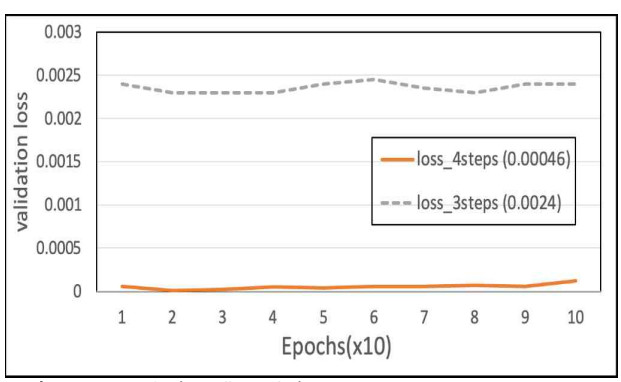

그림 13. 검증 손실 그래프 향상도
Fig. 13. Encreasement validation loss graph using PReLU

그림 12는 전처리 과정을 진행하기 위해 차단할 최적의 주파수를 찾기 위한 실험을 그래프로 나타내었다. 그림 12를 보았을 때 5500Hz에서 차단했을때 가장 높은 SNR값이 도출됨을 볼 수 있다. 또한 그림 12를 통해 2000Hz에서 3500Hz사이에서는 타겟 음성 데이터가 잡음의 데이터보다 더 많음을 확인하였고 이는 잡음의 주파수 분포가 저주파보다 고주파에 더 많이 분포가 되어 있음을 확인할 수 있었다. 또한 7000Hz 이상에서는 타겟음성과 잡음의 데이터가 거의 동일함을 알 수 있었다.

활성 함수를 Leaky ReLU에서 PReLU로 바꾼 뒤 측정한 검증 손실 값을 아래의 그래프에 나타내었다.

기존 3단계에서도 검증 손실 값이 더 내려갈 여지가 있었지만, 과적합의 문제로 더 이상 낮아지지 못하는 단점을 PReLU로 대체함으로써 검증 손실 값을 30% 이상 낮추는 결과값을 확인할 수 있었다. 그림 13은 검증 손실 그래프 향상도의 그래프를 나타내고 있다. 그림 13에서 x축은 epoch 수를, y축은 10번마다 도출된 검증 손실 값을 평균값을 보여주고 있다.

본 논문에서 그림 12를 통해서 5500Hz의 SNR값이 가장

높게 나온 것을 확인할 수 있다. 이 실험에서는 차단하는 주파수를 5500Hz로 고정한 뒤 실험을 계속 진행할 예정이다. 따라서 본 실험에서는 SNR 비가 가장 좋은 5500Hz에서 실험을 진행한다.

얻어낸 데이터에 대해서 평가를 하기 위해서 평가지표로 잡음의 제거 정도를 알 수 있는 SNR와 RMSE를 사용하였다. SNR의 수식 및 계산과정은 식 4와 같이 계산된다.

$$SNR = 20log_{10}\left(\frac{Voice^2}{Noise^2}\right) \quad (5)$$

SNR에서 Voice는 예측 단계에서 추출된 음성 데이터이고, Noise는 음성데이터와 잡음이 결합된 데이터이다. 본 논문에서 SNR의 Voice로 사용할 데이터는 두 가지이다. 첫번째는 기존 연구에 대한 음성 데이터이고, 두 번째는 4-step을 통해 추출된 음성 데이터이다. RMSE는 평균 제곱근 오차로서 원본 데이터와 예측한 데이터의 오차값을 모두 더한 값의 제곱의 평균을 더하고 그 값에 제곱근을 취한 값이다. RMSE의 식은 식 5와 같다.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (6)$$

위의 식에서 yi는 원본 음성 데이터이고, yi^hat은 예측 단계에서 추출된 음성 데이터이다.

4-3 잡음 제거 성능 평가 및 분석

차단할 주파수를 5500Hz로 고정한 뒤 실험한 결과에 대한 SNR 및 RMSE 그래프를 그림 14와 그림 15에 나타내었다. 그림 14와 15는 기존 3단계의 알고리즘과 제안하는 4단계 알고리즘에 대해서 예측단계에서 도출된 결과값에 대해서 SNR과 RMSE 비를 구한 것이다. SNR의 계산은 식 (4)를 통해 계산할 수 있고 RMSE의 계산은 식 (5)를 통해 계산하였다. 이번 실험에서 총 epoch 수는 90번으로 오차를 최대한으로 줄이고자 하였다. 그림 14의 x축은 epochs이고 y축은 SNR값을 나타내었다. 또한 그림 15의 x축은 epochs이고 y축은 RMSE값을 나타내었다. 결과값을 비교해 보았을 때, SNR값은 50%정도 증가하고 RMSE는 30%이상 성능이 개선되었음을 확인할 수 있었다.

그림 16은 2장에서 제시한 DNN, FCN등 기존의 연구들에 대해서 SNR값의 평균을 계산해서 비교한 표이다. 기존의 연구들의 SNR값은 평균적으로 3-4사이로 나타나고 있다. 본 연구에서 제안한 4단계 U-Net 알고리즘은 6.16으로 기존 연구에 비해서 SNR값이 더 높은 것을 확인할 수 있다.[3][4][5][6]

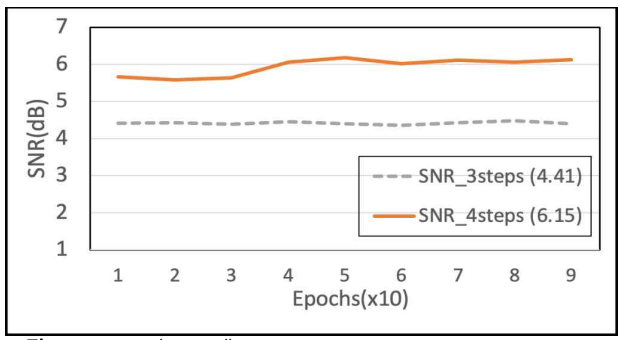

그림 14. SNR 비교 그래프
Fig. 14. SNR comparion graph

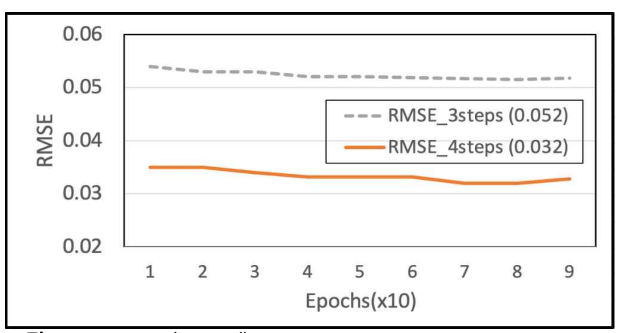

그림 15. RMSE 비교 그래프
Fig. 15. RMSE comparison graph

	DNN	FCN	Proposed U-Net
Average SNR	3.15	3.92	6.15

그림 16. SNR 수치 비교
Fig. 16. SNR numerical comparison

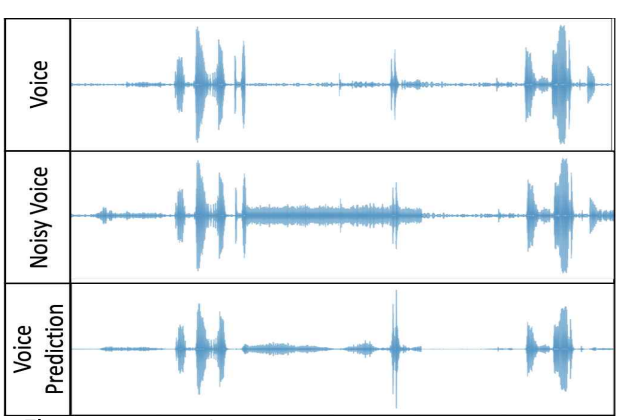

그림 17. Amplitude 비교 waveform
Fig. 17. Amplitude waveform compare

그림 14와 그림 15를 통해서 5500Hz 이상의 주파수를 차단한 전처리의 데이터가 SNR과 RMSE 값을 향상시키는 것을 볼 수 있었고, 그림 17은 Mel-Spectrogram을 통해서 예측과정을 통해 잡음이 제거된 데이터가 원본 데이터와 유사한 것을 확인할 수 있었다.

	DNN	FCN	Proposed U-Net
Average SNR	3.15	3.92	6.15

그림 18. 그림 15의 peak값 평균치
Fig. 18. Average peak value in Figure 15

그림 17은 위에서부터 순서대로 원본 음성 데이터, 잡음이 섞인 데이터, 알고리즘을 거친 예측 단계에서 추출된 잡음이 제거된 음성 데이터중 일부이다. 또한 그림 18은 그림 17에 대해서 peak값의 평균을 각 음성 데이터별로 구한 것이다. 그림 18에서 보는 것처럼 Noisy Voice의 피크값의 평균값이 높은 걸로 보아 Noise가 많이 섞여 있고, Voice Prediction의 수치가 Voice의 값과 비슷한 것으로 보았을 때 소음의 제거가 잘 되었다고 판단할 수 있다.

V. 결론 및 향후 연구

본 논문에서의 제안하는 U-Net 신경망에서는 고주파 영역에서의 잡음의 제거가 잘 되지 않는 문제점을 저역 통과 필터를 이용해서 우선적으로 고주파 영역의 데이터를 제거 후에 알고리즘을 실행시킴으로써 SNR과 RMSE와 같은 평가지표가 상승되는 것을 확인하였고 활성함수를 가중치가 업데이트되는 과정에서 최적의 파라미터를 찾는 새로운 PReLU로 교체함으로써 과적함을 방지하여 검증 손실값 또한 낮추는 걸과를 확인했다.

향후 연구에서는 최근에 음성 향상의 알고리즘으로 많은 각광을 받고 있는 GAN(Generative Adversarial Nets) 알고리즘과 이번 연구를 결합한, U-Net 베이스의 판별자를 GAN 알고리즘에 적용함으로써 원본의 음성 데이터에 가장 가까운 음성 데이터를 추출할 수 있을 것으로 예상된다.[15]

참고문헌

[1] H. Purwins, B. Li, and J. Schluter, “Deep Learning for Audio Signal Processing,” *IEEE Journal of Selected Topic in Signal Processing,* Vol. 13, pp. 206-219, April 2019. https://arxiv.org/abs/1905.00078

[2] J. Devliin, M. -W. Chang, K. Lee, K. Toutanova, “BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv:1810*.04805, May 2019. https://arxiv.org/abs/1810.04805

[3] Strake, M., Defraene, B., Fluyt, K. et al, “Speech enhancement by LSTM-based noise suppression followed by CNN-based speech restoration,” *Springer Series in Information Sciences,* 49, December 2020.

[4] E. Yeredor, A. Koldovsky, and Z. Tichavsky, “Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR," *Lecture Notes in Computer Science,* Vol. 9237, pp. 91-99, August 2015. https://link.springer.com/chapter/10.1007/978-3-319-2248 2-4_11

[5] Y. Xu, J. Du, L .-R, Dai and C. -H.Lee, “A Regression Approach to Speech Enhancement Based on Deep Neural Networks,” I*EEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 23, no. 1, pp. 7-19, January 2015.

[6] Z. Ouyang, H. Yu, W. -P.Zhu, B. Champagne, “A Fully Convolutional Neural Network for Complex Spectrogram Processing in Speech Enhancement,” *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* May 2019, pp. 5756-5760. 10.1109/ICASSP.2019.8683423

[7] J . Benesty, S. Makino, and J. Chen, “ The unimportance of phase in speech enhancement” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 30, pp. 679-681, August 1982.

[8] J.Shen et al, “Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions,” 2018 *IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP),* April 2018, pp. 4779-4783. 10.1109/ICASSP.2018.8461368

[9] D. Griffin, J. Lim, “Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. 32, pp. 236-243, April 1984.

[10] O. Ronneberger, P. Fischer, and T. Brox, “U-Net : Convolutional Networks for Biomedical Image Segmentation,” *International Conference on Medical Image Computing and Computer-Assisted Intervention* Vol. 9351, pp. 234-241, November 2015. https://link.springer.com/chapter/10.1007/978-3-319-2457 4-4_28

[11] A.F. Agarap, “Deep Learning using Rectified Linear Units(ReLU),” *arXiv*:1803.08375, March 2018. https://doi.org/10.48550/arXiv.1803.08375

[12] J. Xu, Z. Li, B. Du, M. Zhang, J. Liu, “Reluplex made more practical : Leaky ReLU,” *S2020 IEEE Symposium on Computers and Communications (ISCC),* July 2020, pp. 1-7 10.1109/ISCC50000.2020.9219587

[13] W. Qingjie, W. Wenbin, “Research on image retrieval using deep convolutional neural network combining L1 regularization and PReLU activation function,” *IOP Conference Series : Earth and Environmental Science,* Vol. 69, May 2017 https://iopscience.iop.org/article/10.1088/1755-1315/69/1/012156

[14] M. Nelson, and A. Hoover, “Notes on using Google

Colaboratory in AI education," *ITiCSE* :pp. 533-534, June 2020. https://dl.acm.org/doi/10.1145/3341525.3393997

[15] I. Goodfellow, J. Pouget-Abadie, M. Mizra, B. Xu, D. Warde-Farley, S.Ozair, A. Courville and Y. Bengio, "Generative adversarial nets," *Advanced in Neural Information Processing Systems,* pp. 2672-2680, 201. https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

김세하(Se-Ha-Kim)

2016년～현　재: 강원대학교 IT대학 전기전자공학과 재학

※관심분야 : 인공지능(AI), 딥러닝, Speech Enhancement

김동회(Dong-Hoi Kim)

2005년 : 고려대학교 전파공학과 (공학박사)

1989년 1월～1997년 1월: 삼성전자 전임연구원
2000년 8월～2005년 8월: 한국전자통신연구원 선임연구원
2006년 3월～현　재: 강원대학교 IT대학 전기전자공학과 교수
2020년 6월～2020년 6월: 강원대학교 정보화본부장 등
※관심분야 : 무선 네트워크 및 사물인터넷(IoT) 등