# Hifi-GAN 기반의 음성 대역폭 확장

Speech Bandwidth Extension with Hifi-GAN

지도교수 : 유승주

이 보고서를 공학학사 학위 논문

대체 보고서로 제출함.

2021 년 9 월 02 일

서울대학교 공과대학

산 업 공 학 과

황 보 진 경

2021 년 9 월

# Hifi-GAN 기반의 음성 대역폭 확장

Speech Bandwidth Extension with Hifi-GAN

지도교수 : 유승주

이 보고서를 공학학사 학위 논문

대체 보고서로 제출함.

2021 년 9 월 02 일

서울대학교 공과대학

산 업 공 학 과

황 보 진 경

2021 년 9 월

# 초 록

최근 무선 이어버즈를 사용하여 통화하는 상황이 증가하면서 음성 품질을 향상하기 위한 연구가 활발히 이뤄지고 있다. 네트워크 및 송수신 기기의 코덱 지원 여부에 따라 원음의 대역폭 제한 혹은 코텍 아티팩트로 인한 음질의 손실이 발생한다. 이를 해결하기 위해 본 연구에서는 디바이스 단에서 음성의 명료도를 높일 수 있는 GAN 기반의 음성 대역폭 확장 딥러닝 모델을 제안한다. TTS 보코더 분야의 SOTA 로 알려진 Hifi-GAN 을 기반으로 광대역 음성 신호를 초광대역 음성 신호로 확장하였으며, 정성적 성능평가를 통해 제안된 BWE 기술을 적용한 경우 30%가량의 MUSHRA 점수 향상을 이룩하였다. 또한, 학습 데이터에 존재하지 않는 화자에 대해서도 뛰어난 일반화 성능을 보여주었다. 여러 번의 코덱 부호화 · 복호화를 거쳐서 코덱 아티팩트가 존재하는 음성 신호에 대해서도 같은 방법론을 적용하여 효과적으로 코덱 아티팩트를 복원하고 음성 대역폭 확장을 이뤄낼 수 있음을 증명하였다.

**주요어 : 음성 대역폭 확장, 딥러닝, GAN, 코덱 아티팩트 복원**

# 목 차

# Hifi-GAN 기반의 음성 대역폭 확장

## 1 Introduction

대부분의 스마트 기기들이 50~14,000Hz 의 주파수 대역을 표현하는 초광대역 신호(SWB, Super Wide Band)의 음질을 제공하는 것에 반해, 네트워크 송수신 과정을 거치면서 음질 저하가 발생한다. 네트워크를 통해 음성이 실시간으로 전달되는 과정에서는 전송 지연을 줄이고 효율을 높이기 위해 코덱을 사용한다. 이 과정에서 원음의 대역폭 제한 및 코덱의 부호화·복호화 과정에서 발생하는 아티팩트(artifact)로 인해 음질의 손실이 발생한다.

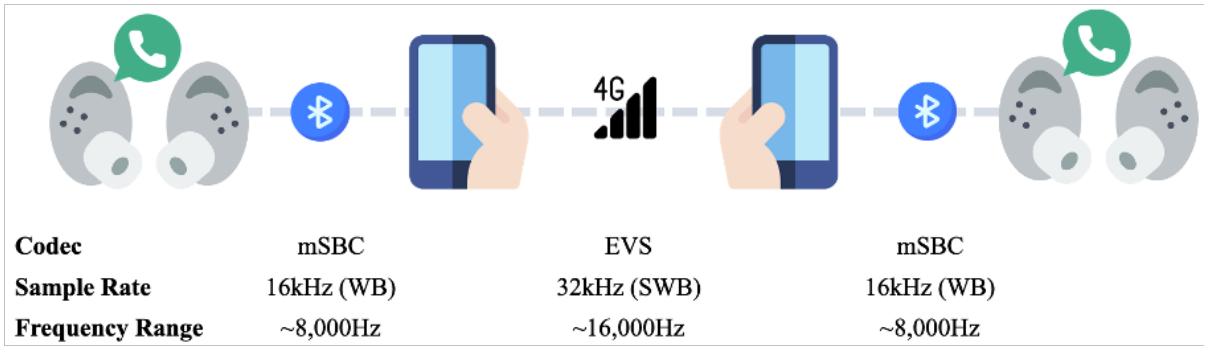



Figure 1 무선 이어버즈를 착용한 두 사람이 통화하는 경우 음성 신호의 일반적인 전송 과정을 그린 모식도이다. 음성 통신 과정에서 사용되는 코덱은 각각 블루투스와 이동통신 표준화 기술 협력 기구(3GPP , 3rd Generation Partnership Project)에서 현재 표준으로 채택된 코덱이다.

최근에는 Figure 1 과 같이 무선 이어버즈를 사용하여 통화하는 상황이 증가하였다. 블루투스 통신 과정에서 mSBC(modified SBC) 코덱을 이용한 부호화·복호화 과정을 거치기 때문에 일반적으로 통화 음성의 품질은 광대역 신호(WB, Wide Band)로 제한된다. WB 신호는 50~7,000Hz 주파수 대역의 음성 신호를 의미하며, 모바일 통신에서는 VoLTE 또는 HD Voice 라 불린다. WB 신호는 대면 의사소통이나 전문 스튜디오 녹음 품질과 비교하면 다소 먹먹하게 느껴진다. [1] 음성의 품질 저하를 최소화하며 전송하기 위한 고효율 코덱들이 등장하고 있지만, 네트워크의 불균일성 및 송수신 기기에 탑재된 코덱의 호환 여부 등으로 인해 실생활에 적용되기는 제약사항이 많다. 모든 네트워크와

송수신 기기가 고효율 코덱을 지원하기까지는 오랜 시간과 비용이 소요되기 때문에 종단 기기에서 음성의 명료도를 높일 수 있는 음성 대역폭 확장 기술이 필요하다.

음성 대역폭 확장(BWE, Bandwidth Extension)은 낮은 주파수(LF, Low Frequency) 부분의 신호를 이용해 높은 주파수(HF, High Frequency) 부분의 신호를 다시 생성하여 음성 신호의 품질을 향상하는 기법이다. [2] 본 연구에서는 Hifi-GAN 을 기반으로 WB 음성 신호의 멜 스펙트로그램(Mel Spectrogram)을 이용하여 SWB 음성 신호를 합성하는 BWE 딥러닝 모델을 제안한다. 실험을 통해 제안하는 BWE 기술이 화자에 의존하지 않고 효과적으로 음질 향상을 이뤄낼 수 있으며, 정성적으로 성능 평가를 수행한 결과 제안된 BWE 기술을 적용한 음성의 경우 입력 음성에 비해 30%가량의 MUSHRA 점수 향상이 있음을 확인하였다.

# 2 Background

## 2.1 Bandwidth Extension

주파수 대역의 제한은 음질 저하의 주요한 원인으로 작용하기 때문에 다양한 인공 음성 대역 확장(ABWE, Artificial Bandwidth Extension) 기술이 제안되어 왔다. 기존의 ABWE 연구들은 대부분 협대역(NB, Narrow Band)의 음성을 광대역으로 확장하는 것을 목표로 하고 있다. 음성의 중요한 요소들이 4,000~8,000Hz 주파수 대역에 존재하기 때문에 ABWE 는 눈에 띄는 음질의 향상을 보여준다. [3] 초광대역 확장(SWBE, Super-wide Bandwidth Extension)은 WB 신호에 존재하지 않는 8,000~16,000Hz 주파수 대역을 복원하는 것을 목표로 한다.

16kHz 의 샘플링 레이트의 음성은 나이퀴스트 이론(Nyquist Theorem)에 따라 8,000Hz 의 주파수 대역까지 표현할 수 있으며, 이는 음성 콘텐츠를 대부분 포함하면서 음성 처리 과정이 그렇게 무겁지 않다. 즉, 명료성과 계산 비용 사이의 절충안을 이루는 "최적 지점 (sweet spot)"이라 여겨진다. 그러나 음성의 사실성이 부족하므로 일부 사용자들은 만족스럽지 못한 청취를 경험한다. [4] 더 넓은 주파수 대역을 처리하기 위해서는 자원과 비용이 충분하지 않기 때문에 WB 음성 신호를 SWB 음성으로 확장하기 위한 SWBE 기술이 필수적이다.

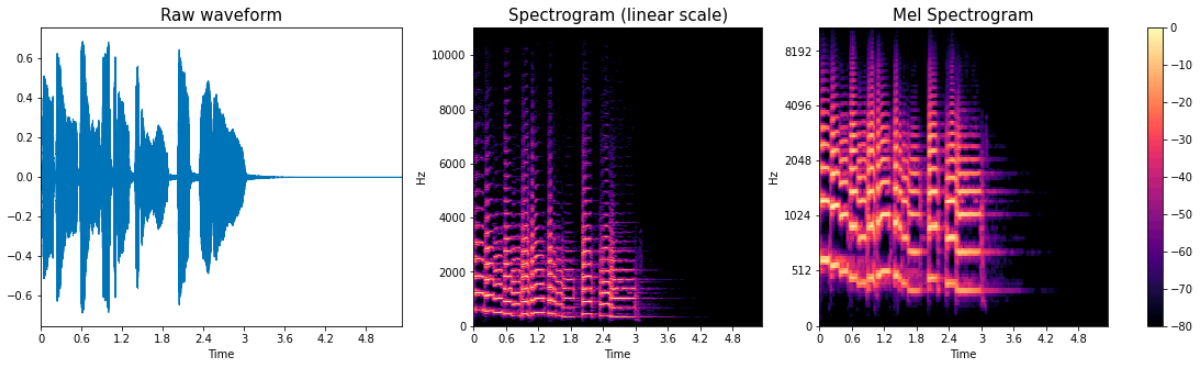
SWBE 와 관련된 연구는 많지 않다. 이는 SWBE 가 AWBE 에 비해 느껴지는 음질의 향상 폭이 크지 않으며, WB 신호도 음성 콘텐츠를 모두 담기에는 충분하므로 오히려 대역폭 확장 과정에서 발생하는 아티팩트들이 음질 저하 요인으로 작용할 수 있어서 더 도전적인 과제이다. [3]

SWBE 와 관련하여 다음 연구들이 진행되었다. [5] 에서는 현존하는 주파수 요소들에 피치 스케일링(pitch scaling)을 적용하여 누락된 주파수 대역을 복원하였으며, [3] 에서는 WB 신호에 선형 예측 기법을 적용하여 주파수 응답이 평탄한 소스(residual information)와 포락선(envelope)을 얻어내는 전통적인 소스 필터(source filter) 모델을 이용한다. 최근에는 DNN 을 도입한 연구들이 이뤄지고 있다. [6] 에서는 DNN 과 ConvNet 을 사용하여 EVS(Enhanced Voice Services) 코덱을 거친 WB 음성 신호를 SWB 음성 신호로 복원하는 모델을 제안하였다. 더 나아가, [4] 에서는 feed-forward WaveNet 구조와 생성적 적대 신경망(GAN, Generative Adversarial Network)을 기반으로 샘플링 레이트가 16kHz 인 음성 신호를 샘플링 레이트가 48kHz 인 하이파이(Hi-Fi, High Fidelity) 음성 신호로 확장하는 연구가 이뤄졌다.

## 2.2 TTS Pipeline and Hifi-GAN

본 연구는 음성 합성(TTS, Text-to-Speech) 분야에서 사용되는 방법론을 답습하였다. TTS 는 일반적으로 어쿠스틱 모델(Acoustic Model)과 보코더(Vocoder)라고 불리는 두 개의 단계로 구성되어 있다. 첫 번째 단계는 텍스트로부터 멜 스펙트로그램과 같은 저해상도의 중간 표현단계를 예측하는 것이고, 두 번째 단계는 중간 표현 단계로부터 원시 오디오 파형(raw waveform audio)을 합성하는 것이다. [7] 본 연구에서는 두 번째 보코더 단계에 사용되는 모델을 활용하여 BWE 를 구현하고자 한다.

소리의 특성을 제대로 파악하기 위해서는 시간, 소리의 세기 및 높낮이에 대한 정보가 모두 필요하기 때문에 국소 푸리에 변환(STFT, Short-Time Fourier Transform)을 적용하여 **Figure 2** 의 (b) 와 같이 스펙트로그램으로 나타낸다. 스펙트로그램에서 x 축은 프레임(시간)을, y 축은 주파수를, 색은 주어진 프레임에서 해당 주파수의 크기를 의미한다. 인간이 낮은 주파수 대역의 차이를 높은 주파수 대역보다 민감하게 반응하는 특성을 반영하기 위해 주파수에 Mel-Scale 을 적용하여 **Figure 2** 의 (c) 와 같이 멜 스펙트로그램이 많이 사용된다. [8]




(a) (b) (c)

**Figure 2** (a)는 원시 오디오 파형을, (b)는 STFT를 적용하여 나온 Spectrogram, (c)는 Mel Spectrogram의 예시이다.

GAN 은 데이터의 분포를 추론하고 생성하는 Generator 와 생성된 데이터가 학습 데이터일 확률을 예측하는 Discriminator 로 구성되어 있다. Generator 는 생성된 데이터가 Discriminator 를 속일 확률이 최대가 되도록, Discriminator 는 진위를 잘 구분할 수 있도록 두 행위자 최소 최대 게임(two-player minimax game) 전략을 사용하여 학습이 진행된다. [9] WaveGAN, GAN-TTS, MelGAN, Parallel WaveGAN, Hifi-GAN 등 많은 보코더들이 GAN 을 이용하여 뛰어난 성능의 음성 합성을 이루어 냈다. [10]

Hifi-GAN 은 현재 보코더 분야에서 SOTA(State of art)를 달성한 모델로, **Figure 3** 과 같이 한 개의 Generator 와 두 개의 Discriminator 로 이루어져 있다. Generator 는 Mel-spectrogram 을 입력으로 받아서 원하는 출력의 길이가 나올 때까지 Transposed Convolution 연산을 거치며 업샘플링(upsampling) 한다. 이때 MRF(Multi-Receptive Field Fusion) 모듈을 통해 병렬로 다양한 길이의 패턴을 파악한다. 사실적인 음성을 재현하기 위해서는 음성의 주기적 특성과 장기 의존성(Long-term Dependency) 문제를 해결해야 한다. 이를 위해 MPD(Multi-Period Discriminator)는 음성 신호에서의 다양한 주기적인 패턴을, MSD(Multi-Scale Discriminator)는 음성 신호에서의 연속적인 패턴을 관찰하여 음성 신호의 진위를 판별한다. [7]

## 3 GAN based Bandwidth Extension

본 절에서는 BWE 기술 구현을 위해 데이터 처리 방법 및 모델 훈련 과정을 설명한다. 이미지 생성에서는 하나의 화소(pixel)에 오류가 생겨도 이미지를 인식하는 데 큰

어려움으로 작용하지 않는다. 그러나 주파수 도메인은 로그 스케일이기 때문에 스펙트로그램 상에서의 작은 선형 변환일지라도 청취에 있어서 심각한 왜곡을 일으킨다. 또한, 시간 도메인에서의 변화는 지직거리는 소리나 아티팩트로 들릴 수 있다. 따라서 소리를 생성할 때는 높은 수준의 세부사항(high level of a detail)을 보존하는 것이 중요하다. [11] Hifi-GAN 은 원하는 길이의 출력이 나올 때까지 계층적으로 음성 신호의 패턴을 생성하기 때문에 BWE 에 효과적으로 적용할 수 있다.

## 3.1 Architecture

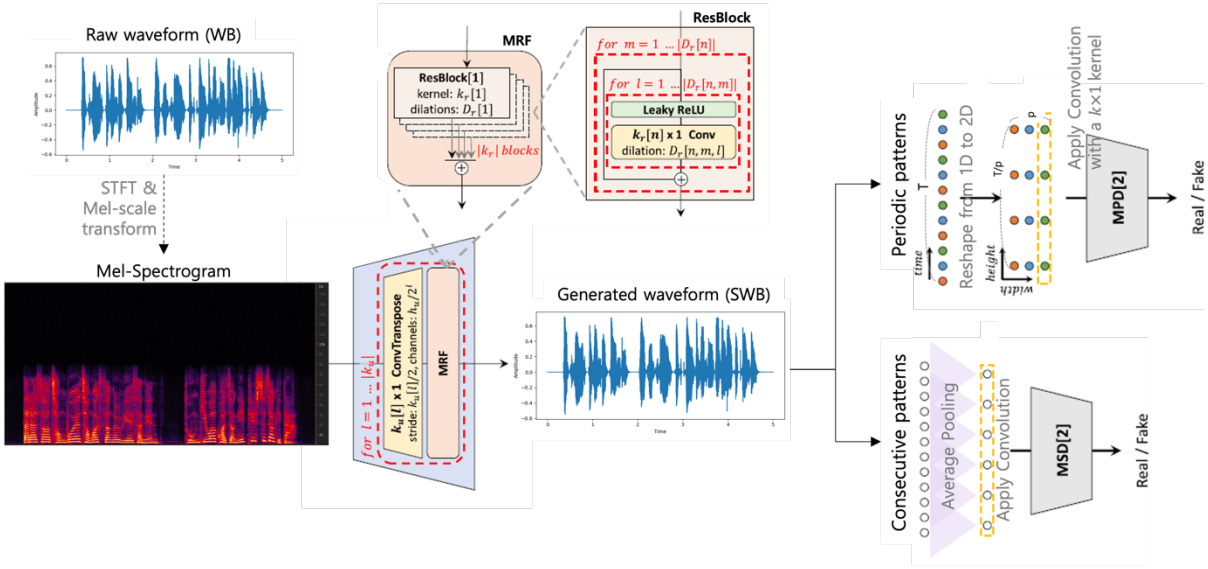



**Figure 3** Hifi-GAN 기반 WB-to-SWB BWE 모델 모식도

본 연구에서 제안하는 BWE 모델의 구조는 **Figure 3** 과 같다. 먼저, WB 음성 신호를 멜 스펙트로그램으로 변환한다. Generator 는 멜 스펙트로그램으로부터 SWB 음성 신호를 생성하며, MPD 와 MSD 는 생성된 신호의 진위를 판별하도록 훈련된다. 멜 스펙트로그램으로부터 음성 신호를 생성하는 Generator 의 역할은 TTS 모델과 큰 차이가 없기 때문에 사전 학습된 모델을 사용하여 전이 학습(Transfer Learning)을 통해 BWE 에 사용할 수 있다.

그러나 기존의 Generator 는 입력 음성 신호에 STFT 를 적용할 때 사용한 window 에 해당하는 길이의 음성 샘플을 생성하지만, BWE 에 사용되는 Generator 는 출력 음성 신호의 샘플링 레이트가 증가하기 때문에 입력 신호와 출력 신호에서 같은 window 에 해당하는 샘플의 수가 달라지는 문제가 발생한다. 이는 Generator 로부터 나오는 출력

음성 신호의 샘플링 레이트가 32kHz 가 되도록 모델 구조를 수정하거나 입력 음성 신호를 스펙트로그램으로 변환할 때 사용하는 window 사이즈를 조절하여 해결할 수 있다. 본 연구에서는 여건상 모델의 구조 수정이 힘들어 출력 음성 신호의 샘플링 레이트가 44kHz 인 사전 학습된 모델*을 전이 학습하는 후자의 방법을 택하였다.

모델의 하이퍼 파라미터를 조절하여 생성된 음성의 품질과 생성 속도 사이의 절충안을 찾아야 한다. 앞서 언급한 사전 학습된 모델에서는 $h_u = 512, k_u = [16,16,4,4], k_r = [3,7,11], D_r = [[1,3,5]] \times 3,\ MPD = [3,5,7,11,17,23,37]$와 같은 하이퍼파라미터를 사용하였다. 입력으로는 80 차 멜 스펙트로그램을 사용하였다. FFT 는 2048, window 는 2048, hop 은 512 를 사용하였다. AdamW ( $\beta\_1 = 0.8, \beta\_2 = 0.99, \lambda = 0.01, lr = 2 \times 10^{-4}, lr\ decay = 0.999$)로 학습하였다. 모델의 파라미터 수는 약 13.96M 개이며, 하나의 NVIDIA M40 GPU 로 약 9 일간 135k step 정도 학습을 진행하였다.

## 3.2 Dataset

실험에 사용한 데이터셋은 14 명의 남성, 13 명의 여성 화자로 이루어진 16,290 개의 짧은 음성 클립으로 구성되어 있으며, 총 길이는 약 26 시간이다. 오디오 포맷은 32-bit float 형식의 wav 파일이다. 입력 데이터로 사용하기 위해 샘플링 레이트 16kHz 로 다운샘플링(downsampling)한 후 FFT, window, hop 은 각각 2048, 743, 162 를 사용하여 80 차 멜 스펙트로그램을 만들었다. 생성된 음성 신호는 사전 학습된 모델의 설정에 따라 44kHz 의 샘플링 레이트를 가지기 때문에 본 연구의 목적에 맞춰 16-bit int 형식의 32kHz 로 다운샘플링한다. 또한, 코덱 아티팩트 복원 가능성을 확인하고자 mSBC, EVS, mSBC 코덱의 부호화·복호화 과정을 차례로 거친 음성 파일에 대해서도 위와 같은 작업을 수행하였다. 전체 데이터셋 중 90%는 실제 훈련에 사용하였으며, 나머지 10%는 검증 데이터셋으로 사용하였다.

* https://drive.google.com/drive/folders/1sv7AQjyR_LQn-s128THjvy9GTgU-fegS

# 4 Results

## 4.1 Graphical Comparison

Figure 4 는 모델의 성능을 평가하기 위해 코덱의 부호화·복호화 과정을 거쳤는지 여부, 학습 데이터셋에 화자의 존재 여부에 따라 총 네 가지 샘플에 대해 본 연구에서 제안하는 방법을 적용한 후 각각의 음성 신호에 대한 스펙트로그램을 나열한 것이다.

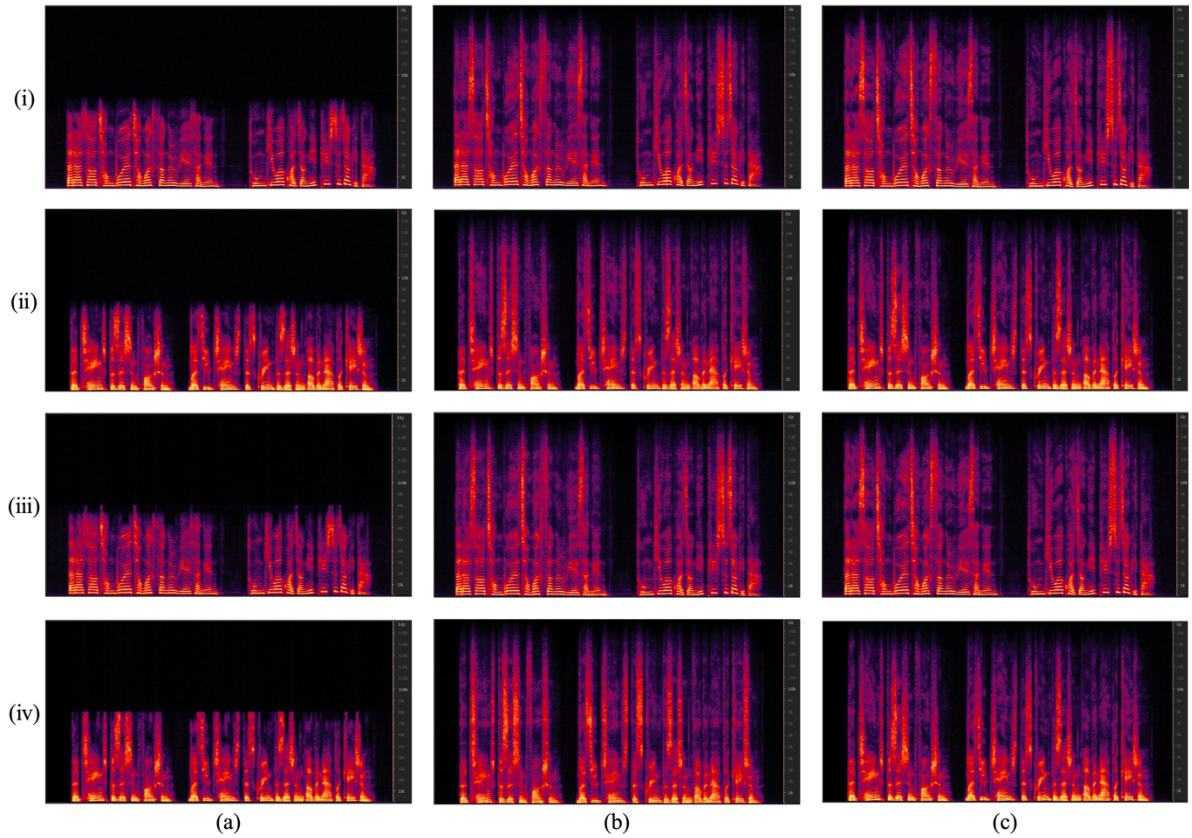



**Figure 4** 각 행은 차례로 (ⅰ) BWE with seen speaker (ⅱ) BWE with unseen speaker (ⅲ) Multiple codec artifacts BWE with seen speaker (ⅳ) Multiple codec artifacts with unseen speaker에 해당하는 음성 신호의 스펙트로그램이다. 각 열은 차례로 (a) 입력으로 들어가는 샘플링 레이트 16kHz인 음성 신호 (b) 훈련된 모델로 생성한 결과 (c) Ground-Truth인 샘플링 레이트 32kHz인 음성 신호를 의미한다. 각 스펙트로그램의 y축은 0~16kHz 범위를 선형 스케일로 나타낸다.

첫 번째 열은 입력 음성 신호의 스펙트로그램으로, 샘플링 레이트가 16kHz 인 WB 신호이기 때문에 주파수가 8kHz 이상인 음성은 표현하지 못한다. 고주파수 대역이 존재하지 않기 때문에 음성이 담고 있는 내용적인 면에서의 손실은 없지만, 명료도가 떨어지다. 가운데 열의 생성된 음성은 눈으로 보아도 마지막 열의 Ground-Truth 와 상당히 유사한 스펙트로그램을 가지며, 청취 결과 명료도와 공간감이 향상된 것을

확인할 수 있다. 또한, **Figure 4** 의 (ii) 와 같이 학습 데이터셋에 존재하지 않는 화자에 대해서도 뛰어난 일반화 성능을 보여준다.

여러 번의 코덱을 거치면서 생긴 아티팩트들이 입력 음성에 존재하는 **Figure 4** 의 (iii) 와 (iv) 의 경우에도 이전과 비교하면 잡음이 간혹 들리는 등 성능이 소폭 하락했지만 양호한 성능을 보여준다.

## 4.2 Subjective Evaluation

BWE 의 정성적 성능 평가를 위해 학습 데이터셋 안에 존재하지 않는 데이터 중에 **Table 1** 과 같이 무작위로 7 개의 샘플을 추출하였다. 5 명의 화자는 학습 데이터셋에 존재하는 화자이고, 2 명의 화자는 학습 데이터셋에 존재하지 않는 화자로 모델의 일반화 성능을 확인하고자 하였다. 각각의 샘플에 대해 **Table 2** 와 같은 처리를 하여 MUltiple Stimuli with Hidden Reference and Anchor(MUSHRA) 테스트를 진행하였다.

**Table 1** MUSHRA 테스트용 데이터셋 정보

| No | Type | Speaker |
|---|---|---|
| 1 | Seen | Korean Male |
| 2 | Seen | English Female |
| 3 | Seen | English Male |
| 4 | Seen | Korean Male |
| 5 | Seen | Korean Female |
| 6 | Unseen | English Female |
| 7 | Unseen | English Male |

**Table 2** 데이터 처리 방법

| Type | Bandwidth |
|---|---|
| Band-limited | 8kHz (WB) |
| Extend Bandwidth | 16kHz (SWB) |
| Pass through codec | 8kHz (WB) |
| Restore codec loss | 16kHz (SWB) |
| w/o Fine-tuning (Anchor) | 16kHz (SWB) |
| Hidden Reference | 16kHz (SWB) |

MUSHRA 테스트는 음성의 품질을 평가하기 위한 방법론이다. 실험 참여자들은 라벨링이 가려진 음성을 들으면서 Reference 와 동일한 Hidden Reference 를 찾아서 100 점을, 가장 음질이 나쁜 음성인 Anchor 를 찾아 50 점을 부여한다. 이 두 가지 음성을 기준으로 나머지 음성들의 점수를 평가한다. 본 연구에서는 Anchor 를 정확히 찾은 5 명의 참가자를 대상으로 평가를 진행하였다.

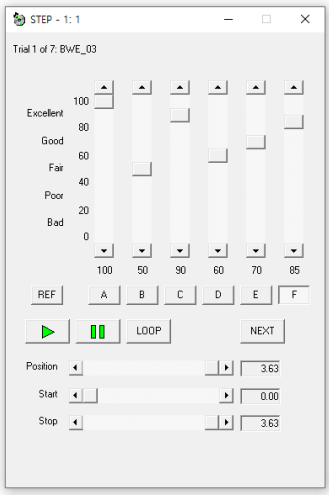



Figure 5 MUSHRA 테스트 화면

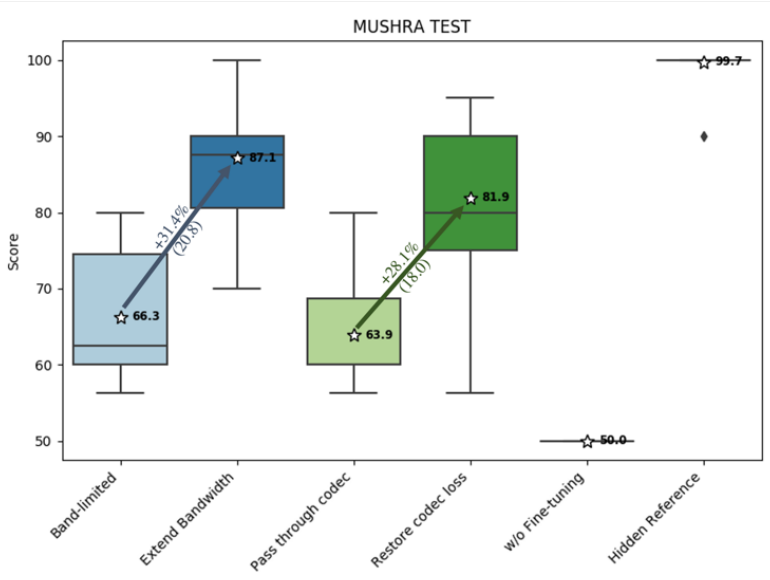



Figure 6 MUSHRA 테스트의 결과를 상자 그림으로 시각화하였다. 별은 각 데이터셋의 MUSHRA 점수 평균이다.

Table 3 MUSHRA 테스트의 점수 및 순위 결과

| Type | Size | Score | | Rank | |
|---|---|---|---|---|---|
| | | Mean | Std | Mean | Std |
| Band-limited | 35 | 66.311 | 7.578 | 4.329 | 0.382 |
| **Extend Bandwidth** | 35 | **87.143** | 6.957 | **2.243** | 0.520 |
| Pass through codec | 35 | 63.911 | 6.293 | 4.586 | 0.477 |
| Restore codec loss | 35 | 81.893 | 9.317 | 2.786 | 0.610 |
| w/o Fine-tuning (Anchor) | 35 | 50 | 0 | 6 | 0 |
| Hidden Reference | 35 | 99.714 | 1.690 | 1.057 | 0.202 |

**Figure 6** 과 **Table 3** 은 MUSHRA 테스트 결과이다. 코덱 아티팩트가 있는 경우와 없는 경우 모두 약 30%의 점수 향상을 보였다. 코덱 아티팩트를 복원하는 모델의 경우 입력에 노이즈가 있으며, 학습 시간도 짧았기 때문에 성능이 다소 낮은 것으로 보인다. 그럼에도 불구하고 청취 결과 큰 점수 차이가 나지 않는 것으로 보아 같은 방법론을 통해 충분한 데이터셋과 학습 시간을 가지면 코덱 아티팩트를 효과적으로 개선하는 모델을 얻어낼 수 있을 것이라 기대한다.

# 5 Conclusion

본 연구에서는 대역이 제한되거나 네트워크 송수신 과정에서 다양한 코덱을 거치면서 음질 저하가 발생한 음성 신호를 복원 및 확장하기 위해 BWE 딥러닝 모델을 구현하고 성능을 평가하였다. WB 음성 신호의 멜 스펙트로그램을 입력으로 받아 SWB 원시 오디오 파형을 합성하기 위해 TTS 보코더 중 SOTA 로 알려진 Hifi-GAN 의 사전학습된 모델을 이용하여 전이 학습하였다. 그 결과 효율적으로 대역폭 확장을 이뤄낼 수 있으며, 학습 데이터셋에 존재하지 않는 화자에 대해서도 양호한 성능을 보여주어 일반화가 가능함을 확인하였다. GAN 을 기반으로 하는 BWE 는 코덱 아티팩트가 존재하는 입력 음성 신호에 대해서도 상대적으로 짧은 학습 시간에 비해 뛰어난 성능을 보여주었다. 그러나 생성 속도가 느리다는 점이 한계로 여겨진다. 현재 CPU 에서의 초당 0.38 초(16.73k sample)의 음성이 생성되기 때문에 실시간으로 BWE 를 적용하기는 어려움이 있다. 이는 사전 학습된 모델을 사용하면서 실제 출력 음성 신호보다 더 많은 샘플을 생성해야 하며, 생성된 음질을 높이는 데 치중한 하이퍼파라미터 세팅을 사용하였기 때문이다. 또한, 일반적으로 음성 통화는 지연에 민감하기 때문에 end-to-end delay 를 300ms 이하로 유지하기 위한 causal 한 신호처리가 매우 중요하다. 즉, 미래 신호를 최대한 적기 사용해야 하기 때문에 한번에 처리하는 프레임의 크기를 줄이는 것이 중요하다. 따라서 추후에 이러한 부분을 고려한 개선안을 통해 향후 다양한 디바이스에 GAN 기반의 BWE 가 적용되기를 기대한다.

# 참 고 문 헌


[1] R. V. Cox, S. F. De Campos Neto, C. Lamblin and M. H. Sherif, "ITU-T coders for wideband, superwideband, and fullband speech communication [Series Editorial]", IEEE Communications Magazine, vol. 47, no. 10, pp. 106-109, 2009.

[2] 심규홍, 이민재, 성원용, "깊은 재귀형 신경망을 이용한 음성 대역폭 확장", 한국통신학회 학술대회논문집, pp. 961-962, 2017.

[3] Bachhav, Pramod, Massimiliano Todisco, and Nicholas Evans, "Efficient super-wide bandwidth extension using linear prediction based analysis-synthesis", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5429-5433, 2018.

[4] J. Su, Y. Wang, A. Finkelstein and Z. Jin, "Bandwidth Extension is All You Need", Speech and Signal Processing (ICASSP), pp. 696-700, 2021.

[5] Beiser, B., & Vary, P, "Artificial bandwidth extension of wideband speech by pitch-scaling of higher frequencies", INFORMATIK 2013-Informatik angepasst an Mensch, Organisation und Umwelt, pp. 2892-2901, 2013.

[6] Abel, Johannes, Ernst Seidel, and Tim Fingscheidt, "Enhancing the EVS Codec in Wideband Mode by Blind Artificial Bandwidth Extension to Superwideband", 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 281-285, 2018.

[7] Kong, J., Kim, J., & Bae, J, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis", Advances in Neural Information Processing Systems, 2020.

[8] 문지영, 고형석, "DeepASMR: 딥러닝 기반의 ASMR 플랫폼", 한국정보과학회 학술발표논문, pp. 494-496, 2020.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets", Advances in neural information processing systems, pp. 2672-2680, 2014.

[10] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, “A survey on neural speech synthesis”, arXiv preprint arXiv:2106.15561, 2021.

[11] Andreas Jansson, Eric J. Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyd, "Singing voice separation with deep U-Net convolutional networks", Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pp. 323-332, 2017.

## Abstract

# Speech Bandwidth Extension with Hifi-GAN

Jinkyoung Hwangbo

Department of Industrial Engineering

College of Engineering

Seoul National University

As the number of phone calls using wireless earbuds has increased, recent research has been actively conducted to improve speech quality. The quality of speech is degraded because of bandwidth limitation and multiple codec artifacts when transmitting through heterogeneous networks. This paper proposes a GAN-based bandwidth extension deep learning model, which can extend the bandwidth and clarify speech on the device stage. Wideband speech signals can be extended to super wideband signals by a proposed model based on Hifi-GAN, known as SOTA in the TTS vocoder fields. MUSHURA evaluation shows that the SWB speech generated by proposed model gets about 30% higher score than input WB speech regardless of the speaker. Furthermore, the proposed model has the ability to extend bandwidth even though the input speech has multiple codec artifacts.

**Keywords: Bandwidth Extension, Deep learning, GAN, Restoring tandem codec artifacts**