공 학 석 사 학 위 논 문

높은 소음 환경에서 강건한 음성 인식을 위한 ASR 시스템 개발

2024 년 2 월

국 립 부 경 대 학 교 대 학 원

산업및데이터공학과

김 민 성

공 학 석 사 학 위 논 문

높은 소음 환경에서 강건한 음성 인식을 위한 ASR 시스템 개발

지도교수 최 성 철

이 논문을 석사 학위논문으로 제출함.

2024년 2월

국 립 부 경 대 학 교 대 학 원

산업및데이터공학과

김 민 성

김민성의 공학석사 학위논문을 인준함.

2024 년 2 월 16 일

위 원 장 공학박사 서 원 철 (인)

위 원 공학박사 유 태 선 (인)

위 원 공학박사 최 성 철 (인)

목 차

표 목차

그림 목차

Enhancing Robust Speech Recognition in High-Noise Environments with ASR System Development

Min Sung Kim

Deparment of Industrial and Data Engineering,
The Graduate School, Pukyoung University

Abstract

Speech recognition is a task that takes speech as input and outputs text. Speech recognition is used in various places in the real world, such as AI speakers, voice memos, and more, and there is room for further development. For this purpose, various studies based on deep learning are being conducted. However, these studies are trained with data containing relatively quiet noises rather than real-world environments with strong noises including ambient sounds or noisy situations, resulting in poor recognition rates in noisy environments. We improved the performance of the Whisper model by using a pre-trained model that removes the noise through a process of compressive restoration, and showed that the method of learning by synthesizing noisy speech with denoised speech shows better performance than the model that learned only noisy speech, showing that the proposed method is effective for DNN-based speech models, and proposes a framework for creating noise robust speech recognition models.

Ⅰ. 서론

1. 연구 배경

음성인식은 사람의 목소리가 녹음된 음성에서 발화에 해당하는 텍스트를 전사하는 작업이다. 음성인식은 이미 다양한 형태로 우리의 일상속에서 사용되고 있다. 우리의 일상 생활에서 가장 많이 사용하는 기기 중 하나인 스마트폰은 인공지능 개인 비서라는 이름으로 음성인식이 적용된 대표적인 사례이다. 그 외에 인공지능 스피커, 커넥티드 카 등 많은 기기에 음성인식이 적용되어 있으며 CLOVA note와 같이 서비스로 제공되어 사용되고 있다.

우리 일상의 음성인식 기술은 현재까지 꾸준한 발전을 거듭하고 있다. 초기 음성인식 기술은 군사용으로 개발되어 사용되었고 낱말 단위의 표현 정도만 가능하였다. 1980 년대 Hidden Markov Model(HMM)의 등장 이후 Hidden Markov Model-Gaussian Mixed Model (HMM-GMM)을 이용한 음성인식이 주를 이뤘다. Hidden Markov Model(HMM)은 현재 상태(state)는 이전 상태(state)에만 영향을 받는다는 가정인 Markov process 를 기반으로한 통계적 모델이다. 음성인식에서의 HMM 은 음성신호가 Markov Model 에 의해 발생하였다는 가정하고 학습 단계에서 모델의 파라미터를 추정하며 이를 바탕으로 미지의 음성에 가장 적합한 단어나 음소를 찾는 것이다.

HMM-GMM 기반의 음성인식 모델은 크게 Acoustic Model 과 Language Model 으로 구성되며 입력과 출력은 각각 음성(wavefo

rm)과 word sequence 이다[1]. 음성인식 모델은 입력 음성 신호에 대해 가장 높은 확률을 갖는 음소의 시퀀스를 추정하는 문제이다. 이때 입력인 발화된 문장은 단어로 구성되며 음소라는 발음의 단위를 갖는다. 이 음소에서 단어를 추정하는 것은 Acoustic Model 을 통해 모델링하며 문장을 이루는 단어를 모델링하는 것은 Language Model 을 통하여 진행된다. 음소를 추정하기 위한 Acoustic Model 에서의 HMM(Hidden Markov Model)은 한 단어를 구성하는 음소가 주어졌을 때 각 음소에 해당하는 글자를 관측할 확률을 계산하며 HMM 의 state 에 음소가 주어질 때 특정 acoustic feature 가 관측될 확률을 GMM 을 통해 모델링한다. 음성인식에서는 하나의 모델을 통해 모든 발화에 대한 성질을 반영하기 어렵다는 단점이 있다. 따라서 여러 HMM 을 사용하여 발화를 모델링하지만 데이터 내의 새로운 단어나 낮은 빈도를 가지는 발화에 대해서는 HMM 이 인식하지 못하거나 잘 학습되지 않는 단점이 있다.

2000 년대 인공신경망에 관련된 연구가 있었지만 인공신경만의 깊이와 형태로 인해 여전히 GMM-HMM 기반의 음성인식보다 성능이 떨어졌다. 하지만 Graphic Processing Unit(GPU)와 같은 하드웨어의 등장으로 컴퓨팅 성능의 향상되어 더 깊어진 신경망의 연산을 가능하게 하였다. 더 깊어진 신경망을 통해 음성인식에도 더욱 깊어진 신경망이 적용되어 현재까지도 활발히 연구되고 있다.

현재까지 Deep Learning 은 다양한 분야에서 활발히 연구되고 있고 음성인식에 Deep Learning 을 적용한 많은 연구들이 이어지고 있다. 특히 Deep Learning 에서 음성인식은 Automatic Speech Recognition (ASR), Speech-To-Text(STT) 라는 이름을 가진 문제로 연구들이 수행되고 있다. 음성인식의 목표는 사람의 목소리, 음성을 정확하게 텍스트로 변환하는 것이다. 정확한 음성인식은 시스템과 사람 간의 상호작용을 위한 필수적인 시스템이 되었다. 이러한 음성인식이 포함된 시스템을 잘 활용하기 위해서는 음성인식 모델을 설계시 실제 사용되는 여러 상황들과 다양한 환경을 고려하여 설계되어야 한다.

2. 연구 목표 및 내용

음성인식 모델은 음성인식 서비스가 사용되는 실제 환경을 고려하여 다양한 잡음에 대해 정확한 전사가 가능해야 한다. 하지만 지금까지 연구되어 온 음성인식 모델들은 현실의 모든 상황을 반영하지 못했으며 소음이 없는 환경을 가정하여 비교적 소음의 세기가 작은 데이터로 학습이 되어 소음에 대해 강건하지 못하다. 본 연구에서는 주변 소음을 포함한 다양한 종류의 소음과 세기에 강건한 모델을 만들기 위한 방법을 제안한다. 소음에 강건한 모델을 구축하기 위한 음성인식 모델의 학습용 데이터셋으로는 AI HUB 의 극한 소음 음성인식 데이터를 사용하였다. 극한 소음 음성인식 데이터는 음성인식 성능 개선을 위해 다양한 극한 소음 환경에서 발화된 음성데이터가 수집, 정제, 가공된 학습용 데이터셋이다. 본 연구에서는 극한 소음이 포함된 음성 데이터(original data)에서 Speech Enhancement 테스크를 수행하여 소음을 제거한다. 소음이 제거된 음성(denoised audio)을 소음이 포함된 original data 와 합하여 새롭게 생성된 음성(mixed audio)을 모델의 학습에 사용한다. 소음을 포함한 원본 데이터와 소음이 제거된 데이터를 이용하여 간단하지만 효과적인 합성 기술을 통해 새롭게 생성된 mixed audio 는 소음이 제거된 음성 보다 더 자연스럽고 사람의 발화에 집중된 데이터를 만들고 이를 이용하여 음성인식 모델을 fine-tuning 하여 소음에 보다 더 강건한 모델을 만든다. 강건한 모델 구축을 위해 mixed audio 를 학습하는 방법의 타당성을 검증하기 위해 음성인식 모델을 original data,

denoised data, mixed data에 대해 각각 fine-tuning시키고 학습된 모델을 original data, denoised data, mixed data에 대해 각각 평가를 진행하여 성능을 측정한다. 이 실험을 통해 소음이 제거된 음성은 부자연스럽고 학습 데이터로 사용했을 때 모델의 성능을 떨어뜨린다는 것을 증명한다. 또한 원본 데이터와 합성된 데이터로 각각 학습된 모델의 성능을 평가하여 단순하지만 효과적인 합성 기술이 딥러닝 기반의 음성인식 모델의 성능을 향상시키며 소음에 강건한 모델을 구축할 수 있다는 것을 보여준다.

Ⅱ. 선행 연구

2.1 Speech Enhancement

음성인식 시스템은 크게 Front-End, Back-End 로 나뉜다. Front-End 는 음성인식 시스템에서의 입력인 음성을 텍스트로 전사하는 모델에 들어가기 전 입력 오디오의 품질을 향상시키는 Speech Enhancement(SE), 여러 신호가 섞인 음성에서 각각의 source 를 분리하는 작업인 Source Separation(SS)등을 수행한다. Back-End 는 실질적인 음성을 텍스트로 전사하는 작업인 Speech Recognition 을 수행하는 부분이다.

음성 향상(Speech Enhancement)은 음성인식 시스템에서 중요한 역할을 수행하는 모듈이다. SE 는 음성인식 시스템에서 Back-End 의 실질적인 음성인식을 담당하는 모델의 입력을 잡음을 제거하여 더 좋은 품질의 입력을 제공하는 것을 목표로 한다. 이러한 과정에서 degraded speech signal 의 품질의 향상, noise reduction, noise 를 제거하는 작업을 진행한다. SE 모듈을 통해 음성인식 시스템의 Back-End 에 더 좋은 데이터를 제공하여 음성인식 성능을 향상시킬 수 있다.

음성인식의 Front-End 에 해당하는 모듈 역시 Deep Neural Network(DNN)의 발전과 더불어 DNN 기반의 음성 향상 모델과 음성 향성 성능 역시 발전하였다. Speech Enhancement 의 대표적인 모델 중 하나인 Conv-TasNet[2]는 Encoder, Decoder 구조를 가진 Speech Source Separation 문제를 위해 제안된 모델이다. Demucs[3]는 Conv-TasNet 을 speech 가 아닌 음악에 적용하여 여러 종류의 악기가 섞여 있는 오디오에서 특정 악기의 소리만을 추출하는 music source separation 을 위해 제안된 모델이다. U-Net 구조를 차용하여 Encoder(downsampling block), Decoder(upsampling block)가 쌓인 구조이며 Encoder, Decoder 의 같은 인덱스를 가지는 블록들은 skip-connection[4]으로 연결되어 있다.

2.1.1 Denoiser

Denoiser 는 Music Source Separation 문제를 해결하기 위해 제안된 Demucs 모델 구조를 이용하여 다양한 소음을 제거하기 위해 제안된 모델이다. Laptop CPU 에서도 실시간 음성 향성 처리가 가능하도록 상대적으로 낮은 연산량을 목표로 개발되었다[5].

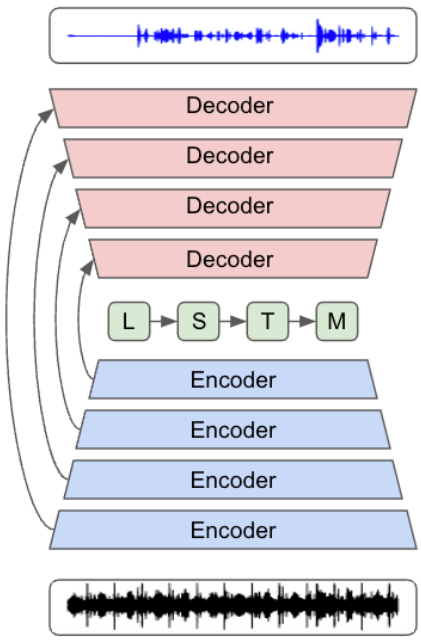

Figure 1. Denoiser 구조

2.1.2 Observation Adding

음성인식 시스템의 Front-End 에서는 음성 데이터에 포함된 소음을 제거하는 등의 음성 데이터의 향상에 대해 초점이 맞춰져 있다. 하지만 향상된 음성이 실제 음성인식 모델에 미치는 영향에 대한 분석은 잘 이루어 지지 않았다. Iwamoto 등(2022)는 소음이 제거된 음성과 실제 소음이 선형 결합(linear combination)으로 표현될 수 없는 부자연스러운 표현을 갖는 artifact 라고 주장하며 Speech Enhancement 를 통해 향상된 음성을 음성인식 모델의 학습에 사용하면 성능의 저하를 초래할 수 있다고 주장하며 이를 극복하기 위해 원본 음성과 SE 를 통해 소음이 제거된 향상된 음성을 결합한 단순하지만 효과적인 방법인 Observation Adding(OA) 제안하였다[6]. 해당 논문에서는 Speech Enhancement 를 수행하는 Front-end 에서 Denoising-TasNet[7]을 사용하였으며 실질적인 음성인식 테스크를 수행하는 Back-End 에서는 모델 전체가 DNN 으로 구성된 모델이 아닌 Deep Neural Network-Hidden Markov Model(DNN-HMM) hybrid 모델을 사용하였다.

2.2 Speech Recognition

음성인식 시스템에서 Back-End는 실제 음성을 텍스트로 전사하기 위해 존재하는 모듈이다. 연산 처리 장치의 성능이 낮았던 과거에는 GMM-HMM의 기반의 음성인식 모델들이 많이 연구되었다. GMM-HMM 기반의 하이브리드 모델은 일반적으로 acoustic model과 language model으로 구성되어 있다. Acoustic model은 음소의 나열을 예측하고 language model은 단어를 통해 적절한 문장을 예측하는 역할을 수행한다. Deep Neural Network의 발전으로 과거와 달리 하나의 모델을 통해 음성을 텍스트로 연결하도록 훈련하는 End-to-End(E2E) 음성인식이 많이 연구되고 있다. E2E 모델은 음성 벡터를 심층 신경망을 통과시켜 텍스트를 예측하도록 한다. E2E 음성인식 모델은 온전히 데이터에만 의존하여 학습하며 모델을 학습하기 위한 음성 데이터에서 별도의 프레임별 정보가 필요 없다는 장점이 있다.

E2E 기반의 음성인식 모델은 크게 Connectionist Temporal Classification(CTC)[8], Transducer[9], Attention based Encoder Decoder(AED)[10]로 나눌 수 있다[11]. CTC는 입력 음성 신호와 타겟 텍스트 간 매핑을 위해 사용되며, 출력 텍스트의 길이와 입력 신호의 길이 사이의 매핑 문제를 해결하는데 사용된다. 이 방법은 RNN[12]과 같은 순환 네트워크를 기반으로 하며, 학습 과정에서 입력과 출력 간의 매핑을 자동으로 학습하

도록 설계되어 입력 데이터에서 별도의 프레임별 정보 없이 전체적인 시계열 정보를 직접 학습할 수 있다. Transducer 는 CTC 와 유사하게 입력과 출력 사이의 연결을 처리하기 위한 모델로, 입력 시퀀스와 출력 시퀀스 사이의 매핑을 다룬다. Trasnducer 는 CTC 와는 달리 입력과 출력의 길이가 다를 수 있는 상황에서 더 유연하게 대응할 수 있으며, 음성 인식에서 매우 효과적으로 사용된다. Attention based Encoder Decoder(AED)는 주로 딥러닝 모델의 인코더와 디코더 부분에 어텐션 메커니즘을 적용하여 시퀀스 간 매핑을 수행한다. 특히 어텐션 메커니즘을 통해 모델은 입력과 출력 시퀀스 사이의 관계를 자동으로 학습한다.

E2E 은 입력 오디오와 텍스트 간의 명시적 alignment 없이 학습할 수 있다는 장점이 있다. 하지만 이러한 장점에도 불구하고 여전히 모델을 학습하기 위해서는 음성 데이터와 음성 데이터의 정답에 해당하는 transcription(전사된 텍스트)가 쌍으로 존재해야 했다. 음성인식에서 오디오의 정답에 해당하는 텍스트를 라벨링하는 것은 시간과 비용이 많이 든다. 이러한 한계로 인해 최근 텍스트가 필요 없이 오디오만을 학습하는 Self-Supervised Learning(SSL) 기반의 모델이 많이 연구되고 있다. SSL 기반의 모델 학습의 핵심은 텍스트가 없는 대용량의 음성, 즉 오디오만을 학습하여 고품질의 오디오 representation 을 갖는 것이며 이렇게 사전학습된 모델을 음성인식 down-stream task 에 fine-tuning 하여 음성인식 모델을 구축한다.

2.2.1 Whisper

최근 E2E 기반의 음성인식을 위한 Whisper[13] 모델이 제안되었다. Self-Supervised Learning 을 통해 학습한 모델은 좋은 representation 을 가지지만 특정 작업을 수행하기 위해서는 추가적인 학습이 필요하며 이때 고품질의 representation 을 가지는 encoder 에 비해 decoder 가 덜 학습되어 성능이 떨어지는 문제가 발생한다고 이야기한다. 이를 해결하기 위해 Whisper 는 weakly-supervised learning 기반으로 학습된 모델을 제안한다. Weakly-supervised learning 은 SSL 과 달리 모델 학습을 위해 음성과 텍스트가 쌍으로 존재해야하며 완벽하게 전사된 텍스트와 이전 음성인식 모델들을 통해 생성된 텍스트를 섞어 사용하는 학습 방법이다. 논문의 저자들은 지금까지 음성인식을 위한 사용된 학습 데이터 중 가장 많은 양의 웹 스케일의 대용량 데이터를 구축하였다. 구축된 데이터를 사용하여 제안된 모델을 사전 학습하였고 대규모의 데이터로 학습되어 일반화 성능이 뛰어나고 fine-tuning 없이 사용할 수 있을 만큼의 음성인식 성능을 보인다고 주장하였다. 이를 통해 주류로 연구되던 Self-Supervised learning 기반의 모델들 보다 더 좋은 성능의 모델을 만들었고 weakly-supervised learning 이 좋은 성능을 낼 수 있다는 것을 다시 한번 보여주었다. 대규모 데이터를 사용하여 학습한 효과로 일반화가 잘된다는 장점이 있다. 하지만 학습을 위해 구축된 데이터는 비교적 소음이 심하지 않은 데이터들이 많을 것이며 이것을

증명하듯 여전히 높은 소음에 대해서는 낮은 성능을 보인다는 한계가 존재한다.

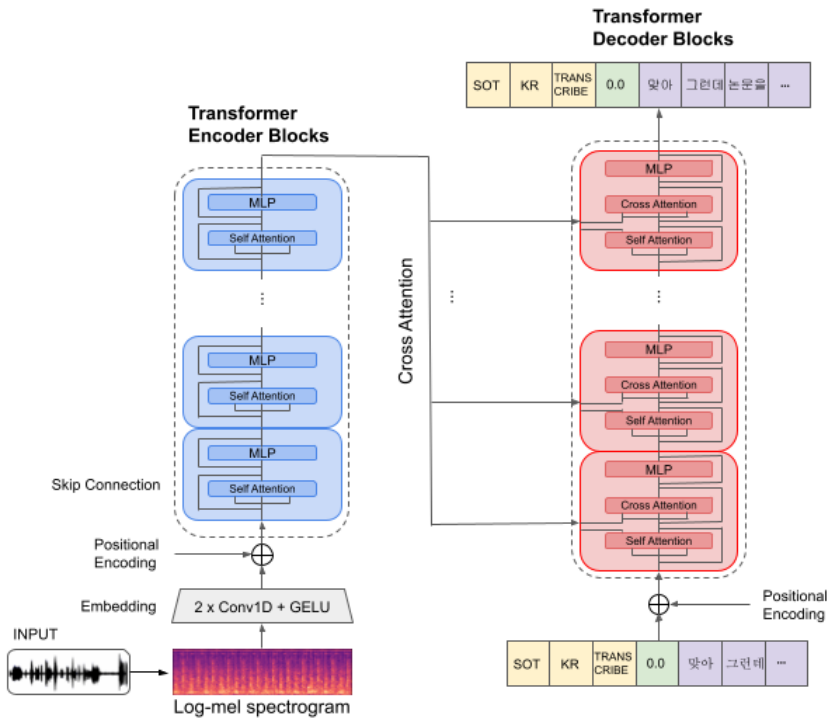

Figure 2. Whisper 구조

Ⅲ. 노이즈 제거 및 fine-tuning 프로세스

3.1 연구 방법 소개

본 연구에서 강건한 음성인식 모델을 만들기 위한 모델 학습 프레임워크는 다음과 같다. 음성인식 시스템의 Front-End 로 Denoiser 를 사용하며 Back-End 에서는 Whisper 를 사용한다. Front-End 의 Denoiser 와 Back-En 의 Whisper 모두 사전학습된 모델을 사용한다. Front-End 에서 Denoiser 를 통해 speech enhancement 를 수행한다. Waveform 의 형태를 입력으로 하는 Denoiser 모델은 원본 오디오를 입력으로 받아 upsampling, downsampling 을 거쳐 소음이 제거된 음성을 생성한다. 소음이 제거된 음성은 부자연스럽고 인위적으로 생성된 모델의 결과물이기 때문에 이를 조금 더 자연스러운 음성으로 만들기 위해 단순하지만 효과적인 OA 를 사용하여 mixed audio 를 만든다. Mixed audio 는 Back-End 모듈의 Whisper 모델의 입력으로 학습이 이뤄진다. 기존 선행 연구들은 DNN 기반의 모델을 사용한지 않았고 제안된 OA 기법 역시 DNN 기반 모델을 선택하지 않았다. 따라서 우리는 해당 기법들이 DNN 기반의 모델에서 역시 효과적인 방법인지를 실험을 통해 검증한다. 실험은 데이터셋에 whisper 를 사용하여 zero-shot 으로 WER, CER 성능을 평가하였다. 이후 Whisper 모델을 소음이 포함된 원본 음성, 소음이제거된 음성, 원본과 소음이 제거된 음성의 합성 결과인 mixed audio

에 대해 각각 fine-tuning 을 한다. 이후 각 데이터셋으로 튜닝된 모델은 test 데이터에서 원본, 소음이 제거된 음성, mixed audio 로 WER, CER 을 통해 평가를 진행하였다.

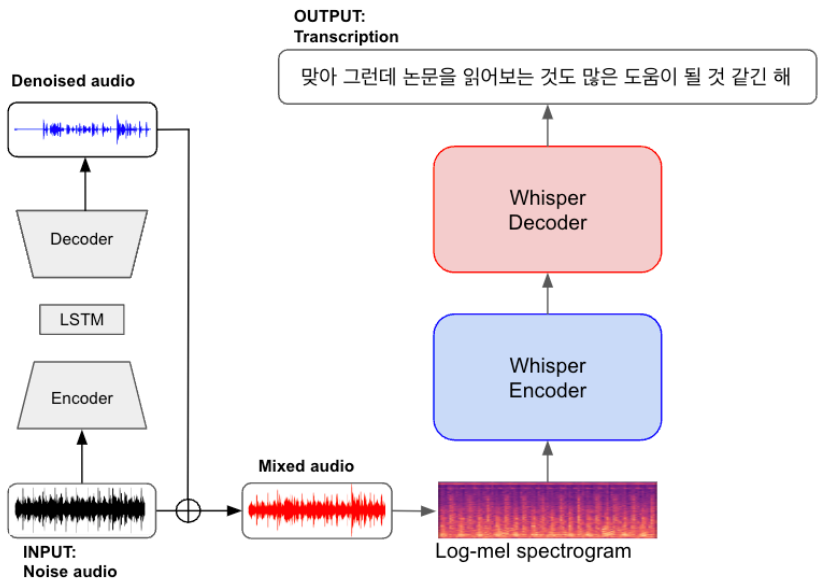

Figure 3. 소음에 강건한 모델 구축 framework

3.2 데이터 설명

3.2.1 극한 소음 음성인식 데이터셋

AI HUB 의 극한 소음 음성인식 데이터셋은 음성인식 모델의 성능 저하의 가장 큰 요인이 되는 극한 소음 환경에서 소음을 제거하는 학습 방법을 적용하여 극한 소음에서도 강건한 음성 인식 성능을 만들기 위한 연구 개발을 목적으로 구축되었다. 극한 소음 음성인식 데이터는 학습을 위한 Train, 평가를 위한 Evaluation 으로 구성되어 있다.

극한 소음인식 데이터는 다양한 소음이 있는 환경에서 녹음된 데이터로 구성된다. 데이터셋 소음의 환경의 클래스는 1.교통수단, 2.공사장, 3.공장, 4.시설류, 5.기타, 6.복합소음으로 총 6 가지로 구성된다. 1.교통수단은 다시 지상운동수단, 철로운송수단, 항공운송수단, 수상운송수간으로 나눌 수 있다. 지상운송수단은 자동차 경기장, 고속도로, 터널, 대형트럭 등 자동차의 소음 환경에서 녹음된 음성데이터이다. 철로운송수단은 열차가 통과하는 선로, 육교근처소음, 플랫폼 기차 통과 소리, 철길 건널목에서 녹음된 음성이다. 항공수단은 항공기, 경비행기, 헬리콥터, 이착륙 소음, 내부소음으로 항공운송수단과 관련된 소음이다. 수상운송수단은 수상택시, 모터 보트 소음등 배의 소음과 관련된 환경에서 녹음

된 음성이다. 2. 공사장 클래스는 경장비소음, 중장비소음 클래스로 분류가 되고 경장비소음은 공사자의 기계음(그라인더, 기계톱, 절단기, 전동해머드릴)과 관련된 소음환경에서 녹음된 음성이며, 중장비소음은 공사장의 중장비 클래스는 공사장 중장비소음(굴삭기, 착암기, 천공기, 항타기)이 나는 환경에서 녹음된 소음이다. 3. 공장은 공장기계음 하나의 클래스를 가지며 공장 기계에서 발생하는 소리(제철소, 자동차공장, 방직공장, 목재소)가 있는 환경에서 녹음된 음성이다. 4. 시설류는 실내시설, 실외시설 클래스를 가진다. 실내시설은 게임장(오락실), 공장연(오케스트라, 국악, 콘서트) 실내 서핑장, 실내 사격장, 실내 경기장 등 실내 레저와 관련된 환경에서 나오는 소리와 함께 녹음된 음성이다. 실외시설은 실외 경기장(야구, 축구, 경마, 경륜), 폐차장, 놀이시설, 행사장 등 실외 시설에서 발생하는 소음 환경에서 녹음된 소음이다. 5. 기타는 실내기타 소음, 실외기타소음 클래스를 가지며 각각 기계실(서버실), 펌프실, 공조시설, 자동차 검사소에서 녹음된 음성이다. 실외기타소음은 여름철 매미소리, 산업용 진공 기계 소리, 싸이렌 소리, 농기계(예초기, 트랙터, 경운기), 무선 모형 엔진 소리, 드론, 천연폭포, 빗소리 소음에서 녹음된 데이터이다. 마지막으로 6. 복합소음은 2 가지이상소음으로 구성된 소음으로 공장소리,빗소리와 자동차소리, 싸이렌 소리와 같이 2 가지 이상 소음이 섞인 환경에서 녹음된 음성이다.

Table 1. 극한 소음 음성인식 데이터 연령 분포

연령	건(수)	구성비
20대	12,386.0	24.5%
30대	12,858.0	25.4%
40대	11,816.0	23.3%
50대	13,552.0	26.8%
합계	50,612.0	100.0%

Table 2. 극한 소음 음성인식 데이터 성별 분포

성별	건(수)	구성비
남성	23,062.0	45.6%
여성	27,550.0	54.4%
합계	50,612.0	100.0%

3.2.2 데이터 전처리

소음에 강건한 음성인식 모델을 학습하기 위해 데이터 전처리를 진행하였다. 극한 소음 음성인식의 각 데이터는 하나의 대화 주제에 대해 두명의 화자가 대화한 것을 녹음한 파일이다. Back-End 의 Speech Recognition 을 수행하는 음성인식 모델인 Whisper 를 학습하기 위하여 음성을 30 초 단위의 chunk 로 segmentation 을 진행하였다. 데이터 전처리 작업은 AI HUB 의 극한 소음 음성인식 데이터에 포함된 meta 에 파일 별 각 화자별 발화 시간과 발화 종료 시간을 기준으로 문장을 segmentation 하였다. 전처리되어 화자별로 나눠진 문장들에서 30 초 이상 발화된 문장들은 학습에서 제외하였다. 이후 학습된 모델을 평가하기 위해 Evaluation 에서 50%를 나누어 학습된 모델을 평가하기 위한 Test 데이터셋으로 구축하였다. Segmentation 과정과 전처리 후 train, validation, test 데이터셋의 개수와 시간은 각각 아래 그림과 표와 같다.

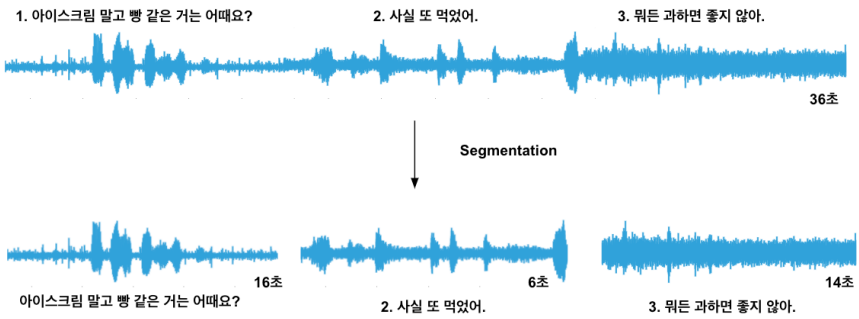

Figure 4. 데이터 전처리 segmentation 예시

Table 3. 극한 소음 음성인식 데이터 전처리 후 구성

데이터셋	개수	시간
Train	124,054	348.69
Validation	11,714	32.60
Test	11,630	32.56

모델의 학습을 위한 데이터셋의 정답에 해당하는 텍스트 데이터는 외래어, 숫자 표기의 경우 한글 표기와 괄호 안의 영어로 중복 표기 되어 성능 측정에 영향을 미치므로 정규식을 사용하여 의미가 같은 괄호 안의 문자들은 제거하였다. 음성인식 결과인 텍스트 구두점과 따옴표, 쉼표 등과 같은 문장 부호들은 역시 음성인식 성능에 직접적으로 영향을 미칠 수 있다. 문장 부호들의 제거를 위해 Whisper 논문의 전처리 과정과 동일하게 제거하였다. 이 과정에 huggingface whisper 라이브러리에서 제공하는 BasicTextNormalizer 를 사용하였다.

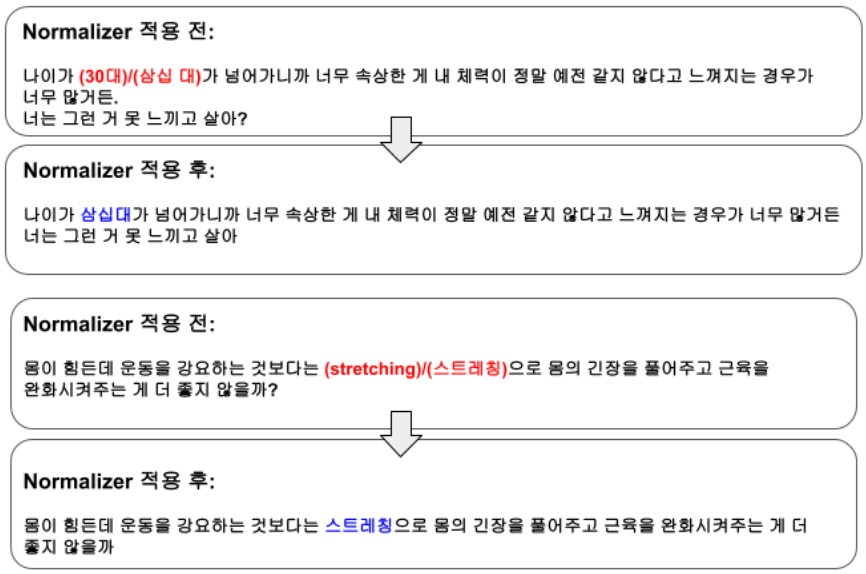

Figure 5. BasicTextNormalizer 예시

3.3 Training Details

극한 소음 데이터는 Front-End의 Denoiser 모델에 의해 소음이 제거된 waveform의 형태로 출력이 된다. 이후 OA 기법을 사용하여 원본 음성과 소음이 제거된 음성이 합쳐져 mixed audio(waveform)이 된다. 하지만 Back-End의 Whisper의 입력은 Spectrogram이므로 waveform을 Spectrogram으로 변환하는 과정을 거쳐야 한다. 이러한 변환은 time-domain인 waveform을 frequency-domain으로 변환시키는 Fourier Transform 함수를 이용하여 진행된다. 모델의 입력으로 들어가기 위해 waveform 형태의 음성이 spectrogram으로 변환되어야 하고 이것은 학습에 직접적으로 관련 없이 시간이 드는 과정이다. 따라서 전처리된 데이터셋들에 대하여 사전에 Spectrogram으로 변환하였다. 사전학습된 Whisper 모델은 parameter의 크기에 따라 whisper-tiny, whisper-base, small, medium, large(v, v2)가 있다.

Table 4. Whisper 모델 사이즈별 파라미터 수

Size	Tiny	Base	Small	Medium	Large
파리미터	39M	74M	244M	796M	1,550M

학습에 사용한 모델은 huggingface의 Whisper-small을 사용하였다. Whisper 모델이 소음에 강건하지 못하다는 것을 확인하기 위해 Whisper-small 모델을 사용하여 test 데이터셋에 대해 zero-shot 성능을 평가하였다. 극한 소음 음성인식의 원본 데이터을 사용하여 whisper 모델을 fine-tuning하고 극한 소음 음성인식 데이터의 일부인 test 데이터에 대한 성능을 측정하였다. 원본 극한 소음 음성인식 데이터에 대해 fine-tuning된 모델의 성능은 본 연구의 목적인 DNN 기반의 음성인식 모델이 OA를 이용해 생성된 mixed audio를 통해 강건한 음성 모델을 구축할 수 있는지에 대해 평가하기 위한 기준으로 잡았다. Speech Enhancement가 음성인식 성능에 부정적인 영향을 미치는 것을 증명하기 위해 Denoiser를 사용하여 소음이 제거된 denoised audio를 사용하여 모델을 fine-tuning하였다. 본 연구의 강건한 음성인식 모델의 핵심인 Mixed Audio를 사용하여 Whisper 모델을 fine-tuning하였다. 각 실험은 16 batch size, warmup steps 500, 10000 training step을 동일하게 학습하였다.

Ⅳ. 실험 결과

4.1 Evaluation Metrics WER, CER

음성인식 성능을 측정하는 데 사용되는 지표에는 WER(Word Error Rate)과 CER(Character Error Rate)이 있다. WER은 음성 인식의 결과와 실제 정답 사이의 단어 수준의 에러 비율을 측정한다. CER은 문자 단위에서의 에러 비율을 나타내며 문자 단위에서의 오류를 측정한다. WER과 CER 모두 음성인식 시스템의 정확성과 성능을 평가하는 데 사용되며, 더 낮은 WER과 CER이 높은 정확성을 나타낸다. 일반적으로 ASR (STT) task에서는 evaluation metric으로 Word Error Rate (WER)을 사용하지만 한국어는 교착어로 조사를 사용하며 다른 언어들과 비교하면 형태소의 구조가 복잡하다. 따라서 한국어 구조의 특성으로 단어 단위의 평가인 WER 보다 문자 수준의 오류 측정 방법인 CER이 더 정확한 평가 방법으로 간주된다.

Table 5. 각 데이터셋으로 fine-tuning된 모델의 성능

Model	Original Audio		Denoised Audio		Mixed Audio	
	WER	CER	WER	CER	WER	CER
Zero-shot	145.56	116.79	359.57	283.55	157.45	122.66
Noise Mdodel	56.58	32.18	71.97	45.25	56.56	32.0
Denoised Model	65.36	39.87	60.19	34.95	59.46	34.65
Mixed Model	57.28	33.31	69.36	43.2	54.08	30.48

4.2 Zero-shot

사전학습된 Whisper 모델의 소음에 대한 강건성을 평가하기 위해 극한 음성 데이터셋을 학습하지 않고 평가하는 zero-shot 평가를 진행하였다. 극한 소음 음성인식 데이터셋으로 fine-tuning 되지 않은 모델의 zero-shot 성능은 original audio 에 대해서 WER: 145.45, CER: 116.79, denoised audio 에 대해 WER: 359.57, CER: 283.55, mixed audio 에 대해 WER: 157.45, CER: 122.66 인 결과를 보였다. Whisper 의 극한 소음 음성인식에 대한 zero-shot 성능은 모든 실험들 중 가장 높은 WER, CER 을 기록하였다.

4.3 Fine-tuning on Original Dataset

Whisper 모델을 극한 소음 음성인식 원본 데이터셋 대해 학습을 진행한 결과이다. 원본 데이터셋에 대해 fine-tuning 된 모델은 original audio 에 대해 WER: 56.58, CER: 32.18, denoised audio 에 대해 WER: 71.97, CER: 45.25, mixed audio 에 대해 WER: 56.66, CER: 32.0 성능을 보였다. Denoised Audio 의 경우 WER, CER 은 각각 71.97, 45.25 로 origina audio 로 fine-tuning 된 모델의 성능 결과 중 가장 낮았다. Original audio 와 mixed audio 의 WER 은 각각 56.58, 56.66 으로 비슷

한 성능을 보였고 CER 의 경우 각각 32.18, 32.0 으로 Mixed Audio 에서 더 좋은 성능을 보였다.

4.4 Fine-tuning on Denoised Dataset

소음이 제거된 음성인 denoised audio 로 fine-tuning 된 Whisper 모델의 WER 은 60.19 , CER 은 34.95 이지만 mixed audio 의 WER 은 59.46, CER 은 34.65 로 Denoise Audio 로 fine-tuning 되었지만 Mixed Audio 에 대해 가장 좋은 성능을 보였다. Original Audio 와 Mixed Audio 로 fine-tuing 된 모델들과 비교하면 zero-shot 을 제외한 fine-tuning 된 모델들의 모든 결과 중 가장 높은 WER, CER 을 기록했다.

4.5 Fine-tuning on Mixed Dataset

본 연구의 소음에 강건한 음성인식 모델 구축을 위한 방법인 original audio 와 소음이 제거된 denoise audio 를 합하여 생성된 mixed audio 로 fine-tuning 된 모델은 original audio 에 대해 WER 57.28, CER 33.31, denoised audio 에 대해 WER 69.36, CER 43.2, mixed audio 에 대해 WER 54.08, CER 30.48 을 성능을 보였다. 특히 mixed audio 에 대해서는 모든 실험 결과에서 가장 낮은 WER, CER 을 보였다.

Ⅴ. 결론

Whisper의 zero-shot 성능은 모든 실험 결과들 중 가장 높은 WER, CER을 수치를 보이는 것으로 보아 모델이 소음에 강건하지 못하다는 것을 알 수 있었다. Denoised audio에 대해 가장 높은 WER, CER을 가지는 것을 통해 소음이 제거된 음성을 학습 없이 사용하는 것이 모델의 성능에 부정적인 영향을 미친다는 것을 보여준다. Denoised audio에 대해 fine-tuning된 모델의 denoised audio에 대한 WER은 60.19, CER은 34.95, original audio로 fine-tuning된 모델의 original audio에 대한 WER은 56.58, CER은 32.18, mixed audio로 fine-tuning된 모델은 mixed audio에 대해 WER은 54.08, CER은 30.48의 성능을 보였다. 이 중 denoised audio에 대해 fine-tuning된 모델이 denoised audio에 대해 가장 낮은 성능을 보여준다. Fine-tuning된 데이터와 동일한 종류의 데이터로 평가되었지만 denoise audio에서 가장 낮은 성능을 보이는 것으로 보아 이는 Speech Enhancement를 통해 향상된 음성이 모델의 학습에 부정적인 영향을 미치며 SE의 결과물인 denoised audio를 음성인성을 위한 학습용 데이터로 직접 사용하는 것은 효과적인 방법이 아니라는 것을 보여준다. Mixed audio로 fine-tuning된 모델의 mixed audio에 대한 WER, CER은 모든 실험들과 비교하여 가장 좋은 성능을 보인다. 또한 original audio에 fine-tuning된 모델의 original audio에 대한 WER과 mixed audio에 대한

WER 의 성능이 0.02 의 미세한 차이를 보였으며 mixed audio 에 대한 CER 은 original audio 에 0.18 낮은 좋은 성능을 보였다. 실험의 결과를 통해 제안된 mixed audio 의 방법은 DNN 기반의 소음에 강건한 모델을 구축하는 데 효과적인 방법이라는 것을 보여준다.

Ⅵ. 부록

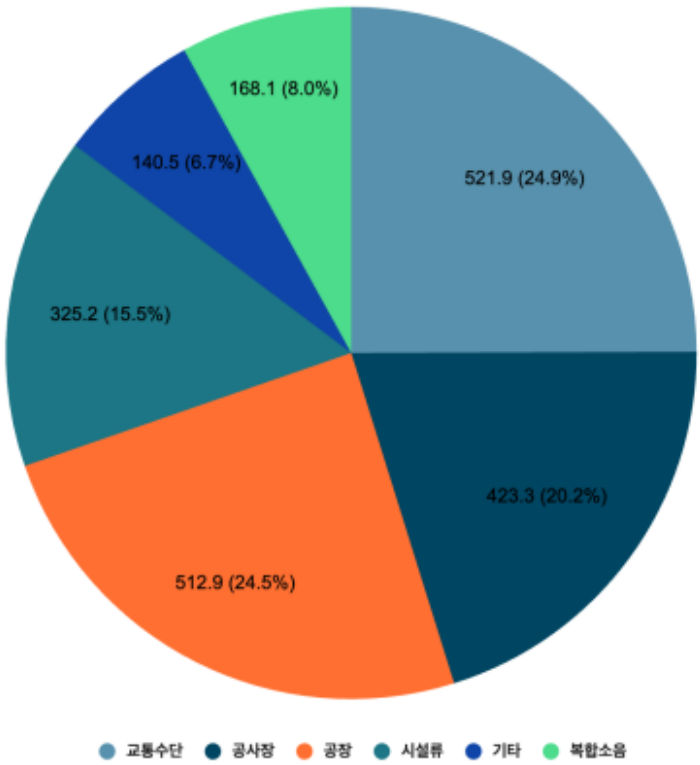

Figure 6 극한 소음 음성인식 데이터 구축 시간

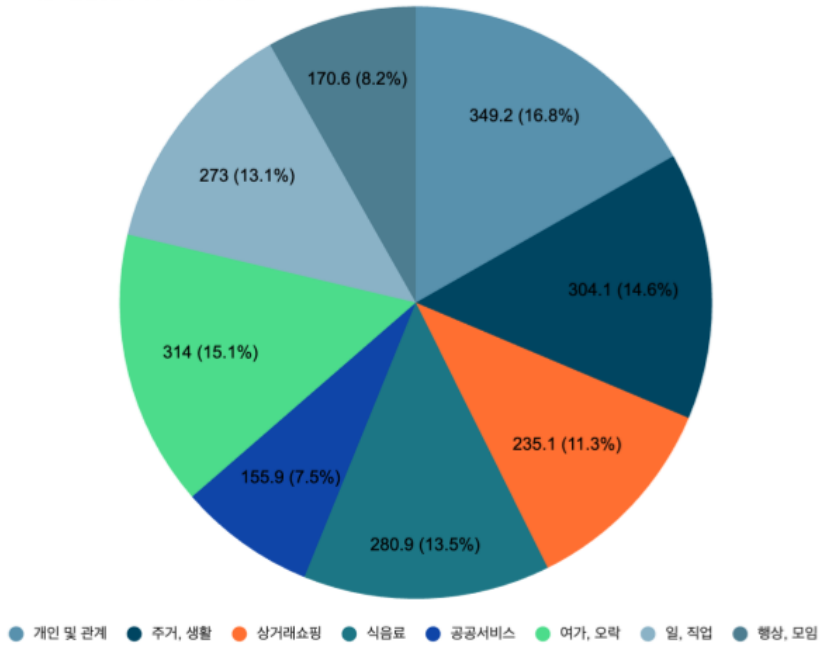

Figure 7 극한 소음 음성인식 데이터 대화 주제 분포

참고 문헌

[1] Jurafsky Dan (2000)Speech & languageprocessing. PearsonEducation India: 1.

[2] Y. Luo and N. Mesgarani (2019) "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 8, pp. 1256-1266

[3] D'efossez, A., Usunier, N., Bottou, L., & Bach, F.R. (2019). Music Source Separation in the Waveform Domain.ArXiv, abs/1911.13254.

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition.2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778.

[5] Défossez, A., Synnaeve, G., & Adi, Y. (2020). Real Time Speech Enhancement in the Waveform Domain.ArXiv, abs/2006.12847.

[6] Iwamoto, K., Ochiai, T., Delcroix, M., Ikeshita, R., Sato, H., Araki, S., & Katagiri, S. (2022). How Bad Are Artifacts?: Analyzing the Impact of Speech Enhancement Errors on ASR.Interspeech.

[7] Kinoshita, K., Ochiai, T., Delcroix, M., & Nakatani, T. (2020). Improving Noise Robust Automatic Speech Recognition with Single-Channel Time-Domain Enhancement Network.ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7009-7013.

[8] Graves, A., S. Fernandez, F. Gomez, and J. Schmidhuber (2006) Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. ICML.
[9] Graves, A., A.-r. Mohamed, and G. Hinton (2013) Speech recognition with deep recurrent neural networks. ICASSP.
[10] Chan, W., N. Jaitly, Q. Le, and O. Vinyals (2016) Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. ICASSP.
[11] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, N. Jaitly (2017) A comparison of sequenceto-sequence models for speech recognition. Interspeech: 939-943
[12] Williams, Ronald J, Hinton, Geoffrey E. Rumelhart, David E. (1986). "Learning representations by back-propagating errors. Nature
[13] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision.ArXiv, abs/2212.04356.