JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 1, pp. 41-49, January 2024
https://doi.org/10.9708/jksci.2024.29.01.041

A Study on the Impact of Speech Data Quality on Speech Recognition Models

Yeong-Jin Kim*, Hyun-Jong Cha**, Ah Reum Kang***

*Student, Dept. of Smart ICT Convergence, Pai Chai University, Daejeon, Korea
**Professor, Dept. of Software Engineering, Pai Chai University, Daejeon, Korea
***Professor, Dept. of Information Security, Pai Chai University, Daejeon, Korea

[Abstract]

Speech recognition technology is continuously advancing and widely used in various fields. In this study, we aimed to investigate the impact of speech data quality on speech recognition models by dividing the dataset into the entire dataset and the top 70% based on Signal-to-Noise Ratio (SNR). Utilizing Seamless M4T and Google Cloud Speech-to-Text, we examined the text transformation results for each model and evaluated them using the Levenshtein Distance. Experimental results revealed that Seamless M4T scored 13.6 in models using data with high SNR, which is lower than the score of 16.6 for the entire dataset. However, Google Cloud Speech-to-Text scored 8.3 on the entire dataset, indicating lower performance than data with high SNR. This suggests that using data with high SNR during the training of a new speech recognition model can have an impact, and Levenshtein Distance can serve as a metric for evaluating speech recognition models.

▸ **Key words**: Speech Recognition, Signal-to-Noise-Ratio(SNR), Levenshtein Distance Algorithm, Meta Seamless M4T, Google Cloud Speech-to-Text

[요 약]

현재 음성인식 기술은 꾸준히 발전하고 다양한 분야에서 널리 사용되고 있다. 본 연구에서는 음성 데이터 품질이 음성인식 모델에 미치는 영향을 알아보기 위해 데이터셋을 전체 데이터셋과 SNR 상위 70%의 데이터셋으로 나눈 후 Seamless M4T와 Google Cloud Speech-to-Text를 이용하여 각 모델의 텍스트 변환 결과를 확인하고 Levenshtein Distance를 사용하여 평가하였다. 실험 결과에서 Seamless M4T는 높은 SNR(신호 대 잡음비)을 가진 데이터를 사용한 모델에서 점수가 13.6으로 전체 데이터셋의 점수인 16.6보다 더 낮게 나왔다. 그러나 Google Cloud Speech-to-Text는 전체 데이터셋에서 8.3으로 높은 SNR을 가진 데이터보다 더 낮은 점수가 나왔다. 이는 새로운 음성인식 모델을 훈련할 때 SNR이 높은 데이터를 사용하는 것이 영향이 있다고 할 수 있으며, Levenshtein Distance 알고리즘이 음성인식 모델을 평가하기 위한 지표 중 하나로 쓰일 수 있음을 나타낸다.

▸ **주제어**: 음성인식, 신호 대 잡음비, Levenshtein Distance 알고리즘, Meta Seamelss M4T, Google Cloud Speech-to-Text

• First Author: Yeong-Jin Kim, Corresponding Author: Ah Reum Kang
*Yeong-Jin Kim (jin971001@naver.com), Dept. of Smart ICT Convergence, Pai Chai University
**Hyun-Jong Cha (hjcha@pcu.ac.kr), Dept. of Software Engineering, Pai Chai University
***Ah Reum Kang (armk@pcu.ac.kr), Dept. of Information Security, Pai Chai University
• Received: 2023. 10. 30, Revised: 2024. 01. 25, Accepted: 2024. 01. 25.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근 음성인식 기술은 꾸준한 발전을 거듭하여 현대 사회의 다양한 분야에서 널리 활용되고 있다[1]. 음성인식 기술은 음성명령 시스템, 음성 검색, 실시간 자막 시스템 등 다양한 응용 분야에서 사용된다. 이에 따라 음성인식 모델의 성능 향상이 더욱 중요한 연구 주제로 떠오르고 있다. 특히 한글 음성인식 모델의 연구와 개발은 대한민국만이 사용하는 한글에 관한 연구의 필요성이 대두되고 있다.

그러나 실제 음성 데이터는 주변의 잡음 및 간섭 등의 다양한 환경조건에 의해 영향을 받는다. 그렇기 때문에 음성인식 모델은 실제 환경의 변수들을 고려해여 개발되어야 한다[2]. 특히 주변의 잡음은 음성 데이터의 품질에 큰 영향을 미친다. 대표적으로 Joseph과 Cheyenne은 지난 2020년 이후 COVID-19로 인해 모든 사람이 마스크를 끼고 다니게 되면서 음성인식이 배경 소음의 수준에 따라 마스크의 유형별로 미치는 영향을 조사하였다[3].

음성인식 모델의 성능이 환경 잡음과 SNR에 얼마나 영향을 받지 않는지를 분석하는 것이 중요하다. 높은 SNR은 음성 신호의 선명도와 정확성을 높일 수 있으며, 음성인식 모델의 성능을 향상시킬 수 있다. 그러므로 한글 음성인식 모델의 성능 향상을 위해서는 환경 잡음에 대한 영향과 적절한 SNR 조건의 연구가 필요하다. 이미지 인식을 위한 데이터셋은 ImageNet 데이터셋과 같이 대용량 데이터셋이 많이 존재한다. 하지만 오디오 데이터셋은 이미지 데이터셋에 비해 그 수가 많지 않다. 이를 해결하기 위해 Google에서는 Youtube 영상을 기반으로 데이터셋을 제작하였다[4]. 하지만 해당 데이터셋은 영어만 포함하고 있다. 임연수 외는 한국어 음성 명령어 인식기 개발에 필요한 데이터를 웹에서 자동으로 추출하고, 학습 데이터로 사용할 수 있는 데이터를 자동으로 선별하는 방법을 소개하였다[5]. 그러나 위 연구에서 수집한 데이터의 양은 매우 적은 편이다. 따라서 한국어 데이터셋이 더 필요한 경우에는 데이터를 직접 수집해야 하는 경우도 존재한다. 데이터셋을 직접 제작할 때, 조용한 환경에서 녹음을 진행한다. 또한 연구 목적에 따라 소음이 있는 환경에서 녹음을 진행한다. 이 외에도 조용한 환경에서 녹음한 음성에 후처리로 소음을 넣을 수 있는데, 조용한 환경에서 녹음한 음성에 후처리로 소음을 넣게 될 때 의도치 않게 잡음이 너무 많이 들어간 데이터가 섞여 훈련에 영향을 미칠 수 있기에 이 데이터들을 다시 정제해 주는 작업이 필요하다.

본 연구는 한글 음성인식 모델의 성능향상을 위해 환경 잡음과 SNR이 한글 음성 인식 모델의 성능에 미치는 영향과 새로운 음성인식 모델이 효과적인지 알아보고자 한다. 따라서 다양한 SNR 조건에서 음성 인식 모델의 성능을 분석하고 SNR 조건에 따른 모델의 최적 성능을 도출하기 위한 방법을 제안한다. Hans와 David는 소음 환경에서의 음성인식 시스템을 평가하기 위해 잡음 신호를 -5dB에서 20dB로 제한해서 실험하였다[6]. 본 논문에서는 더욱 다양한 음성 데이터 환경을 고려하여 잡음 신호의 기준을 두지 않고, 잡음이 포함된 전체 데이터와 SNR 수치 상위 70%의 데이터로 데이터셋을 나누어서 실험하였다. 또한, SNR이 음성인식 모델의 학습 데이터에 미치는 영향을 고려한다. 본 논문은 만들어진 학습 데이터의 다양성과 품질이 음성인식 모델의 일반화 성능에 미치는 영향을 평가한다, 기존의 연구에서 사용하지 않았던 Meta사의 Seamless M4T 모델을 자주 사용하는 Google의 Cloud Speech to Text API와 비교해 보았다. 또한 기존의 연구에서는 사용하지 않던 평가지표인 Levenshtein Distance을 활용한다.

2장에서는 본 연구와 관련된 기술과 선행연구에 대해 기술한다. 3장에서는 전체적인 시스템의 개요와 환경, 실험 과정에 대해 기술한다. 또한 4장에서는 실험의 결과에 대해 분석하고, 5장에서는 결과 해석과 향후 연구에 대해 기술한다.

II. Preliminaries

1. Related works

음성인식 모델의 성능을 향상하거나 기존의 음성-텍스트 변환 모델의 성능을 비교한 연구들은 크게 음성인식 모델의 성능향상, 음성인식과 STT를 활용한 프로그램 구현, 음성인식 모델의 비교로 나뉜다.

이예진 외[7]는 한국어 데이터에서 노이즈를 합성한 음성 데이터를 잡음제거 알고리즘으로 제거한 후 Google Cloud Speech-to-Text 라이브러리를 이용하여 음성 데이터를 텍스트 데이터로 변환하고 WER(Word Error Rate), CER(Character Error Rate), PESQ(Perceptual Evaluation of Speech Quality)등을 사용하여 성능을 평가하였다.

김보경 외[8]는 OTT 서비스의 자막으로 쓰이는 음성인식 모델의 정확도를 올리기 위해 Vocal Remover를 이용하여 비음성 오디오를 추출한 뒤 Google의 음성인식 서비스를 이용하여 텍스트 변환을 진행한 후 추출된 단어의 개수를 비교하였다. 음성 신호 필터를 통해 추출된 오디오 데이터가 원본 오디오 데이터에 비해 평균적으로 더 많은

Table 1. Studies to Improve the Performance of Voice Recognition Models

Category	Content
Improved performance of Voice Recognition Model	Noise removal from audio data with background noise[7]
	Comparing results by extracting noise using Vocal Remover and proceeding with text conversion[8]
	Using augmentation algorithm with formant enhancement[9]
	Proposing a method to differentiate between speech signals and noise signals through correlation operations[10]
	Proposing a deep learning model that combines acoustic scene classification techniques and location-based technologies to achieve environment-specific audio enhancement[11]
	Proposing a method to enhance learning by incorporating mouth shape image data[12]
	Suggesting a method incorporating pronunciation considerations using BERT[13]
	In a dataset where three speakers are speaking simultaneously performance degradation is observed regardless of the presence of noise[14]
Implementation of a program utilizing Speech Recognition and STT	Proposing a system for recording everyday-life audio, utilizing an STT model to convert it into text, and storing it on a server[15]
Comparison of Speech Recognition Models	Comparing 14 models with the addition of white noise and pub noise[16]

단어를 추출하는 것을 확인했다.

강병휘와 노동건[9]은 음성의 공명 주파수이자 언어의 명료도에 영향을 미치는 포먼트 주파수에 가중치를 부여하는 포먼트 강화를 활용한 데이터 증강 알고리즘을 사용하여 음성인식의 성능을 높일 수 있음을 확인하였다.

임지원 외[10]는 음성인식 모델의 인식률을 높이기 위해 음성 신호를 추출하여 주파수 성분을 분석하고 오디오 신호 사이의 주파수 영역 correlation 연산을 통해 음성 신호와 노이즈 신호를 구분하는 방법을 제안하였다.

강병휘와 노동건[11]은 환경 특수적으로 발생할 수 있는 잡음을 효과적으로 제거하기 위해 음향 장면 분류 기법과 위치 정보 활용 기술을 결합한 음성인식 모델을 제안하였다. 이 모델을 기존의 모델과 비교한 결과 PESQ값은 평균 0.06 이상, STOI(Short-Time Objective Intelligibility) 값은 평균 0.015 이상의 향상된 품질을 보였다.

J.C. Hou 외[12]는 음성인식 기술의 향상을 위해 오디오 정보 처리에 중점을 두는 것이 아니라 입모양 이미지를 함께 학습시키는 모델을 제안하였다. 이를 통해 음성인식 모델의 성능을 높이는 데 성공하였다.

박영미와 김철연[13]은 음성인식 정확도를 개선하기 위해 BERT를 활용하여 발음도 함께 고려하는 방법을 제안하였다.

Joris Cosentino 외[14]는 음성인식 모델이 한 데이터셋에서 높은 성능을 달성하더라도 다양한 화자로부터 제작한 음성 데이터셋이나 다른 환경에서 녹음된 음성 데이터셋에 대한 성능은 떨어지는 것을 확인하였다. 특히 세 명의 말이 overlap된 데이터셋에서는 노이즈의 유무와 상관없이 성능이 떨어지는 것을 확인하였다.

최정윤 외[15]는 모바일 기기로 일상생활에서의 음성을 녹음하고 텍스트로 변환하여 서버에 저장하는 시스템을 제안하였다. 이 시스템에서는 네이버나 카카오의 음성인식 모델보다 Google의 음성인식 모델이 문장의 전체적인 이해도와 인식률이 높아 Google의 음성인식 모델을 사용하였다.

A. Radford 외[16]는 음성인식 모델이 self-supervision이나 self-training 기술 없이 weakly-supervised 사전 학습을 통해 좋은 결과를 얻을 수 있음을 확인하였다. 또한 음성언어 모델링 연구에 기여하기 위해 openAI Whisper를 공개하였다.

2. Related Technologies

2.1 SNR

SNR은 음성인식과 신호 처리에서 중요한 개념이다. “Signal-to-Noise-Ratio”의 약자로, 한국어로는 “신호 대 잡음비”로 사용된다. SNR은 신호와 잡음 사이의 상대적인 강도를 나타내며 클수록 원하는 신호가 뚜렷하게 들리는 것을 의미한다. SNR은 주로 음성데이터에서 신호(원하는 음성)와 잡음(배경 소음, 잡음, 간섭 등) 사이의 상대적인 세기를 비교하는 데 사용된다. SNR은 일반적으로 dB 단위로 표현된다. SNR은 음성 인식 및 오디오 처리 알고리즘에 큰 영향을 미치며 수식 1로 계산된다[17].

$$SNR = 10*\log(\frac{P_{signal}}{P_{noise}}) \quad (1)$$

여기서 P_{noise}는 잡음의 세기, P_{signal}은 신호의 전력이다. 일반적으로 높은 SNR은 좋은 음질이나 정확한 음성인식을 나타낸다.

본 연구에서는 각 음성 데이터의 SNR을 계산하여 높은 SNR을 가진 데이터셋과 잡음이 포함된 전체 데이터셋을 비교하는 연구를 진행하였다.

2.2 Sound to Text

STT(Sound to Text)는 음성 데이터를 텍스트로 변환하는 기술이다. 이 기술은 음성인식(TTS, Text to Speech) 또는 음성 투표와 함께 자연어 처리 분야에서 중요한 역할을 한다. STT 시스템은 음성 신호를 분석하여 음성에서 음소, 단어 또는 문장을 인식하고, 이를 텍스트로 변환하는 과정을 거친다[18].

STT 기술의 주요 응용 분야 중 하나는 음성 비서 및 음성 명령 인식이다. 이를 통해 사용자들은 음성으로 컴퓨터, 스마트폰, 스마트 홈 기기와 상호 작용할 수 있으며, 텍스트로 변환된 음성을 기반으로 원하는 작업을 수행할 수 있다. 또한 음성 대화형 시스템, 자동 음성 번역, 음성 검색 및 음성 데이터의 텍스트 변환과 같은 다양한 응용 분야에서 STT 기술은 중요한 역할을 한다. 이를 통해 다국어 간 커뮤니케이션, 음성 데이터의 텍스트화, 음성 데이터 분석 등 다양한 활용 가능하다.

2.3 Levenshtein Distance

Levenshtein Distance 알고리즘은 두 문자열 간의 차이를 나타내는 데 사용되는 알고리즘이다[19]. 이 알고리즘은 Dynamic programming을 사용하여 두 문자열 간의 최소 편집 횟수를 효율적으로 계산하는 알고리즘이다.

Levenshtein Distance 알고리즘의 원리는 세 가지이다.

- 2D 배열을 초기화하고, 각 셀에는 해당 위치까지의 부분 문자열 간의 편집 거리를 저장한다.
- 배열을 채우기 위해 각 셀에 대해 삽입, 삭제, 대체 연산을 고려하여 최소 편집 횟수를 계산한다.
- 배열의 마지막 셀에는 최종적인 편집 거리가 저장된다.

Levenshtein Distance를 계산하는 함수는 세 가지 연산을 사용한다[20].

- 삽입(Insertion) : 문자열에 문자를 추가하는 연산
- 삭제(Deletion) : 문자열에서 문자를 제거하는 연산
- 대체(Substitution) : 문자열의 문자를 다른 문자로 교체하는 연산

Levenshtein Distance 알고리즘은 철자 교정, 검색엔진에서의 검색어 유사성 측정, 노이즈가 있는 데이터에서 정확한 일치를 찾는 등 다양한 자연어 처리 분야에서 활용된다. 또한 음성인식에서 사용되는 성능 지표 중 WER와 CER은 Levenshtein Distance에서 점수를 얻어와서 사용하기도 한다.

본 연구에서는 각 모델에서의 성능 비교를 위해 Levenshtein Distance 라이브러리를 사용하였다.

3. STT Models

3.1 Seamless M4T

Seamless M4T는 Meta AI에서 개발한 번역과 전사가 가능한 최초의 올인원 다국어 멀티모달 AI 모델로 2023년 8월에 공개되었다. Seamless M4T는 메타버스 및 가상현실(Virtual Reality, VR) 증강현실(Augmented Reality, AR)과 같은 가상환경에서 사용자의 음성 대화와 상호작용을 원활하게 지원하기 위해 설계되었다. 이 기술은 사용자가 음성으로 정보를 검색하거나 가상환경 내에서 명령을 내릴 때, 음성 입력을 신속하고 정확하게 텍스트로 변환하여 응용 프로그램 및 플랫폼에서의 다양한 기능을 활용할 수 있도록 한다[21].

Seamless M4T는 음성분석, 음성인식 순으로 작동한다. 먼저, 사용자의 음성 입력이 수집되고 음성 신호 분석을 거친다. 이 과정에서 음성의 주파수, 음높이, 발음 및 억양과 같은 다양한 특징을 추출하여 분석한다. 음성인식 과정에서는 분석된 음성 데이터를 텍스트로 변환하는 과정이 진행된다. 이때 Seamless M4T는 다양한 언어와 억양을 인식하며, 다국어 환경에서도 효과적으로 작동한다.

3.2 Google Cloud Speech-to-Text

Google Cloud Speech-to-Text는 Google Cloud 플랫폼에서 제공되는 음성인식 서비스로, 음성 데이터를 텍스트로 변환하는 핵심 도구 중 하나이다. 이 서비스는 사용자가 음성을 입력하면 정확하게 텍스트로 변환하여 다양한 응용 분야에 활용할 수 있다[22].

Google Cloud Speech-to-Text가 음성인식을 수행하는 주요 방법은 동기 인식, 비동기 인식, 스트리밍 인식으로 나뉜다.

- 동기 인식: 오디오 데이터를 Speech-to-Text API로 보내고 해당 데이터를 인식하여 모든 오디오가 처리된 후에 결과를 반환한다. 동기식 인식 요청 대상은 길이가 1분 이하인 오디오 데이터로 제한된다.
- 비동기 인식: 오디오 데이터를 Speech-to-Text API로 보내고 장기 실행 작업을 시작한다. 이 작업을 사용하여 주기적으로 인식 결과를 폴링할 수 있다. 최대 480분 길이의 오디오 데이터에 비동기식 요청을 사용한다.
- 스트리밍 인식: gRPC(Google Remote Procedure Call) 양방향 스트림에 제공되는 오디오 데이터를 인식한다. 스트리밍 요청은 마이크에서 라이브 오디오 캡처 용도와 같은 실시간 인식 용도로 설계되었다.

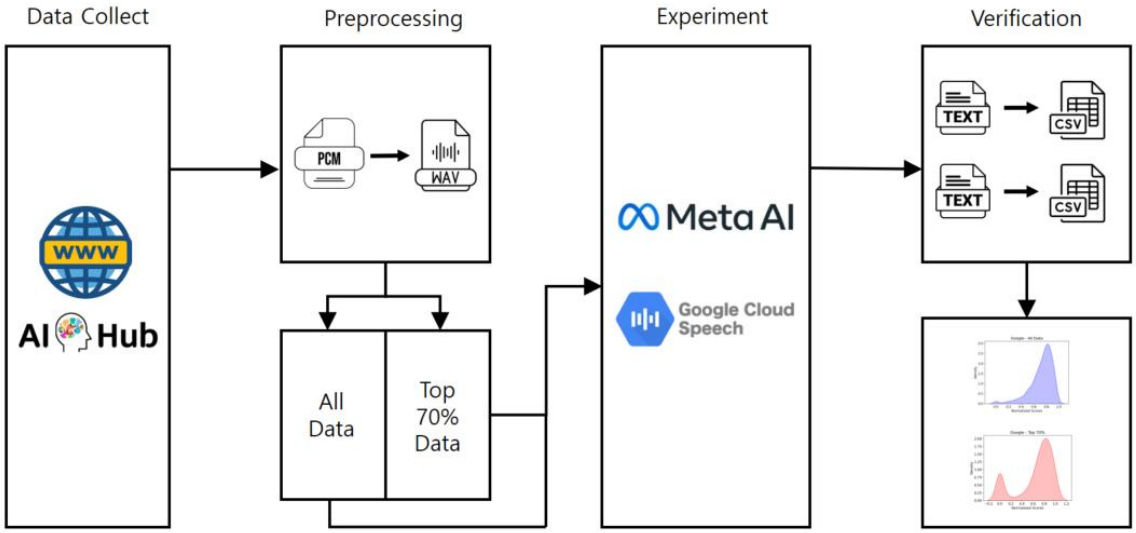

Fig. 1. System Overview

스트리밍 인식은 오디오 캡처 중에 중간결과를 제공하므로, 사용자가 계속 말하는 중에도 결과를 표시할 수 있다.

III. The Proposed Scheme

1. System Overview

본 연구에서는 그림 1과 같이 데이터 수집, 전처리, 실험, 검증으로 나뉜다. 데이터 수집 단계에서는 실험에 사용할 데이터셋을 수집한다. 전처리 단계에서는 수집한 데이터셋들의 SNR을 계산한 후, SNR 값 중 상위 70%을 기준으로 데이터셋을 나누었다. 수집한 데이터셋은 전부 pcm(Pulse Code Modulation) 파일인데, 실험을 위한 모델에서 pcm 파일을 지원하지 않아 모두 wav(Waveform audio format) 파일로 변환하였다. 실험 단계는 SNR 값 상위 70%의 데이터셋과 전체 데이터셋으로 나누어서 Meta AI에서 지원하는 모델인 Seamless M4T와 Google Cloud Speech의 Speech-to-Text 모델에 적용하였다. 검증 단계는 각 모델을 통해 생성된 text 파일을 Levenshtein Distance를 통해 두 모델의 정확도를 서로 비교해 보았다.

2. System Environment

본 연구에서 사용한 서버 및 장비의 시스템 환경은 모델의 실험과 성능 평가를 위해 표 2와 같이 구축되었다.

Table 2. System Environment

Item	Value
OS	Window 10 Pro 64bit
CPU	Intel i9-12900
Memory	64.0GB
GPU	NVIDIA GeForce RTX 3080
Python	Python 3.9
Library	Levenshtein, AduioSegment, Speech_v1p1beta1

3. Speech Data Collection and Utilization

음성 데이터는 여러 소스로부터 수집되었다. 연구의 신뢰성과 일반화 가능성을 확보하기 위해 다양한 데이터를 수집하였다. 수집된 데이터는 표 3과 같다.

Table 3. List of Collected Datasets

Sources	Data Name	Time(About)
AiHub	한국어 음성	193 hours
Zenodo	Zenodo 환경음	22.6 hours
Blogs	기타 환경음	0.6 hours
AiHub	소음 환경 음성인식 데이터	703 hours

대부분의 데이터는 AiHub에서 수집하였다. AiHub는 AI 기술 및 제품/서비스 개발에 필요한 AI 인프라(AI 데이터, AI SW API, 컴퓨팅 자원)를 지원함으로써 누구나 활용하고 참여하는 AI 통합 플랫폼이다[23]. 실험에서는 AiHub 사이트에서 제공되는 한국어 음성, 환경 음성 데이터 중에서 한국어 음성 데이터셋과 소음 환경 음성인식 데이터셋을 사용하였다. 한국어 음성 데이터셋은 실내 환경에서 녹

음된 음성이며, 소음 환경 음성인식 데이터셋은 실제 환경에서의 소음 및 간섭이 포함되어 있다. 따라서 이 데이터셋들은 높은 SNR에서부터 낮은 SNR까지 다양한 특성을 가지며 음성인식 모델의 성능을 평가하기에 알맞다.

Zenodo는 과학 및 연구 분야에서 데이터와 코드를 저장하고 공유하는 플랫폼이다. Zenodo에서는 공개된 환경음 데이터를 수집하였다. 이 자료는 다양한 생활 환경 및 자연환경 소리 등의 여러 다양한 환경음을 포함하는 특징을 가지고 있다.

기타 블로그 및 온라인 자료에서도 데이터를 수집하였다. 이러한 데이터는 실제 환경에서 녹음된 환경 소리를 포함하고 있어 다양한 환경에서 음성 인식 모델의 성능을 평가하는 데 유용하다.

다양한 데이터 소스를 활용하여 데이터셋을 수집했지만 AiHub에서 다운로드한 데이터셋 외에 다른 데이터셋에는 전사문이 없어 정확한 정답을 비교하기 어렵다는 단점이 존재했다. 따라서 실험에는 AiHub에서 다운로드한 데이터셋만 이용하였다.

4. Preprocessing

실험을 시작하기에 앞서, 몇 가지 준비 단계가 필요하였다. 우선 Google Cloud Speech-to-Text는 전처리 과정이 필요하진 않지만, 동일한 조건에서의 실험을 위해 Seamless M4T와 동일한 전처리 과정을 추가했다. 연구에 사용한 데이터셋은 한 개 문장에 해당하는 pcm 음성파일과 txt 전사문으로 구성되어 있다. 정확한 결과를 비교하기 위해 데이터에 있는 전사 규칙에 의해 쓰인 전사 기호나 문장 부호를 표 4와 같이 삭제하였다. 또한 Seamless M4T는 pcm 파일을 지원하지 않으므로 모든 pcm 파일을 wav 파일로 변환하였다.

Table 4. Comparison Before and After Transcription Symbols and Punctuation Removal

Before Removal	After Removal
b/ 빨리 가야 돼.	빨리 가야 돼.
o/ b/ 네+ 네. b/	네 네.
n/ 있잖아. n/ 그/ 너 해외여행 간 적 있어?	있잖아. 그 너 해외여행 간 적 있어?

또한 SNR 값에 따른 정확도를 비교하기 위해 모든 wav 파일에 대해 SNR 값을 계산하였다. 그림 2는 데이터셋의 SNR 값을 csv 파일로 정리한 결과이다.

	A	B		A	B
1	File Name	SNR (dB)	1	File Name	SNR (dB)
2	KsponSpeech_000001.wav	33.903	2	KsponSpeech_028002.wav	32.586
3	KsponSpeech_000002.wav	40.435	3	KsponSpeech_028004.wav	41.705
4	KsponSpeech_000003.wav	47.739	4	KsponSpeech_028005.wav	38.75
5	KsponSpeech_000004.wav	39.293	5	KsponSpeech_028006.wav	41.396
6	KsponSpeech_000005.wav	33.801	6	KsponSpeech_028007.wav	35.363
7	KsponSpeech_000006.wav	41.913	7	KsponSpeech_028008.wav	48.137
8	KsponSpeech_000007.wav	34.687	8	KsponSpeech_028009.wav	37.271
9	KsponSpeech_000008.wav	31.506	9	KsponSpeech_028010.wav	31.057
10	KsponSpeech_000009.wav	32.696	10	KsponSpeech_028011.wav	41.668

Fig. 2. Distribution of SNR in the Dataset

5. Experiment

전처리된 데이터셋은 Seamless M4T와 Google Cloud Speech-to-Text API를 사용하여 각각 STT 작업을 수행하였다. 작업을 수행하고 난 뒤 저장된 텍스트 파일과 AiHub에서 수집한 데이터셋에 있는 전사문을 각각 비교하였다. 비교하는 방법은 Levenshtein Distance 알고리즘을 사용하였다. 그림 3은 Aihub에서 제공한 전사문과 Google Cloud Speech to Text의 결과를 비교한 그림이다.

A	B
answer	google
아니면 캐슬에서 라면.	아니면 캐슬에서 라면
낮에 쪼끔 봤으면은.	낮에 조금 봐 쓰면은
아 그 공모전? 공모전. 인스타는 라이프 플러스.	아 그 공모전 공모전 인스타는 라이크 플러스
들어봤는데? 그거 그 일제시대 때 그 사람이지?	들어가는데 그거 그 일제시대 때 그 사람이
아니. 나가는데 너무 늦어.	많이 남았는데 너무 늦어
아 그럼 뒷자린 잘 안 들려. 그럼 인제 환불 들어가고.	그럼 뒷자리 잘 안 들려 그럼 이제 환불 들어가고
응 지금 이론이니까.	응 지금 일요일이니까

Fig. 3. CSV Generated Based on Experimental Results

IV. Verification

Table 5. Comparison of Average Levenshtein Distance for Each Model Based on SNR Values in the Dataset

Model	All Data	Top 70% SNR Data
Seamless M4T	16.6481	13.6306
Googld Cloud Speech-to-Text	8.3870	10.0278

실험 결과를 csv 파일로 정리하고 비교한 결과, 표 5에서 알 수 있듯이 Seamless M4T 모델에서는 SNR이 상위 70%인 데이터의 Levenshtein Distance 값이 평균 3.0175점 낮았다. 이는 SNR 값이 높을수록 음질이 더 좋다는 것을 나타내기 때문에 SNR 상위 70%의 데이터를 텍

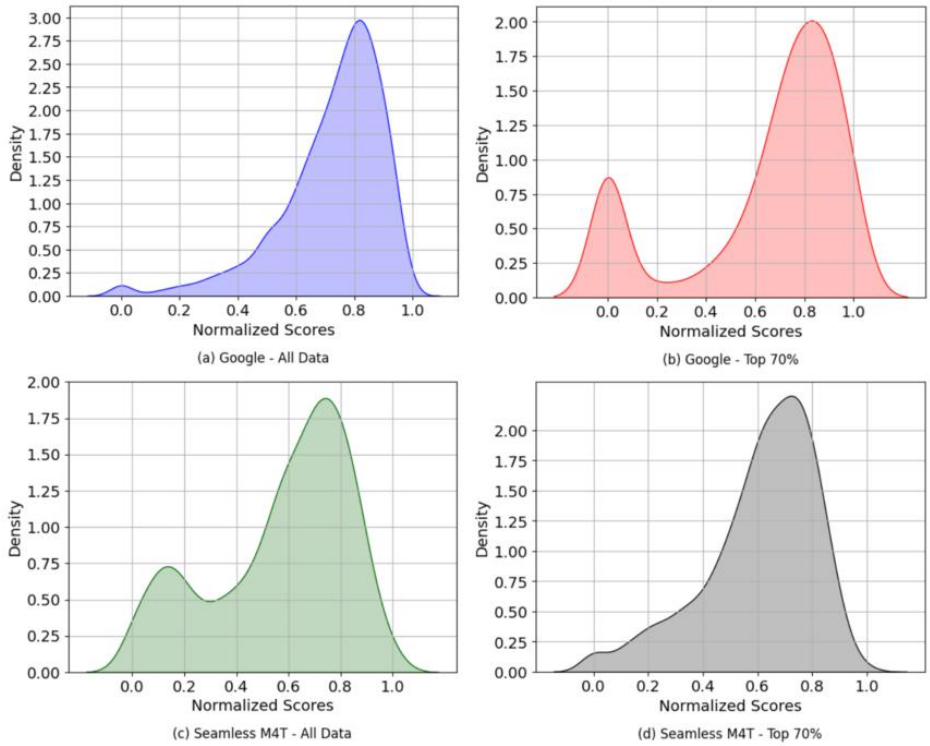

Fig. 4. Comparison of Density Plots for Each Model

스트로 변환할 때 Seamless M4T 모델의 결과가 좋게 나온 것으로 해석된다.

반면 Google Cloud Speech-to-Text의 결과는 SNR 상위 70%의 데이터의 Levenshtein Distance 값이 1.6408점 높았다. 그림 4는 Seamless M4T 모델과 Google Cloud Speech to Text에 전체 데이터셋과 SNR 상위 70%의 데이터셋으로 STT를 진행했을 때, 각 실험의 Levenshtein Distance 결과를 Density Plot을 통해 확인한 그림이다. 표 6은 그림 4에서 각 모델의 Density Plot의 평균값을 정리한 표이다. 표와 그림을 보면 Google Cloud Speech-to-Text 모델은 SNR 상위 70%의 데이터가 전체 데이터셋 보다 Levenshtein Distance의 평균 점수가 0.0762 높은 것을 확인할 수 있으며, Seamless M4T 모델은 전체 데이터셋의 Levenshtein Distance의 평균 점수가 0.0382 높은 것을 확인할 수 있다.

Table 6. Average Vaule of Density Plot data in Figure 4

Model	All Data	Top 70% SNR Data
Seamless M4T	0.5716	0.6098
Googld Cloud Speech-to-Text	0.7231	0.6469

Google Cloud Speech-to-Text 모델은 데이터 로깅을 통해 사용자가 사용한 음성 data를 기록하여 학습시킴으로써 음성인식 서비스를 개선한다. 또한 지속적으로 추가 학습을 진행하기 때문에 최근에 출시한 Seamless M4T 모델보다 Google Cloud Speech-to-Text 모델의 점수가 8.2611 더 좋은 것으로 추론할 수 있다.

V. Conclusions

본 연구는 SNR이 STT 성능에 미치는 영향을 조사하기 위해 2023년 8월에 새롭게 공개된 Meta 사의 Seamless M4T 모델과 가장 널리 알려진 Google Cloud Speech to Text 모델을 Levenshtein Distance를 통해 비교하였다.

실험 결과로 높은 SNR을 가진 상위 70% 데이터에서는 Seamless M4T 모델이 13.6306으로 낮은 Levenshtein Distance 값을 보였다. 이 결과는 높은 SNR 데이터에서 모델이 효과적으로 작동했다는 것을 의미한다. 또한 Google Cloud Speech-to-Text 모델은 동일한 데이터에서 10.0278로 높은 Levenshtein Distance 값을 나타냈다. Google Cloud Speech-to-Text 모델은 데이터 로깅을 통해 사용자가 사용한 음성 데이터를 기록하여 학습시킴으로써 음성인식 서비스를 개선한다. 그렇기 때문에 최근에 공개된 모델인 Seamless M4T 모델보다 Google

cloud Speech-to-Text 모델이 더 인식률이 높게 나온다고 해석할 수 있다. 두 모델의 차이로 모델 특성에 따른 훈련 데이터의 선택이 중요하며, 여러 음성인식 모델의 데이터셋 인식정도에 대한 연구가 필요하다.

실험 결과를 비교해 보았을 때 Seamless M4T 모델은 Google Cloud Speech to Text 모델보다 Levenshtein Distance가 전체 데이터셋에서는 8.2611, SNR 상위 70% 데이터셋에서는 3.6028 정도 성능이 떨어지지만, 1분에 0.016달러의 비용을 청구하는 Google Cloud Speech to Text와는 다르게 Seamless M4T 모델은 무료라는 장점이 있다. 비용적인 측면을 생각한다면 새로운 STT 모델인 Seamless M4T가 Google Cloud Speech to Text 모델과 비교해도 충분히 연구에 활용될 수 있을 것이다.

이번 연구에서는 이미 공개된 훈련이 되어 있는 STT 모델만을 사용하여 결과를 비교하였다. 앞으로의 연구에서는 훈련 단계에서 SNR의 영향을 더 깊게 조사하고, 모델의 정확도를 향상시키기 위한 방법을 모색할 예정이다. 또한 이러한 결과는 음성인식 모델을 선택할 때 SNR 수준을 신중하게 고려해야 함을 강조한다.

ACKNOWLEDGEMENT

This work was supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(IITP-2024-RS-2022-00156334)

REFERENCES

[1] Chandolikar, N., Joshi, C., Roy, P., Gawas, A., & Vishwakarma, M. "Voice Recognition: A Comprehensive Survey," 2022 International Mobile and Embedded Technology Conference (MECON). pp. 45-51. Noida, India. March 2022. DOI: 10.1109/MECON53876.2022.9751903

[2] Chen, J., Wang, Y., & Wang, D. "A Feature Study for Classification-based Speech Separation at Low Signal-to-Noise Ratios," IEEE/ACM Transactions on Audio, Speech, and Language Processing. Vol. 22. No. 12. p. 1993-2002. September 2014 DOI: 10.1109/TASLP/2014.2359159

[3] Toscano, J. C., & Toscano, C. M. "Effects of Face Masks on Speech Recognition in Multi-talker Babble Noise," PloS one 16.2 February 2021. DOI: e0246842

[4] Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & Ritter, M. "Audio set: An Ontology and Human-labeled Dataset for Audio Events," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 776-780. New Orleans, LA, USA. March, 2017. DOI: 10.1109/ICASSP.2017.7952261

[5] Yeonsoo, L., Deokjin S., Jeong-sik, P., Yuchul. J., "An Automatic Data Construction Approach for Korean Speech Command Recognition," Journal of The Korea Society of Computer and Information Vol.24. No. 12. pp. 17-24, December 2019. DOI: 10.9708/jksci.2019.24.12.017

[6] Hirsch, H. G., & Pearce, D. "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," Automatic Speech Recognition: Challenges for the New Millenium ISCA ITRW ASR2000. Paris, France. September, 2000. DOI: 10.21437/ICSLP.2000-743

[7] Yejin L., Myungjin Son., Juhee Kim., SungWoo Byun., Seokpil Lee., “A Comparison of the Performance of Noise Cancellation Methods for Improving Speech Recognition Accuracy in Noisy Environment” Information and Control Symposium, pp. 257-258, Busan, Korea, 2020.

[8] Bokyoung Kim, Seongbae Lee, and Kyuheon Kim, "Deep Learning-based Filter for Speech Separation to Enhance STT Performance," Proceedings of Symposium of the Korean Institute of communications and Information Sciences, pp. 157-158, Gyeongju, Korea, November 2022.

[9] Seung Gwan L., and Sangmin L., "Data Augmentation for DNN-based Speech Enhancement," Journal of Korea Multimedia Society, Vol. 22, No. 7, pp. 749-758, 2019.

[10] Jiwon L., Yonghae H., and Kyuheon K., "Noise Filtering Method Based on Voice Frequency Correlation to Increase STT Efficiency," a collection of papers from The Korean Institute of Broadcast and Media Engineers academic presentation , pp. 176-179, 2021.

[11] Byung Hee K., and Dong Kun N., "A Deep Learning based Speech Quality Enhancement Scheme Using Environmental Sound Classification and Location Information," Journal of KIISE, Vol. 50, No. 4, pp. 344-350, 2023. DOI: 10.5626/JOK.2023.50.4.344

[12] Hou, J, C., Wang, S. S., Lai, Y. H., Taso, Y., “Audio-Visual Speech Enhancement using Multimodal Deep Convolutional Neural Networks”., IEEE transactions on Emerging Topics in Computational Intelligence

[13] Youngmi P., and Chulyun K., "A Study on the Application of Language Model to Improve Speech Recognition Accuracy," Proceedings of the Korean Information Science Society Conference, pp. 287-289, Jeju, Korea, December 2021.

[14] Joris C., Manuel P., Samuele C., Antonie D., Emmanuel V., “LibriMix: An Open-Source Dataset for Generalizable Speech Separation”, arXiv preprint arXiv:2005.11262., 2020, DOI: 10.48550/arXiv.2005.11262

[15] Jungyoon C., Haeyoung G., Soobin O., and Jongwoo L., "CCVoice: Voice to Text Conversion and Management Program Implementation of Google Cloud Speech API," KIISE Transactions on Computing Practices, Vol. 25, No. 3, pp. 191-197, 2019. DOI: 10.5626/KTCP.2019.25.3.191

[16] Radford A., JW Kim., Xu T., Brockman G., “Robust Speech Recognition via Large-Scale Weak Supervision”, International conference on Machine Learning, pp.28492-28518, April 2023. DOI: 10.48550/arXiv.2212.04356

[17] Don H. Johnson "Signal-to-Noise Ratio," Scholarpedia 2006 DOI: 10.4249/scholarpedia.2088

[18] Microsoft, What is Speech to Text? https://learn.microsoft.com/en-us/azure/ai-services/speech-service/speech-to-text

[19] Levenshtein, Vladimir I. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals," Soviet Physics Doklady. Vol. 10. No. 8. February 1966.

[20] Max Bachmann, Python-Levenshtein, https://github.com/maxbachmann/Levenshtein

[21] Barrault, Loïc, et al. "SeamlessM4T-Massively Multilingual & Multimodal Machine Translation," Meta, August, 2023. DOI: 10.48550/arXiv.2308.11596

[22] Google, Google Cloud Speech-to-Text, https://cloud.google.com/speech-to-text?hl=ko

[23] Aihub, Aihub Introduce, https://aihub.or.kr/intrcn/intrcn.do?currMenu=150&topMenu=105

Authors

Yeong-Jin Kim received the B.S degree in cyber security from Pai Chai University, Daejeon, South Korea, in 2023. He currently pursuing M.S degree in Smart ICT Convergence at Pai Chai University. He is interested in artificial intelligence, malware, and voice recognition.

Hyun-Jong Cha received the M.S. and Ph.D. degree in Compuger science and Defense Acquisition Program from Kwangwoon University, South Korea, in 2008 and 2014. He is a professor in the Department of Software Engineering at Pai Chai University in Daejeon, South Korea. His current research interests include information security, artificial intelligence, IoT, and blockchain.

Ah Reum Kang received the M.S. and Ph.D. degrees in information security from Korea University, South Korea, in 2012 and 2016. She is a professor in the Department of Information Security at Pai Chai University in Daejeon, South Korea. Her current research interests include security, artificial intelligence, malware, medical data analysis, and online game security.