Asia-pacific Journal of Convergent Research Interchange
Vol.9, No.5, May 31 (2023), pp.105-116
http://dx.doi.org/10.47116/apjcri.2023.05.09

Feature Extraction and Analysis of AI-Based Speech Signals for Auditory Rehabilitation

청능재활을 위한 AI 기반 음성신호의 특징 추출 및 분석

Seunghan Ha[1], Sangdo Lee[2]

하승한[1], 이상도

[1] Associate Professor, Department of Nursing Science, Chungbuk Health & Science University, Republic of Korea, shha7774@chsu.ac.kr

[2] Professor, Department of Welfare & Management, Korea National University of Transportation, Republic of Korea, sdlee@ut.ac.kr

Corresponding author: Sandgo Lee

Abstract: The purpose of this study was to extract and analyze features of speech signals using artificial intelligence techniques, with the goal of improving auditory rehabilitation outcomes. Wav2vec 2.0 was used to identify phonemes in 2,000 sound files recorded from individuals with hearing impairment. The average speech intelligibility score was calculated to be over 0.92 %, determined by calculating the difference between a reference sentence and the sentence obtained through speech-to-text (STT). It was possible to distinguish individual phonemes accurately from recorded sound files of speech produced by individuals with hearing impairments, and to measure the intelligibility of their speech by assessing their speech clarity. Through this experiment, we confirmed the potential to distinguish phonemes between individuals with normal hearing and those with hearing impairment, as long as the phonemes are similar to the reference sentence for individuals with normal hearing. Furthermore, it was confirmed that it is possible to differentiate the accuracy of pronunciation in speech produced by individuals with hearing impairments, which further supports the potential for assessing speech intelligibility in this population through phonetic analysis. To validate the effectiveness of artificial intelligence-based speech signal feature extraction and analysis for auditory rehabilitation, a comparative study on phoneme extraction between individuals with normal hearing and those with hearing impairment should be conducted.

Keywords: Auditory Rehabilitation, STT(Speech to text), Phoneme Extration, Speech Intelligibility, GER(Grapheme Error Rate), AI(artificial intelligence)

요약: 이 연구의 목적은 청능재활을 향상시키기 위해 인공지능 기술을 활용하여 음성 신호의 특징을 추출하고 분석하는 것이다. Wav2vec 2.0을 이용하여 청각장애인 2,000명으로부터 얻은 소리 파일에서 음소를 추출하고 말 명료도를 측정하였다. STT(Speech to Text)로 얻은 문장과 스크립트로 작성된 기준 문장의 차이를 말 명료도로 설정하였으며, 이를 기준으로 계산한 결과 평균 말 명료도 점수는 0.92% 이상이었다. 청각장애인의 발화를 통해 녹음된 소리 파일에서 음소를 정확히 구별할 수 있었으며, 청각장애인의 발화를 통해 녹음된 발음의 정확성을 확인하기 위해 말 명료도를 측정할 수 있었다. 이 실험을 통해, 비장애인의 음소가

Received: January 16, 2023; 1st Review Result: March 02, 2023; 2nd Review Result: March 28, 2023
Accepted: April 30, 2023

기준 문장과 유사하다면 비장애인과 청각장애인 사이에서 음소를 구별할 수 있다는 가능성을 확인하였다. 또한, 청각장애인의 발화에서 발음의 정확성을 구분할 수 있는 가능성도 확인하다. 앞으로 청능재활을 위한 인공지능 기반의 음성 신호 특징 추출 및 분석의 유효성 검증을 위해 비장애인과 청각장애인을 대상으로 음소 추출 비교 연구가 진행되어야 한다.

핵심어: 청각재활,음성인식,음소추출,말 명료도,자소오류율,인공지능

1. 서론

최근 보건복지부에서 발표된 『2020년 장애인 실태조사』 에 따르면, 장애인의 인구는 264만명으로 전체인구 대비 5%를 차지하고 있으며, 이중 청각장애인은 전체 장애유형 중 지체장애인(45.1%) 다음으로 많은 비중(15.6%)을 차지하고 있다. 선천적 발달장애, 노화에 따른 청력손상, 청력 손상을 유발하는 외부 요인 등의 다양한 원인으로 인하여 2016년부터 매년 10,000 ~ 40,000명씩 증가하는 추세이며, 장애의 정도가 심한 청각장애인은 매년 1,500여 명씩 감소하고 있으나, 장애의 정도가 심하지 않은 청각장애인은 매년 12,000명 증가하여 보청기 등 청각보조기기와 청능재활을 필요로 하는 청각장애인구는 계속 증가하고 있다[1]. 청각장애로 인해 기능적, 사회적, 심리적인 영향을 겪은 난청인들에게 청력 손실로 인한 부정적 영향을 완화하고 보상하여, 보다 나은 의사소통을 할 수 있도록 청능재활의 환경을 제공한다는 것은, 자신감과 자아존중감을 회복하여 적극적인 사회활동 참여로 평등한 삶을 영위할 수 있게 한다는 측면에서 매우 중요하다. 따라서, 난청인들이 스스로 학습하고 훈련할 수 있는 청능재활 훈련이 필요하며, 이를 위해서는 발음 연습, 유창성 연습, 상호작용 등과 같은 부분에 대한 음성 신호의 정확한 특징 추출 및 분석이 기술개발이 선행되어야 한다[2][3]. 최근 기술의 발달로 인해 인공지능(AI) 기반의 음성인식(Speech recognition) 및 자연어처리(Natural language processing) 기술 등이 널리 사용되고 있다. 그러나, 청능재활을 위한 음성인식 기술의 한계 중 하나는 대화 소리에서 정확한 소음을 구분하여 추출하는 음 분리에 대한 부분이다. 이러한 제한 사항을 극복하기 위해서는 많은 청각장애인 대화 내용을 축적해 놓은 데이터를 활용한 딥러닝(Deep learning) 학습을 통해 인식률을 높이는 것이 필요가 있다[3-5]. 이전의 많은 연구들은 청각보조기기인 보청기 및 인공와우의 기기적 성능 향상의 측면에서 주로 잡음 제거 및 음성 왜곡 개선, 음성을 텍스트로 변환의 측면에서 딥러닝 학습을 결합한 형태의 음성인식 기술개발이 이루어져 왔다. 따라서, 청각장애인의 대화를 정확하게 이해하고 인식하는 능력을 향상시키기 위해 음 분리 기술과 딥러닝 학습을 결합한 음성인식 기술을 개선하는 연구가 절실히 필요하다[6-9]. 본 연구에서는 청각장애인의 청능재활을 위해 청각장애인의 음성 파일로부터 인공지능 기반의 딥러닝 기술인 Wav2vec 2.0 이용한 청각장애인의 음성 신호의 특징을 추출하고 분석하고자 한다[10][11].

2. 이론적 배경

2.1 음성신호 추출 및 분석기술

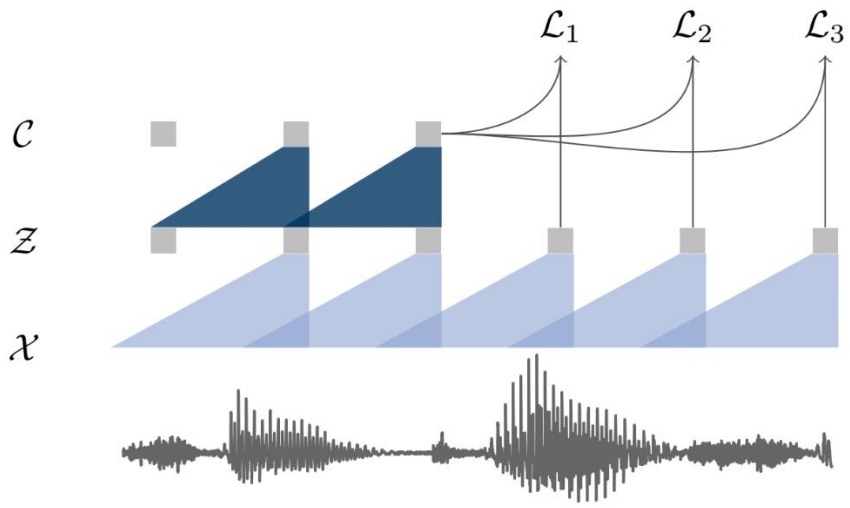
음성신호 특징추출 및 분석기술은 음성신호에서 유용한 정보를 얻기 위해 사용되는

중요한 기술이다. 이 기술은 음성인식, 음성분류, 화자인식, 음성합성 등과 같은 다양한 분야에서 응용된다. 최근 연구 동향을 살펴보면, 딥러닝을 활용한 음성신호 추출 기술이 많이 연구되고 있다. 특히, 컨볼루션 신경망(CNN : Convolutional Neural Network)과 재귀 신경망(RNN : Recurrent Neural Network)을 이용한 음성신호 추출 기술이 많은 연구 결과를 얻고 있다. 이 외에도, 주파수 분석, 선형 예측 분석, 멜 주파수 셋 변환 등의 기술도 여전히 연구되고 있다. 또한, 다중 채널 음성신호 추출 기술도 중요한 연구 분야 중 하나로 부상하고 있다. 이를 통해 음성신호의 공간 정보를 활용하여 음성인식의 정확도를 높이는 것이 가능하다[10]. 본 연구에서는 음성 신호의 특징 추출 및 분석기술로 Wav2vec 2.0을 활용하고자 한다[11]. Wav2vec은 음성 신호의 사전훈련(Pre-training) 및 음성인식과 같은 자연어 처리(NLP) 작업에 대한 전이학습(Transfer learning)을 위한 사용되는 모델이다. Wav2vec은 Facebook AI Research(Fair)에서 개발되었으며, 음성인식에서 뛰어난 결과를 보이고 있다[12][13]. Wav2vec은 기존의 음성 신호 처리 방식과는 다르게, End-to-end 기계 학습 방식을 사용한다. 즉, Wav2vec은 전체 오디오 신호를 입력으로 받아 바로 출력을 예측하는 방식으로 학습한다. 이를 위해 Wav2vec은 Wav2vec과 Wav2vec 2.0으로 두 가지 버전이 있다[12][13]. Wav2vec은 비지도 학습(Unsupervised learning) 방식으로, 대규모 음성 데이터 세트를 사용하여 자동 인코더 모델(Auto-encoder model)로 학습한다. 이 모델은 입력 오디오 신호를 수용하고 난 뒤, 오디오 신호를 재생성하려고 시도하면서 중간 표현값을 학습한다. 이 중간 표현값은 후속 음성처리 작업에서 유용한 정보를 추출할 수 있는 고차원 특징으로 사용된다[12][13]. Wav2vec 2.0은 Wav2vec에서 발전된 모델로, 준지도 학습(Semi-Supervised learning) 방식을 사용한다. 이 모델은 비지도 사전학습(Unsupervised pre-training) 단계에서 적은 양의 레이블링 된 데이터를 사용하여 파인튜닝(Fine-tuning)을 수행한다. Wav2vec 2.0은 단일 모델에서 다중 언어 및 다중 화자 처리가 가능하며, 뛰어난 음성인식 성능을 보이고 있다[11][14]. Wav2vec은 오디오에서 효과적인 특징 추출을 위해 컨볼루션 신경망과 트랜스포머 셀프 어텐션(Transformer self-attention) 메커니즘을 결합한다. 이를 통해 Wav2vec은 효과적인 특징 추출을 수행하고, 그 결과로 NLP 작업에서 높은 정확도를 보여주고 있다[11][14].

2.2 Wav2vec

레이블 없는 데이터로 자기 자신의 특성을 학습하는 것을 자기지도학습(Self-Supervised Learning)이라고 한다. Wav2vec은 신경망(NN: Neural network)를 기반으로 하여, 음성 오디오를 입력으로 받아 사전훈련(Pre-training)을 통해 모델을 최적화한 후, 입력된 음성 오디오 파일에서 특징을 추출하고, 추출된 샘플은 예측할 수 있다. Wav2vec은 모든 언어의 음성에서 특징을 추출할 수 있다. [그림 1]은 Wav2vec의 구조이다[12][13].

Wav2vec은 Encoder network와 Context network로 구성된다. χ는 음성 데이터가 입력되어진 영역으로 Encoder(CNN network)를 통해 특정 Vector로 변환된다. 변환된 벡터는 잠재(Latent) 영역 Z에 저장되어 학습이 진행된다. 음성 데이터의 공통성을 나타내는 상호의존정보(Mutual information)을 최대한 추출하기 위해 컨텍스트 영역(Context region) C가 존재한다. 이러한 구조를 통해 Wav2vec은 다양한 언어의 음성 데이터에서 공통된 특징을 추출할 수 있다[12][13].



[그림 1] Wav2vec의 구조

[Fig. 1] The Structure of Wav2vec

$$\cdot f: \chi \rightarrow Z \quad \text{Encoder network} \tag{1}$$

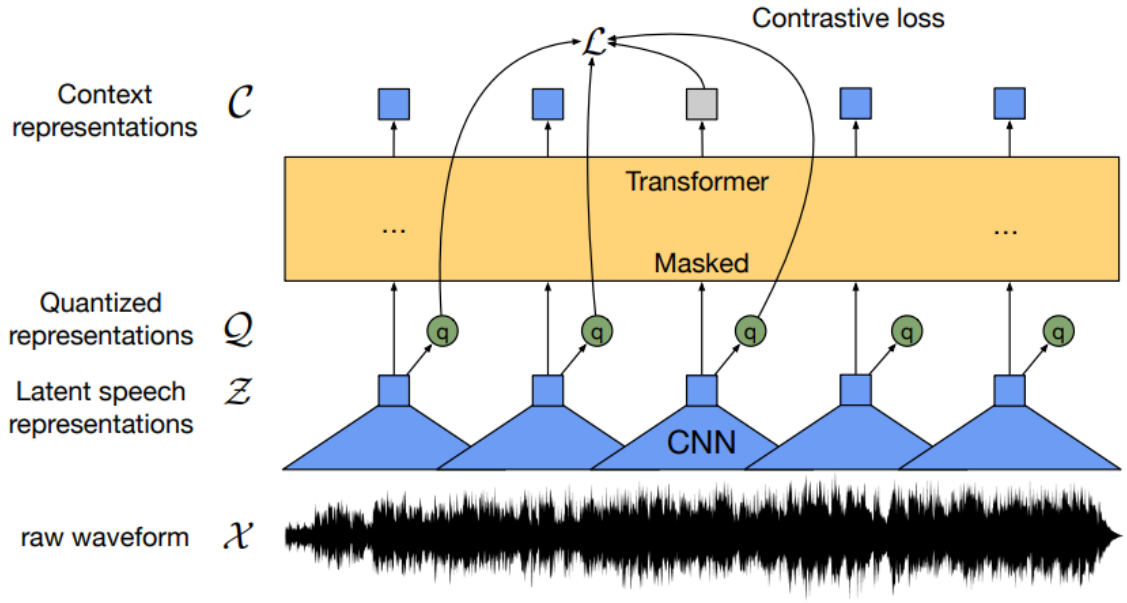
$$\cdot g: Z \rightarrow C \quad \text{Context network} \tag{2}$$

Wav2vec은 BERT의 마스크드 언어 모델(MLM: Masked Language Modeling) 및 대조적예측 코딩(CPC: Contrastive Predictive Coding) 기법에서 영감을 받아, 일반적으로 음향 모델(Acoustic model)에서 사용하는 MFCC(Mel-Frequency Cepstral Coefficients), 필터뱅크(Filter-bank), 스펙트로그램(Spectrogram) 등 별도의 특징 추출 과정을 대체하는 계층을 가진 심층 신경망 모델(DNN: Deep Neural Network model) 이다[12][13].

2.3 Wav2vec 2.0

Wav2vec 2.0은 트랜스포머(Transformer) 구조 기반의 심층신경망 음향 모델(DNN-based acoustic model)이다. Fig 2.은 Wav2vec 2.0의 구조를 나타낸다. 첫번째는 다층 합성곱 특징 인코더(Muli-layer convolution feature encoder) 단계로 $f: \chi \rightarrow Z$ 로 원시 음성 데이터(Raw)인 χ 를 T 시간 간격으로 임베딩하여 잠재된 음성 표현(Latent speech representation) $z_{1,\ldots,}z_T$ 로 변환한다. 이때, 층을 정규화하기 위해서 GELU(Gaussian Error Linear Units)를 사용한다. 두번째는 트렌스포머(Transformer) 단계로 $g: Z \rightarrow C$ 로 맥락화된 표현 $c_{1,\ldots,}c_T$ 로 변환한다. 이때, 위치 임베딩(Positional embedding)에 합성곱 계층(Convolutional layer)을 활용한다. 세번째는 양자화 모듈(Quantization module) 단계로 자기지도학습(Self-supervised training)을 위해 특징 인코더(Feature encoder) f 를 통과한 z 를 PQ(Product Quantization)를 수행한다. 이때, PQ의 서브벡터(Codebook)의 개수를 G라고 하고, 합성곱 신경망(CNN)을 통과한 임베딩 벡터 $z \in \mathbb{R}^{Vxd}$ 라고 가정할 때, V는 타임스텝(Time step)을 d는 합성곱 신경망을 통과한 임베딩 벡터의 크기를 의미한다. 이후, z로부터 크기가 $\mathbb{R}^{Vxd/G}$ 인 G개의 서브벡터를 만들고, PQ를 수행하

여 각 서브벡터의 데이터들을 클러스터링한다. 각 서브벡터에 PQ를 수행할 때, 서브벡터의 엔트리들이 코드북 내에서 정의된 센트로이드(Centroid) 중 가장 가까운 센트로이드의 데이터로 변환된다. 클러스러링 후 변환된 G개의 데이터들을 병합(Concat)하여 크기가 d인 벡터로 변화하고, 단순 선형변환(Linear transformation) $\mathbb{R}^d \rightarrow \mathbb{R}^f$ 을 적용하여 q를 얻는다. 이때, 순전파(Forward propagation)에서는 가장 가까운 센프로이드를 찾는 Argmax 함수를 활용하고, 역전파(Backward propagation)에서는 Gumbel softmax를 사용하여, 양자화(Quantization)을 거친 q를 최종적으로 loss 계산에 활용할 수 있다[11][14].



[그림 2] Wav2vec 2.0의 구조

[Fig. 2] The Structure of Wav2vec 2.0

$$\cdot f : \chi \rightarrow Z \quad \text{Multi-layer convolution feature encoder} \quad (3)$$

$$\cdot g : Z \rightarrow C \quad \text{Transformer} \quad (4)$$

$$\cdot Z \rightarrow Q \quad \text{Quantization module} \quad (5)$$

Wav2vec 2.0은 사전훈련(Pre-training)과 파인튜닝(Fine-tuning) 단계로 구성된다. 첫번째로 사전훈련에서는 마스킹(Masking)이 BERT(Bidirectional Encoder Representations from Transformers)와 비슷한 방식으로 진행되며, 일부 타임스텝에 해당하는 부분의 잠재된 음성 표현 z를 마스킹하고, 남은 부분으로 마스킹한 부분의 양자화된 표현을 유추하는 방식을 사전학습을 진행한다[11][14]. Wav2vec 2.0에서 사용되는 사전훈련 손실함수 L은 다음과 같이 정의한다[11].

$$\cdot \ \mathscr{L} = \mathscr{L}_m + \alpha * \mathscr{L}_d \tag{6}$$

여기서, $\mathscr{L}_m$은 대조손실(Contrastive loss), $\mathscr{L}_d$는 다이버시티 손실(Diversity loss), α는 하이퍼파리미터를 의미한다. 대조손실($\mathscr{L}_m$)은 마스킹된 타임스텝으로 양자화 표현을 유추할 때 활용하는 손실이며, 트렌스포머를 통과한 벡터 c가 양자화된 벡터 q와 잘 일치하면 손실함수(Loss)값이 감소하는 형태이다[11][14].

$$\mathcal{L}m = -\log \frac{\exp(sim(\mathrm{Ct},\mathrm{qt})/k)}{\sum_{q\sim Qt}\exp(sim(\mathrm{Ct},\mathrm{q})/k)} \tag{7}$$

여기서, sim은 코사인 유사도(Cosine similarity) 함수를 의미한다. 다이버시티 손실($\mathscr{L}$d)은 코드북의 엔트리(Entry)들이 균일하게 활용될 수 있도록 손실함수 값이 감소하는 형태이다. 양자화된 음성 표현에 활성화 함수(Softmax 함수)를 씌운 것을 Pg,v로 정의하고, 이를 활용하여 다이버시티 손실($\mathscr{L}$d)을 다음과 같이 정의한다[11][14].

$$\mathcal{L}d = \frac{1}{GV}\sum_{\mathrm{g}=1}^{G} - H(P\mathrm{g}) = \frac{1}{GV}\sum_{\mathrm{g}=1}^{G}\sum_{v=1}^{V} P\mathrm{g},v \log P\mathrm{g},v \tag{8}$$

두번째로 파인튜닝에서는 사전훈련 된 모델에 추가적인 선형 레이어(Linear layer)를 컨텍스트 네트워크 위에 붙이는 등 원하는 작을 수행할 수 있도록 모델을 구성한다. 파인튜닝 과정에서는 음성인식, 음성분류 등 다양하게 정의할 수 있는 하위 과제(Downstream task)에 따라, 하위 과제를 푸는데 적절한 형태로 추가 계층 및 전용 손실 함수를 정의한다. 이때 양자화 모듈은 사용되지 않는다[11][14].

3. 연구방법

3.1 연구대상

본 논문에서는 청각장애인의 음성신호를 추출하고, 분석하기 위한 음성 데이터는 한국지능정보사회진흥원에서 인공지능 학습용 데이터로 제공되는 AI-Hub 사이트의 자료를 사용한다. 사용된 데이터는 청각장애인 2,000명으로부터 수집된 2,000개의 데이터로 구성되어 있다. 본 연구에서는 대량의 음성 데이터를 사용하여 제안된 단위를 사용한 인식 실험의 신뢰성을 높인다. 데이터는 .wav 포맷의 파일로 수집하며, 음성의 특징은 Wav2vec 2.0을 이용하여 음성신호로부터 추출한다. 모든 평가 데이터는 Wav2vec 2.0 모델 학습 및 새로운 단위 구축 실험에 중복되어 사용하지 않는다.

3.2 측정도구

본 연구에서는 청각장애인의 음성신호를 텍스트로 변환(STT, Speech to Text)시켜 주는 인공지능 기반의 방법으로 wav2vec 2.0을 사용한다. Wav2vec 2.0을 이용하여 추출된 음소(Phone)의 정보를 활용하여 청자의 입장에서 화자의 말소리를 알아들 수 있는 정도를

의미하는 말 명료도(Speech intelligibility)를 분석한다. 이때 말 명료도는 정량적으로 측정할 수 있도록 수식화 한다. 말 명료도를 분석하기 위해 목표문장을 대조군인으로 설정하고 이를 기준으로, 인식된 문장을 실험군으로 두어 대조군과 실험군 사이의 자소오류율(GER, Grapheme Error Rate)을 계산한다[15][16]. 말 명료도를 계산하는데는 자소오류율을 사용한다. 자소오류율은 음성 인식기의 성능평가를 위해서 흔히 사용하는 단어오류율(WER, Word Error Rate), 낱말오류율(CER, Character Error Rate)과 같은 평가지표를 자소에 적용한 것으로, 그 원리는 단어오류율, 낱말오류율과 같다. 자소오류율은 식 (9)과 같이 계산한다[15][16]. $\boldsymbol{n0}$는 성 변환 대상 글자 중 특수기호(띄어쓰기, 느낌표, 물음표 등 문장 부호)를 제외한 글자의 수이며, $\boldsymbol{ne}$ 는 잘못 변환된 글자 수를 의미하며 원본의 내용이 빠져 있거나 잘못 변환된 글자, 원본에는 없는 글자가 추가된 글자를 의미한다.

$$\boldsymbol{Error\ rate = \frac{n0 - ne}{n0} \times 100} \tag{9}$$

자소오류율은 오류가 얼마나 많은지 나타내는 지표이므로 말 명료도는 식 (10)과 같이 1에서 자소오류율을 뺀 값으로 정의한다. 자소오류율은 1을 초과하는 값이 존재할 수 있으므로 이 경우에는 말 명료도는 0으로 계산한다.

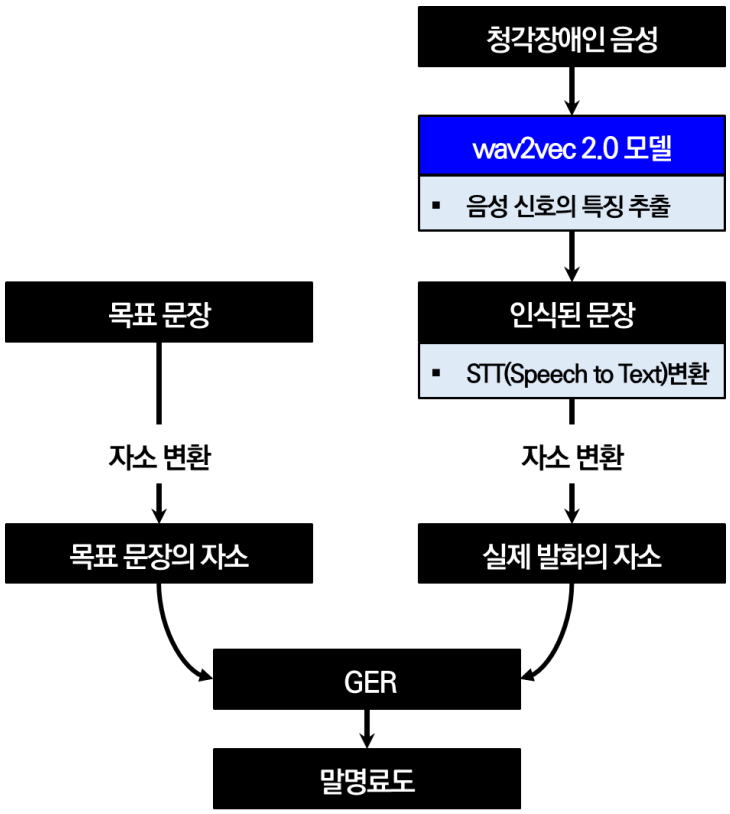
$$\boldsymbol{GER(Grapheme\ Error\ Rate) = \begin{cases} 1 - GER & if\ GER \le 1 \\ 0 & if\ GER > 1 \end{cases}} \tag{10}$$

예를 들어‘자전거'를‘자전거’라고 정확하게 발음한 경우, 이를 자소로 변환하면 ‘ㅈ, ㅏ, ㅈ, ㅓ, ㄴ, ㄱ, ㅓ’가 되고, N은 6, S, D, I는 모두 0이므로 GER은 0, 말 명료도는 1이 된다. 만약에 ‘다던거'로 발음했다면, 자소는 ‘ㄷ , ㅏ, ㄷ, ㅓ, ㄴ, ㄱ, ㅓ’가 된다. 목표한 단어는 같으므로 N은 6으로 동일하고, S는 2, D와 I는 0으로 GER=0.333, 말 명료도는 0.667이 된다.

3.3 연구방법

Wav2vec 2.0을 이용한 말 명료도 분석은 3단계를 걸쳐서 수행하였다.

첫번째 단계에서는 인공지능 모델인 Wav2vec 2.0을 이용하여 .wav 포맷으로 획득된 청각장애인의 음성파일로부터 음소를 추출한다. 두번째 단계에서는 음성파일로부터 추출된 음소를 STT(Speech to text)로 변환하여 인식된 문장을 완성한다. 세번째 단계에서는 청각장애인이 원래 말하고자 했던 문장을 목표문장으로하여 대조군으로 설정하고, 청각장애인의 음성신호로부터 인식된 문장을 실험군으로 설정한다. 네번째 단계에서는 대조군과 실험군으로부터 자소변환을 통해 각각의 자소를 추출하였으며, 이 둘의 자소를 비교하여 자소오류율을 계산한다. 이를 통해 말 명료도를 측정한다. [그림 3]은 청능재활을 위한 청각장애인의 음성신호 특징 추출 및 분석과정을 나타낸다.



[그림 3] 청능재활을 위한 청각장애인의 음성신호 특징 추출 및 분석과정

[Fig. 3] The Block Diagram of Extracting and Analyzing Characteristics of Speech Signals for Auditory Rehabilitation of People with Hearing Impairments

4. 연구결과

본 논문에서는 청각장애인이 원래 발화하고자 했던 목표문장을 대조군으로 설정하였으며, 청각장애인이 발화한 것으로부터 인식된 문장을 실험군으로 설정하여 말소리의 자소를 측정하였다. 대조군의 경우, 목표문장에서 자소변환하여 목표문장의 자소를 추출하였다. 실험군의 경우, 3단계를 거쳐 자소를 추출하였다. 첫번째 단계에서는 Wav2vec 2.0을 이용하여 청각장애인으로부터 획득한 2,000개의 .wav 포맷 파일에서 음성신호의 특징된 음소를 추출하였다. 두번째 단계에서는 음성신호를 안정적으로 인코딩하여, 음성신호를 문장으로 변환할 수 있는 STT로 추출된 음소 변환하여 인식된 문장으로 완성하였다. 세번째 단계에서는 추출된 음성신호로부터 자소를 추출하였다. 대조군의 경우, Wav2vec 2.0으로 추출된 음소를 변환하여 인식된 문장에서 문장을 구성하는 각 음절을 자모단위로 분리하였다. 이후, 분리된 자모단위에서 자소단위로 변환하는 과정인 자소변환을 통해 자소를 측정하였다. 마지막으로 이 자소와 비교 검증하기 설정된 실험군과 대조군의 자료로 자소오류율을 계산하였다. 이를 계산된 자소오류율을 이용하여 말 명로도를 분석하였다. [그림 4]는 말 명료도를 분석하기 위한 소스 코드이고, [그림 5]는 말 명료도의 분석 결과이다.

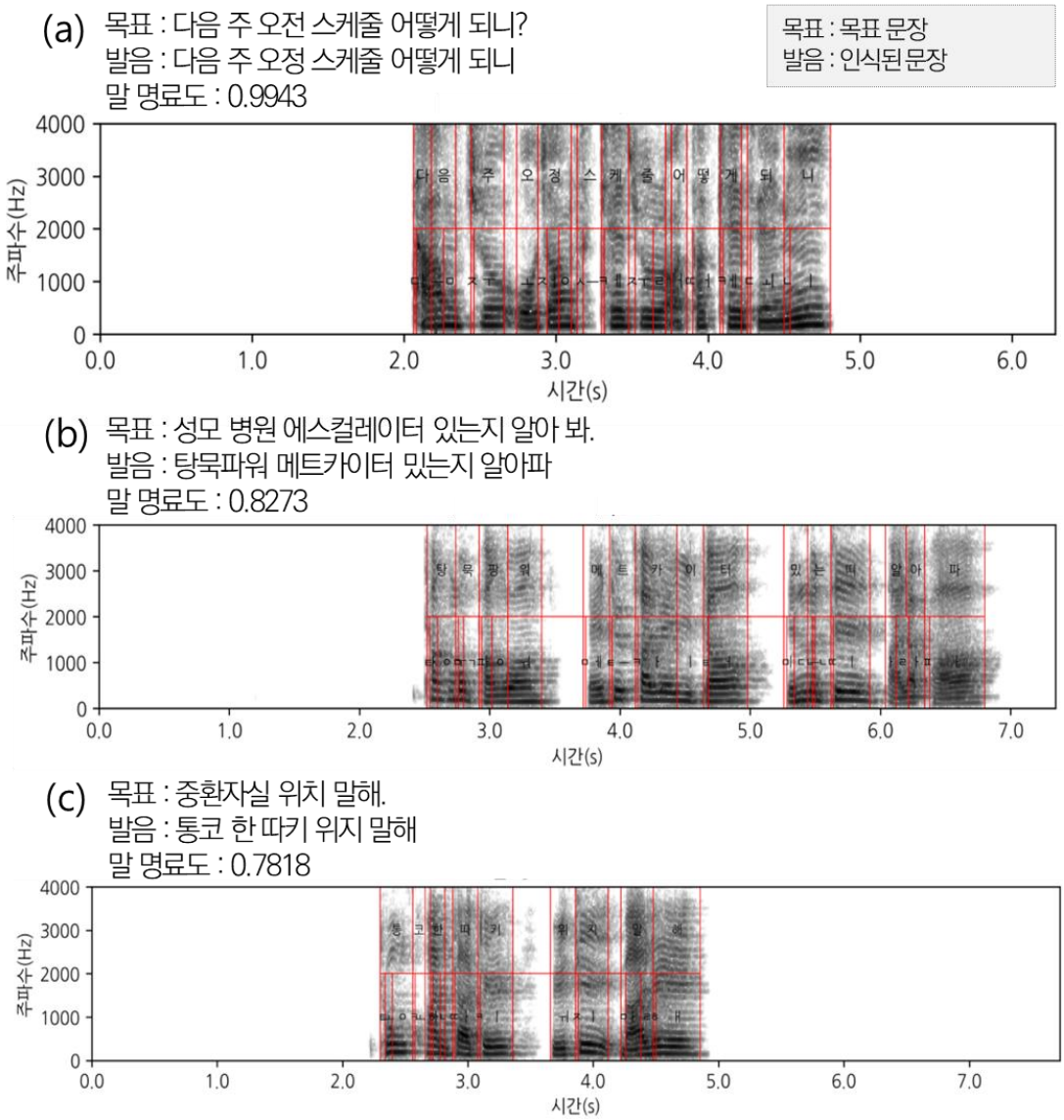
[그림 5]의 (a)에서 목표문장인 '다음 주 오전 스케줄 어떻게 되니?'를 기준으로 발음 문장인 '다음 주 오정 스케줄 어떻게 되니'의 말 명료도를 측정한 결과는 0.9943이었다.

[그림 5]의 (b)에서 목표문장인 ‘성모 병원 에스컬레이터 있는지 알아 봐.’를 기준으로 발음문장인 ‘탕묵파워 메트카이터 밌는지 알아파’의 말 명료도를 측정한 결과는 0.8273이었다. [그림 5]의 (c)에서 목표문장인 ‘중환자실 위치 말해’를 기준으로 발음문장인 ‘통코 한 따키 위치 말해’의 말 명료도를 측정한 결과는 0.7817이었다. [그림 5]의 (a), (b), (c)에서 청각장애인의 발화를 통해 녹음된 소리 파일에서 음소를 정확히 구별할 수 있었으며, 청각장애인의 발화를 통해 녹음된 소리(발음)의 정확성을 확인하기 위해 말 명료도를 측정할 수 있었다.

```
# stt 결과
res = stt(args.file)
json_file = wav_file.parent / pathlib.Path(wav_file.name.replace( “.wav” ,” .json” ))
if json_file.exists():
   with json_file.open( “r” ) as f:
        data = json.load(f)
        transcript = data[ “annotations” ][ “script” ]
        pronounce = data[ “annotations” ][ “pronunciationForm” ]

# cer
cer_metric = load_metric( “cer” )
cer_score = cer_metric.compute(
    predictions = [to_label(res[ “transcript” ])],
    references = [to_label(transcript)])

# plot spectrogram
sound = Sound(args.file)
spectral = SpectralFeatureExtractor(sound)
spectral.get_magnitude()
spec = to_array(spectral.get_spectrogram(spectral.magnifude))
width, height = spec.shape
figsize = (1, height/width) if height>=width else (width/height, 1)
figsize = [s*3 for s in list(figsize)]
plt.figure(figsize=figsize)
plt.fitle(f’ 목표: {transcript}₩n 받음: {res[ “transcirpt” ]}₩n 말 명료도 = {1-cer_score:.4f}’ )

# pause_time
Pause_times = lpBoundary(sound.signal, sound.sample_rate).get_pause_times()
```

[그림 4] 말 명료도 분석을 위한 소스 코드

[Fig. 4] Source Code for Speech Intelligibility



[그림 5] 말 명료도의 분석결과

[Fig. 5] The Results of Speech Intelligibility

5. 토의 및 제언

본 논문에서는 인공지능 기반 Wav2vec 2.0을 이용하여 청각장애인의 음성신호로부터 음소의 특징을 추출하였다[11][14]. 또한, STT를 이용하여 청각장애인이 원래 발화하고자 했던 문장의 목표문장과 청각장애인의 발화한 것으로부터 인식된 문장에서 각각의 말소리 자소를 측정할 수 있었다[11][14]. 이번 연구에서는 비장애인이 구현한 정확한 음소로부터 STT를 통해 획득된 문장을 청각장애인이 원래 발화하고자 했던 문장인 목표문장으로 대체하여 분석했으며, STT를 이용한 비장애인과 청각장애인의 음소 추출 및 구별할 수 있는 기능을 구현하였다. 이를 통해 실제 확보된 데이터에 대해서 해당 정보를 식별할 수 있는 가능성을 타진 할 수 있었고, 음소 추출의 유효성을 확인할 수 있었다. 향후, STT를 통해 획득한 문장을 그대로 비교하는 것이 아니라, 목표 단어나

문장을 표준발음으로의 변환을 통해서 보정하고, 보정된 데이터를 이용하여 STT에 적용하면 보다 더 정확도를 높일 수 있을 것이다. 이러한 기술은 난청인들의 청능훈련을 위한 말하기 연습뿐만 아니라 비장애인들의 발음 유창성 문제를 해결하기 위한 프로그램에도 적용할 수 있다. 또한 음성 인공지능이 음소 단위의 말명료도를 분석하기 때문에 다양한 언어에 적용이 가능하므로 외국어 학습을 위한 발음교정 프로그램으로 개발할 수 있다.

6. 감사의 글

본 연구는 2022년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력 기반 지역혁신 사업의 결과입니다(2021RIS-001).

This research was supported by “Regional Innovation Strategy (RIS)” through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-001).

References

[1] S. Kim, W. Oh, D. Oh, J. Hwang, M. Oh, J. Kim, Y. Lee, D. Kang, S. Kwon, E. Baek, S. Yune, S. Lee, 2020 The survey on disability statistical report, Ministry of Health and Welfare, (2020)
Available from: https://www.prism.go.kr/homepage/entire/downloadResearchAttachFile.do?workKey=000&fileType=USR&seqNo=001&pdfConvYn=Y&researchId=1351000-202000241

[2] Karina F. M. Tao, Tais de C. Moreira, Dona M. P. Jayakody, De Wet Swanepoel, Christopher G. Brennan-Jones, Lize Coetzee, and Robert H. Eikelboom, Teleaudiology hearing aid fitting follow-up consultations for adults: single blinded crossover randomised control trial and cohort studies, International Journal of Audiology, (2021), Vol.60, No.51, pp.549-560.
DOI: https://doi.org/10.1080/14992027.2020.1805804

[3] J. Kwon, Y. Lee, Artificial Intelligence for Persons with Disabilities, The Journal of Korea Elementary Educaton, (2020), Vol.31, Supplement, pp.187-202.
DOI: https://doi.org/10.20972/kjee.31..202008.187

[4] L. C. Yann, Y. Bengio, and G. Hinton, Deep learning, Nature, (2015), Vol.521, No.7553, pp.436-444.
DOI: https://doi.org/10.1038/nature14539

[5] D. Jurafsky, J. H. Martin, Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (3rd ed.), Pearson, (2020)

[6] E. W. Healy, S. E. Yoho, Difficulty understanding speech in noise by the hearing impaired: underlying causes and technological solutions, Proceedings of the 38th Annual International Conference of The IEEE in Medicine and Biology Society, IEEE, (2016)
DOI: https://doi.org/10.1109/EMBC.2016.7590647

[7] A. Chern, Y. H. Lai, Y. P. Chang, Y. Tsao, R. Y. Chang, and H. W. Chang, A smartphone-based multi-functional hearing assistive system to facilitate speech recognition in the classroom, IEEE Access, (2017), Vol.5, pp.10339-10351.
DOI: https://doi.org/10.1109/embc.2016.7590647

[8] C. K. A. Reddy, N. Shankar, G. S. Bhat, R. Charan, and I. Panahi, An individualized super-gaussian single microphone speech enhancement for hearing aid users with smartphone as an assistive device, IEEE Signal Processing Letters, (2017), Vol.24, No.11, pp.1601-1605.
DOI: https://doi.org/10.1109/LSP.2017.2750979

[9] G. Park, W. Cho, K. Kim, and S. Lee, Speech enhancement for hearing aids with deep learning on environmental noises, Appl. Sci., (2020), Vol.10, No.17, pp.6077.
DOI: https://dx.doi.org/10.3390/app10176077

[10] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Chang, and T. Sainath, Deep learning for audio signal processing, , Journal of selected topics and signal processing, (2019), Vol.13, No.2, pp. 206-219.
DOI: https://doi.org/10.1109/JSTSP.2019.2908700

[11] S. Schneider, A. Baevski, R. Collobert, and M. Auli, wav2vec: Unsupervised pre-training for speech recognition, Interspeech, (2019)
DOI: https://doi.org/10.48550/arXiv.1904.05862

[12] A. Baevski, H. Zhou, A. Mohamed, A., and M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in Neural Information Processing Systems, (2020),Vol.33, pp.12449-12460.
DOI: https://doi.org/10.48550/arXiv.2006.11477

[13] H. Yoon, Diagnosis of Parkinson's disease based on audio voice using wav2vec, Journal of Digital Convergence, (2021), Vol.19, No.12, pp.353-358.
DOI: https://doi.org/10.14400/JDC.2021.19.12.353

[14] https://blog.naver.com/andrew9909/222854501348, Feb 28 (2023)

[15] S. Min, K. Lee, D. Lee, D. Ryu, A Study on Quantitative Evaluation Method for STT Engine Accuracy based on Korean Characteristics, Journal of the Korea Academia-industrial cooperation Society, (2020), Vol.21, No.7, pp.699-707.
DOI: https://doi.org/10.5762/KAIS.2020.21.7.699

[16] S. Kim, The Role of Speech Factors in Speech Intelligibility: A Review, Malsori, (2002), Vol.43, pp.25-44.
Available from: http://www.koreascience.or.kr/article/JAKO200213842066256.pdf