한국어 노인 음성 데이터 증강 및 인식 연구

김건희○, 박서윤, 김한샘
연세대학교 언어정보학협동과정
keonhee0510@gmail.com, {seoyoon.park, khss}@yonsei.ac.kr

A Study of Data Augmentation and Auto Speech Recognition for the Elderly

Keon Hee Kim○, Seoyoon Park, Hansaem Kim
Yonsei University, Interdisciplinary program of language and information

요 약

기존의 음성인식은 청장년 층에 초점이 맞추어져 있었으나, 최근 고령화가 가속되면서 노인 음성에 대한 연구 필요성이 증대되고 있다. 그러나 노인 음성 데이터셋은 청장년 음성 데이터셋에 비해서는 아직까지 충분히 확보되지 못하고 있다. 본 연구에서는 부족한 노인 음성 데이터셋 확보에 기여하고자 희소한 노인 데이터셋을 증강할 수 있는 방법론에 대해 연구하였다. 이를 위해 노인 음성 특징(feature)을 분석하였으며, '주파수'와 '발화 속도' 특징을 일반 성인 음성에 합성하여 데이터를 증강하였다. 이후 Whisper small 모델을 파인 튜닝한 뒤 노인 음성에 대한 CER(Character Error Rate)를 구하였고, 기존 노인 데이터셋에 증강한 데이터셋을 함께 사용하는 것이 가장 효과적임을 밝혀내었다.

주제어: 음성 인식, 노인 음성, Whisper, ASR

1. 서론

딥러닝 기술의 출현과 다양한 음성인식 오픈소스 프로그램 덕분에 음성인식 기술은 지난 10년 동안 비약적으로 발전해왔다. 이러한 음성인식 기술은 대부분 청장년 음성 데이터를 기반으로 만들어졌다. 따라서 일반 성인 이외의 음성 데이터셋은 희소하며 비교적으로 연구가 많이 이루어지지 않았다. 한국 사회는 노령 인구가 빠르게 증가하고 있는 추세이다. 통계청에 따르면, 2010년 65세 이상 인구가 약 54만 명에서 2020년 82만 명으로 증가했고, 그 중 85세 이상 초고령자 인구는 두 배 이상 증가하였다. 빠르게 고령화되어가는 한국 사회에 노인음성처리 기술의 확보는 필수이다. 본 연구는 기존 한국어 주류 음성 데이터를 차지하고 있는 성인 음성에 비해 양적, 질적으로 부족한 노인 음성 데이터에 대한 증강과 더불어 음성 인식 모델들의 노인 음성 인식률 향상에 기여하는 방법론을 살펴보는 것을 목표로 한다. 이를 위해 현존하는 한국어 성인 남녀 음성에 대해 노인 음성 feature를 반영한 증강 방법론을 적용하여 노인 음성 데이터셋을 얻는 한편, 증강한 데이터셋에 대해 음성 인식 모델들의 유의미한 성능 향상이 있는지 연구했다. 이를 토대로 노인 음성 인식률 향상에 기여하는 자질(feature)이 무엇인지를 살펴보고, 모델에 대해 어떤 실험 방법론이 유효한지를 중심으로 연구를 진행하였다.

2. 관련 연구

2.1. 노인 음성 특징 관련 연구

노인 음성의 대표적인 특징으로 느린 발화, 긴 휴지, 잦은 비유창성 등이 있다([1-5]). 노화가 진행되면서 혀의 두께, 혀의 움직임의 범위와 지속시간이 줄어들고 발화할 때 반응이 느려지면서 발화 기능에 영향을 주게 된다([2,3]). 또한, 노인들은 단어 사이가 아닌 음절 사이에 휴지를 두는 유창 휴식(fluency breaks)이 증가하며 음절을 길게 늘여 발음하는 경향이 있다([2,3,5]). 나이에 따라 기본주파수가 변화하는 결과를 보여준 연구로는 [1,6]을 들 수 있다. 이 중 [1]에서는 60-80세 노인들의 음성 샘플을 분석하여 노인 남성 기본 주파수가 젊은 남성에 비해 다소 높고 노인 여성 기본 주파수는 젊은 여성에 비해 낮은 것을 관찰하였다. 또한 노인 남성과 노인 여성 사이의 연령별에 따른 기본 주파수 차이는 없었음을 밝혔는데, 이는 60대 이상 노인들은 운동 생리적 기능 저하와 신체적 노화로 인해 더 이상 변화가 나타나지 않기 때문이다.

[7]에 의하면 노인 음성은 청년 음성과 다른 특성들을 가지고 있고 데이터가 적기 때문에 청년 음성으로 학습된 음성인식 기반 서비스에서 노인 사용자의 음성은 인식률이 떨어진다. 노인 음성 데이터 부족으로 인한 낮은 인식률을 개선하기 위한 방법으로는 부족한 데이터를 더 수집하거나, 노인 음성을 청년음성으로 정규화해서 인식

률을 높이거나, 기존에 많이 있는 비노년층 음성 데이터에 노인 음성 특징을 입혀 데이터를 증강하는 방식이 제안되었다.

2.2. 노인 음성 인식률 제고 연구

노인 음성 데이터를 수집한 연구로는 [5]를 들 수 있다. [5]에서는 65세에서 99세까지 초고령층을 포함한 일본 노인 음성 데이터를 수집하여 단어 인식률을 높였다. 평균 연령 67.6세 음성 데이터를 사용하여 학습 모델을 했을 때 단어오류율 (Word Error Rate, WER)이 21.85%에서 17.4%로 감소한 것에 비해 초고령 노인의 음성을 포함하여 학습을 시켰을 때에는 WER이 13.21%로 더 감소한 결과를 보여주었다. 또한 [8]에서는 노인, 어린이 음성 등 희소한 데이터에 대해 추가 데이터셋을 확보해서 미세 조정한 결과, WER가 기존보다 20% 감소한 결과를 보여 음성 인식에 데이터가 매우 중요함을 역설하였다. 그러나 부족한 데이터를 더 수집하여 음성인식 모델을 학습시키는 것은 최선의 방법이기는 하나, 시간과 예산이 많이 들기 때문에 일반적인 연구 환경이나 현실적인 조건 등을 고려했을 때에는 어려움이 수반된다.

데이터셋의 양적 확보에 대한 대안으로 노인 음성을 일반 성인 음성으로 정규화하여 인식률을 높이는 방법들이 연구되었다. 노인 음성 데이터에서 휴지를 찾아 길이를 줄여 전처리한 [9]의 연구는 휴지를 제거한 노인 음성 데이터에서 유의미한 성능 향상 결과를 보여주었다. [10]은 노인 음성의 여러 특징을 정규화하여 음성인식률을 높인 바 있다. 노인 발화 특징에서 발화 속도가 느리다는 점, 휴지가 많다는 점과 노인 남성의 상대적 F2, F3 에너지가 청년 남성보다 낮다는 점에 착안하여, 노인 음성 데이터의 속도를 증가했을 때 인식률이 1%대 증가하였고, 음절 간 휴지를 제거했을 때 여성 노인 음성의 인식률이 4.2% 증가했고, 노인 남성의 상대적 F2와 F3 에너지 값을 청년 남성 에너지 값으로 조정했을 때, 인식률이 6% 증가한 결과를 보여주었다. 노인 음성의 발화 속도, 휴지 빈도 수 그리고 formant 값을 다 정규화 했을 때, 인식률이 12% 오른 것을 보여주었다.

이외에도 기존 데이터를 변형시켜 부족한 데이터를 증강시키는 연구들이 진행되었다. 데이터가 많은 청년 음성을 노인 음성과 비슷하게 합성하여 증강시키는 것이 목적으로, [4]에서는 청년 음성을 위상 보코더를 사용하여 음성 길이를 0.5배 늘려 노인 음성의 발화처럼 변환했다. 노인 음성 데이터를 전처리한 청년 음성 데이터로 증강시켜 사용한 결과 인식 오류율을 낮출 수 있었다. 이러한 일반화 데이터 증강 방법을 사용한 연구가 있는가 하면, 청년 음성과 노인 음성의 미세한 주파수-시간 차이를 생성적 적대 신경망(Generative Adversarial Networks, GANs) 기반으로 데이터를 화자 종속적으로 증강하여 WER을 9.61% 줄인 [11]의 연구도 있다. 본 연구에서는 [4]의 방법을 따라 청장년 음성 데이터를 노인 발화의 특징을 feature로 사용하여 이를 토대로 데이터 증강을 진행한 후, 노인 음성의 인식률을 높일 계획이다. 다만 기존의 선행 연구들이 노인 음성을 청년 음성에 맞추어 정규화하거나 하나의 자질만을 사용하여 증강했던 것과 달리 본 연구는 여러 자질을 단독으로, 혹은 복합으로 사용하여 청년 음성을 변환하는 방법을 사용하였다.

3. 연구 목적 및 방법

본 연구에서는 일반 성인 남녀 음성과 노인음성을 비교하여 다른 음성 특징들을 탐색하여 유의미한 차이가 있는 feature들을 일반 음성 데이터에 증강 시켜 파인튜닝 후 인식률을 높이는 것을 목표로 한다. 연구 질문은 다음과 같다.

◆ 일반남녀와 노인남녀 음성 데이터 간에 발화 속도, 휴지 길이, 주파수에 있어 유의미한 차이가 있는가?
◆ 유의미한 차이를 보인 feature들을 일반 남녀 음성 데이터에 합성하여 파인튜닝하면 음성 인식률이 증가하는가?

3.1. 음성 데이터 자료

현재까지 대중에게 공개된 노인 음성 관련 데이터셋 중 하나인 AIHub의 노인남녀 자유대화 음성(2020)과 일반남녀 자유대화 음성(2020)을 실험 데이터로써 사용하였다. 해당 데이터셋들은 청년 음성을 20대부터 50대까지, 노인 음성을 60대 이상으로 정의하고 있다. 본 연구에서는 노인 음성을 70대 이상으로 한정하여 데이터 정제를 진행하였다. 일반 성인 음성의 경우 연령대에 제한을 두지 않았다. 각 성인/노인 데이터별로 약 15시간씩 음성 파일을 랜덤 샘플링 하였고, 성별 균형을 맞추기 위해 남녀 발화자를 1:1로 맞추었다. 출신 지역 등 인구학적 특성은 현실(real world)과 비슷한 조건을 만들기 위해 고려하지 않았다.

3.2. 음성 특징 비교

성인 음성과 노인 음성을 비교하고자 기존 연구들에서 주요 feature로 언급된 시간당 음절 수(발화 속도), 주

파수를 각 연령대별 음성 파일에 대해 구한 후 이를 평균하였다.

주파수의 경우 향후 피처 엔지니어링을 감안하여 전체 음성 중 최저치에 대해 분석하였다. 피처 간 차이에 대해서는 정말 유의미한 차이를 보이는 지에 대해 독립표본 t-test 검정을 실시하였다.

<표 1> 연령대별 주파수 feature 통계 정보 및 t-검정 결과

최소 주파수	전체	여성	남성
노인	95.96(SD=29.91)	100.79(SD=37.16)	90.53(SD=17.15)
성인	99.19(SD=28.98)	126.76(SD=30.14)	85.14(SD=14.84)
t-test	t(11998)=-6.02 (p<.001***)	t(5196)=26.39 (p<.001***)	t(6800)=13.84 (p<.001***)

주파수의 경우 [1]의 결과와 일치하는 경향을 보였으며, t-test 결과 전체 성별 및 여성, 남성에서 노인과 일반 성인의 주파수 간 유의미한 차이가 있었다. 여성의 경우 노인 여성의 최저 주파수는 100.79Hz로 일반 성인 여성(126.76Hz)보다 약 0.8배 낮은 주파수를 보인 반면, 남성의 경우 노인 남성의 주파수는 90.53Hz, 일반 성인 남성의 주파수는 85.14Hz로 노인 남성이 약 1.1배 정도 높았다.

발화 속도의 경우 1초당 음절 수 (syllable per second;SPS)를 기준으로 삼았으며 지나치게 짧은 발화의 경우 SPS를 온전히 파악할 수 없어 20음절 이상의 발화만을 분석 대상으로 하였다. 일반 성인 음성과 노인 음성의 발화 속도를 비교한 결과, 노인 음성은 1초당 3.15음절, 성인 음성은 3.64음절을 보여 노인 음성이 성인 음성보다 초당 발음 음절 수가 적음을 확인할 수 있었다. 성별별로도 노인이 성인보다 느린 발화를 보였는데, 노인 여성의 경우 성인 여성보다 0.81배 느렸고, 노인 남성은 성인 남성보다 0.93배 느렸다. 이러한 차이가 유의미한지 확인하고자 독립표본 t-test를 시행한 결과, 실제로도 노인/성인 간 유의미한 차이였음을 알 수 있었다.

<표 2> 연령대별 SPS feature 통계 정보 및 t-검정 결과

초당 음절 수	전체	여성	남성
노인	3.15(SD=0.33)	3.03(SD=0.31)	3.3(SD=0.3)
성인	3.64(SD=0.41)	3.76(SD=0.36)	3.56(SD=0.43)
t-test	t(5226)=-47.65 (p<.001***)	t(2639)=-53.84 (p<.001***)	t(2585)=-17.43 (p<.001***)

3.3. 음성 합성

3.2.에서 확인한 것과 같이 노인과 일반 성인 음성 특징을 결정하는 feature는 속도, 주파수 모두이다. 이에 따라 본 연구에서는 일반 음성에 대해 노인 음성의 feature를 반영한 증강 음성을 생성하였다. 생성 시 일반 성인 음성의 발화 속도 조절은 Librosa 라이브러리를 사용하여 여성은 0.81배, 남성은 0.93배 늦춰서 노인 음성 데이터의 속도와 맞추었다. 주파수의 경우 3.2.에서 서술한 바와 같이 Praat의 Parselmouth를 사용하여 일반 성인 여성은 0.8배, 남성 음성에는 1.1배로 주파수를 변형하였다. 이렇게 증강한 음성들을 사전학습에 대해 fine-tuning 데이터로 사용하였다.

4. 실험 및 결과

실험은 Google Colab을 사용하여 진행하였으며, 음성 사전 학습 모델인 Whisper[12]를 사전학습 모델로 선택하였다. Whisper는 open AI가 공개한 사전 학습된 음성 인식 모델로, 약 68만 시간 분량의 데이터로 학습되어 다국어 인식은 물론 다양한 태스크를 수행할 수 있는 능력을 갖추었다. 본 연구에서 Whisper를 선택한 이유는 Whisper 내 한국어 학습이 충분하다는 판단 때문이다. 실제로 한국어는 Whisper의 다국어 음성 인식 학습 데이터(117,113시간) 중 7,993 시간으로 규모 7위, 다국어 번역 데이터(125,739시간) 중 19,938시간으로 규모 1위를 차지하고 있다. Whisper는 파라미터 수에 따라 5개의 모델 종류로 나누어지며, 본 연구에서는 실험 환경을 고려하여 small 모델을 사용하여 연구를 진행하였다.

<표 3> 파라미터별 Whisper 모델

Whisper's Model	Parameters
Tiny	39M
Base	74M
Small	244M
Medium	769M
Large	1550M

Whisper small 모델에 기본 일반 음성, 노인 음성, 그리고 피처 엔지니어링을 거친 데이터로 파인 튜닝을 진행한 후, 노인 음성 인식을 진행하여 성능을 측정하였다. 즉, 성능은 '일반 음성/노인 베이스라인', '단독 피처별(속도, 주파수), 복합 피처별(속도+주파수)', 그리고 '합성 데이터별(노인+ 일반 {속도, 주파수, 속도+주파수})'로 파인튜닝 후 노인 음성을 얼마나 잘 인식하였는지를 나타낸다. 음성 인식 성능 지표로는 CER(Character Error Rate)을 사용하였는데, 이는 형태소 단위의 언어인 한국어의 특성을 반영한 성능 지표를 사용한 것이다. 실험은 Huggingface에 게시되어 있는

'openai/whisper-small' 모델을 사용하였으며, 실험 시 사용한 하이퍼파라미터는 아래와 같다.

<표 4> 하이퍼파라미터 명세

train batch: 8 gradient accumulation steps: 2 learning rate: 2e-5 warmup-steps: 50 epochs: 5 fp16: True eval. batch: 8

4.1. 베이스라인 실험

본 연구에서 노인 음성 데이터셋만 학습한 모델과 일반 성인 음성 데이터셋만 학습한 모델의 노인 음성 인식 성능을 baseline으로 사용하였다. 노인 음성 데이터셋으로만 파인튜닝을 진행한 모델은 CER이 4.62%로 나타났으며, 일반 성인 음성 데이터셋으로만 학습한 모델의 오류율은 7.49%을 기록했다. 이를 통해 노인 음성을 인식할 때 일반 성인 음성으로만 학습하는 것만으로는 충분하지 않으며, 더 많은 노인 데이터 혹은 노인 음성 증강 데이터가 필요함을 알 수 있다.

<표 5> 베이스라인 CER 성능

	노인 음성 단독 학습	일반 음성 단독 학습
CER	**4.62**	7.49

4.2. feature engineering 실험

다음으로는 일반 음성에 대해 피처 엔지니어링을 수행한 데이터셋을 파인튜닝 데이터로 사용하여 음성 인식을 진행하였다. 피처 엔지니어링 데이터셋은 일반 음성에 노인 음성 특성을 반영한 데이터셋으로, 주파수, 발화 속도, 그리고 두 개를 모두 수행한 데이터셋 총 3종류이다. '주파수 변형 데이터셋'은 일반 여성 음성 주파수에 0.8배를, 남성 음성에는 1.1배를 가한 데이터로 구성되어 있으며, '발화 속도 변형 데이터셋'은 여성 남성 각각 0.81배, 0.93배 느리게 한 데이터이다. '두 개를 모두 수행'한 데이터셋은 주파수 변형 데이터셋에 발화 속도를 변형하여 구축하였다. 실험 결과는 아래 표와 같이 나타났다.

<표 6> 피처 엔지니어링 데이터셋 CER 성능

	일반 음성 주파수 변형	일반 음성 속도 변형	일반 음성 주파수+속도
CER	**7.64**	8.31	8.71

피처 엔지니어링 데이터셋 중 가장 낮은 CER을 보인 데이터셋은 주파수 변형 데이터셋으로, 청년 베이스라인과 비슷한 CER을 보였다. 반면 속도 변형이나, 두 가지 피처를 모두 변형한 데이터셋은 청년 데이터셋을 하회하는 CER을 보여 주파수 변형보다는 음성 인식 성능이 낮은 피처임을 알 수 있다. 위 실험을 통해 일반 음성에 다양한 피처 엔지니어링을 진행한 데이터셋 단독으로는 노인 음성에 대한 음성 인식률을 높일 수 없는 것을 확인할 수 있었다.

4.3. 혼합 데이터셋 실험

이 실험에서는 노인 음성 데이터와 음성 합성 데이터를 50:50 비율로 혼합한 데이터셋을 학습에 사용하였다. 노인 음성 데이터와 일반 성인 음성의 주파수 변환 데이터(혼합 데이터셋 A), 노인 음성 데이터와 속도 변환 데이터(혼합 데이터셋 B), 노인 음성 데이터와 주파수+속도 변환 데이터(혼합 데이터셋 C)로 구성하여 실험을 진행하였다. 실험 결과 혼합 데이터셋 A의 CER은 4.91%, B는 4.89%, C는 5.02%를 기록하였다.

<표 7> 혼합 데이터셋 CER 성능

	혼합 데이터셋 A(주파수)	혼합 데이터셋 B(속도)	혼합 데이터셋 C(속도+주파수)
CER	4.91	**4.89**	5.02

위 결과를 통해 피처를 변형한 데이터셋을 단독으로 사용하는 것보다는 노인 데이터와 혼합하여 사용하는 것이 유의미한 성능 향상을 이끌어내는 것을 알 수 있었다. 세부적으로는 속도를 변환한 혼합 데이터셋 B가 가장 낮은 CER을 보여주기는 했으나, 주파수를 변환한 혼합 데이터셋 A보다 근소하게 낮아 단독 피처 엔지니어링을 진행한 데이터를 기존 노인 데이터에 섞어 사용하는 것이 성능 향상에 유의미한 영향이 있음을 확인하였다. 주파수와 속도 모두 변환한 데이터셋의 경우 4.2.의 경향과 유사하게 A, B, C 데이터셋 중 가장 성능이 좋지 않았으나, 일반 음성 단독 학습이나 혼합 없이 피처 엔지니어링 데이터셋만을 사용한 결과보다는 향상된 결과를 보여주었다.

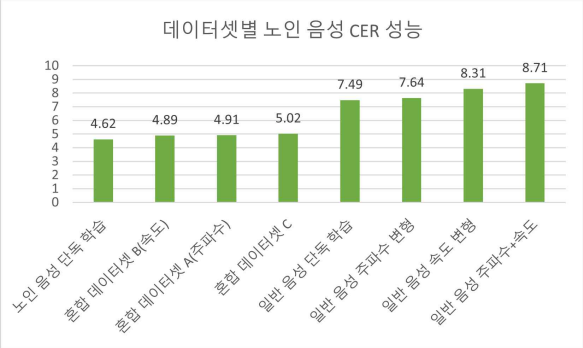

그림 1 데이터셋별 노인 음성 CER 성능

실험 결과 노인 음성 인식 학습 데이터로 직접적으로 노인 음성 데이터셋을 사용하는 것이 가장 성능이 높다는 것을 확인하였다. 그러나 현실에서 노인 데이터셋은 희소하며, 구축에는 비용이 들기에 기존의 일반 성인 음성을 변형하여 증강 데이터로 사용하는 방법이 필요하다. 실험 결과, 노인과 성인 음성을 구분짓는 특성 중 노인 음성 인식 향상에 영향을 주는 피처는 '주파수'였다. 또한 피처 엔지니어링이 된 데이터셋을 단독으로 사용하는 것보다는 기존의 데이터셋과 증강한 데이터셋을 섞어서 사용하는 것이 효과적이며, 주파수와 속도 모두 성능 향상에 긍정적인 영향을 주는 것을 확인할 수 있었다.

5. 결론

본 연구는 음성 데이터 양이 부족하여 음성 인식 성능이 떨어지는 70대 이상 노인의 음성 인식률을 높이기 위해 노인 음성을 증강할 수 있는 방법론에 대해 연구를 진행하였다. 이를 위해 노인 음성과 일반 성인 음성을 비교하여 유의미한 차이가 있는 음성적 특징을 찾아내었다. 노인 음성과 성인 음성의 차이점은 주파수와 발화 속도였으며, 본 연구는 이 특징들을 피처 엔지니어링을 통해 일반 성인 음성 데이터에 합성시켜 데이터를 증강하였다. 이렇게 증강한 데이터로 음성 인식 모델을 학습시켜 어떤 방법론이 노인 음성 인식률을 높이는지 실험을 진행하였으며, 실험 결과 피처 엔지니어링을 진행한 데이터셋을 단독으로 사용하는 것보다는 기존 데이터와 섞어서 사용하는 것이 노인 음성 인식률 향상에 도움이 되는 것을 확인하였다.

연구를 통해 노인 음성 데이터셋의 희소성을 극복할 수 있는 일반 음성 증강 방법론을 제시하고, 실험을 통해 효과적인 방법을 제시했다는 점에서 의의가 있다. 향후에는 본 연구에서 다룬 특성 외에도 노인 음성에서 두드러지는 특성들을 탐색하여 증강 방법론을 실험하는 한편, 피처 엔지니어링을 통한 합성 음성의 음성 인식률을 제고할 계획이다.

참고문헌

[1] 허명진, & 신명선. (2010). 노인 음성의 음향학적 특성. 언어치료연구, 19(2), 41-51.

[2] 박지웅, 이승준, & 권순일. (2013). 노인음성인식을 위한 전처리에 관한 연구. 한국정보처리학회 학술대회논문집, 20(2), 1646-1648.

[3] Soonil Kwon, Sung-Jae Kim, and Joon Yeon Choeh. 2016. Preprocessing for elderly speech recognition of smart devices. Computer Speech & Language, 36:110-121.

[4] 윤수연, 김태인, 나종환, & 이보원. (2022). 위상 보코더를 활용한 데이터 증강 및 노인 음성인식 성능 비교. 대한전자공학회 학술대회, 1167-1170.

[5] Fukuda, M., Nishizaki, H., Iribe, Y., Nishimura, R., & Kitaoka, N. (2020, May). Improving speech recognition for the elderly: A new corpus of elderly japanese speech and investigation of acoustic modeling for speech recognition. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 6578-6585).

[6] Lee, S. J., & Gwon, S. I. (2014). 노인의 음성인식 성능 개선을 위한 노인음성 분석. Communications of the Korean Institute of Information Scientists and Engineers, 32(11), 16-20.

[7] 김준우 and 정호영. (2020). 제한된 학습 데이터를 사용하는 End-to-End 음성 인식 모델. 말소리와 음성과학, 12(4), 63-71.

[8] 김연군, 김형진, & 이정우. (2021). 노인, 어린이 음성을 대상으로 한 한국어 음성 인식 모델에 관한 연구. 한국통신학회 학술대회논문집, 608-609.

[9] 나종환, 윤수연, 서지영, & 이보원. (2022). 휴지(pause) 제거 미세조정을 사용한 노인 음성인식 성능 비교. 대한전자공학회 학술대회, 953-955.

[10] 손귀영, 백성욱 and 권순일. (2016). 응급상황 음성을 통한 성별간의 발화행태 특성 분석. 한국차세대컴퓨팅학회 논문지, 12(1), 55-65.

[1] Jin, Z., Geng, M., Deng, J., Wang, T., Hu, S., Li, G., & Liu, X. (2022). Personalized adversarial data augmentation for dysarthric and elderly speech recognition. arXiv preprint arXiv:2205.06445

[11] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning (pp. 28492-28518). PMLR.