KIPS Trans. Softw. and Data Eng.
Vol.12, No.8 pp.365~370
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2023.12.8.365

Comparison of Classification Performance Between Adult and Elderly Using Acoustic and Linguistic Features from Spontaneous Speech

SeungHoon Han† · Byung Ok Kang†† · Sunghee Dong†††

ABSTRACT

This paper aims to compare the performance of speech data classification into two groups, adult and elderly, based on the acoustic and linguistic characteristics that change due to aging, such as changes in respiratory patterns, phonation, pitch, frequency, and language expression ability. For acoustic features we used attributes related to the frequency, amplitude, and spectrum of speech voices. As for linguistic features, we extracted hidden state vector representations containing contextual information from the transcription of speech utterances using KoBERT, a Korean pre-trained language model that has shown excellent performance in natural language processing tasks. The classification performance of each model trained based on acoustic and linguistic features was evaluated, and the F1 scores of each model for the two classes, adult and elderly, were examined after address the class imbalance problem by down-sampling. The experimental results showed that using linguistic features provided better performance for classifying adult and elderly than using acoustic features, and even when the class proportions were equal, the classification performance for adult was higher than that for elderly.

Keywords : Spontaneous Speech, Acoustic Features, Linguistic Features, Speaker Age Recognition

자유대화의 음향적 특징 및 언어적 특징 기반의 성인과 노인 분류 성능 비교

한 승 훈† · 강 병 옥†† · 동 성 희†††

요 약

사람은 노화과정에 따라 발화의 호흡, 조음, 높낮이, 주파수, 언어 표현 능력 등이 변화한다. 본 논문에서는 이러한 변화로부터 발생하는 음향적, 언어적 특징을 기반으로 발화 데이터를 성인과 노인 두 그룹으로 분류하는 성능을 비교하고자 한다. 음향적 특징으로는 발화 음성의 주파수(frequency), 진폭(amplitude), 스펙트럼(spectrum)과 관련된 특징을 사용하였으며, 언어적 특징으로는 자연어처리 분야에서 우수한 성능을 보이고 있는 한국어 대용량 코퍼스 사전학습 모델인 KoBERT를 통해 발화 전사문의 맥락 정보를 담은 은닉상태 벡터 표현을 추출하여 사용하였다. 본 논문에서는 음향적 특징과 언어적 특징을 기반으로 학습된 각 모델의 분류 성능을 확인하였다. 또한, 다운샘플링을 통해 클래스 불균형 문제를 해소한 뒤 성인과 노인 두 클래스에 대한 각 모델의 F1 점수를 확인하였다. 실험 결과로, 음향적 특징을 사용하였을 때보다 언어적 특징을 사용하였을 때 성인과 노인 분류에서 더 높은 성능을 보이는 것으로 나타났으며, 클래스 비율이 동일하더라도 노인에 대한 분류 성능보다 성인에 대한 분류 성능이 높음을 확인하였다.

키워드 : 자유 대화, 음향적 특징, 언어적 특징, 화자 연령 분류

1. 서 론

사람은 상대방의 음성을 듣고 대략적인 연령을 가늠할 수 있다. 이는 연령에 따라 호흡, 조음, 언어표현 등에서 차이가 존재하기 때문이다[1]. 이러한 차이를 이용하여 화자의 연령을 분류하는 기술은 소셜 미디어나 전자 상거래에서 고객 맞춤형 서비스를 제공하는데 기여할 수 있다. 또한, 인구 고령화로 인한 노인 인구의 증가로 인해 노년층 대상 서비스의 필요성이 증가하고 있으며, 여기에 노인 발화에 대한 분류 기술이 접목될 수 있다. 예를 들어, 음향적 특징과 언어적 특징에 따라 연령 집단을 분류하는 기술은 주로 노인에게 발생하는 인지기능과 관련된 치매, 우울증 등의 질환을 조기에 발견하고 예방하기 위한 목적으로 활용될 수 있다[2-5]. 인공지능 에이전트와의 대화를 통해 인지기능 검사를 진행할 때, 인지기능 장애 발생 위험은 성인보다 노인에게서 더 크게 나타나므로, 양성 진단에 있어 노인 사용자에게는 성인보다 높은 가중치를 두어야한다. 또한, 이와 같은 발화 기반 서비스를 일상적인 환경에서 사용할 경우 같은 공간 내에 다수의 화자가 존재해

※ 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 국가과학기술연구회 창의형 융합연구사업(No.CAP21052-300)의 지원을 받아 수행된 연구임.
※ 이 논문은 2023년 한국 소프트웨어공학 학술대회의 "자유대화 과제에서 음성적 특징과 언어적 특징 기반의 성인과 노인 분류 성능 비교"의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 고려대학교 뇌공학과 석사과정
†† 비 회 원 : 한국전자통신연구원 책임연구원
††† 비 회 원 : 한국전자통신연구원 선임연구원
Manuscript Received : April 24, 2023
First Revision : June 5, 2023
Second Revision : June 29, 2023
Accepted : July 19, 2023
* Corresponding Author : SeungHoon Han(tht1102@etri.re.kr)

노인 사용자 외의 다른 화자의 발화가 함께 기록되는 문제가 발생할 수 있다. 이러한 경우, 분석 대상의 발화를 식별하고 서비스의 성능을 개선하기 위해 노인 발화를 식별하는 과정이 필요하다. 따라서 성인과 노인에 대한 화자 연령 분류 기술의 중요성이 늘어날 것으로 보인다.

따라서 본 논문은 KCSE 2023 우수논문[6]을 확장하여, 한국어 발화 데이터로부터 음향적 특징과 언어적 특징을 추출하고 이를 SVM(Support Vector Machine) 분류기와 MLP (Multi Layer Perceptron) 분류기를 사용해 화자를 성인 혹은 노인으로 분류하는 실험을 진행한다. 또한, 데이터 불균형을 해소하였을 때의 성인과 노인 각 클래스에 대한 분류 성능을 F1 점수를 통해 확인한다.

본 논문의 공헌은 다음과 같다. 첫째, 본 논문에서는 화자 연령 분류를 위한 자동 특징 추출 방법을 제안한다. 제안하는 프레임워크는 openSMILE(open-source Speech and Music Interpretation by Large-space Extraction)[7] 및 KoBERT (Korean Bidirectional Encoder Representations from Transformer)[8]를 이용하여 발화 데이터로부터 화자 연령 분류에 필요한 음향적 특징과 언어적 특징을 자동으로 추출한다.

둘째, 본 논문에서는 각 모달리티에 동일한 분류기를 사용하여 얻은 분류 성능을 비교함으로써, 한국어 사전학습 모델을 통해 추출된 언어적 특징을 사용한 화자 연령 분류 기술의 성능 향상 가능성에 대해 검토한다.

셋째, 본 논문에서는 화자 연령 분류를 위한 언어적 특징 추출 시 발화 오류 정보를 활용하는 방법을 제안한다. 본 논문의 실험 결과에서는, 발화 오류 정보가 담긴 전사문으로부터 추출한 언어적 특징과 SVM 분류기를 이용함으로써, 성인과 노인 분류에 있어 선행 연구 대비 1.1%의 성능 개선이 이루어졌다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로서 음향적 특징을 사용한 화자 연령 분류 연구와 언어적 특징을 사용한 화자 연령 분류 연구를 소개한다. 3장에서는 본 논문에서 제안하는 화자 연령 분류 방법에 대해서 설명한다. 4장에서는 실험 내용 및 결과에 대해 다룬다. 마지막으로 5장에서 결론 및 향후 연구를 제시하며 마무리한다.

2. 관련 연구

기존의 화자 연령 분류에 관한 연구는 주로 음향적 특징을 사용하였다. 기존에는 노화에 따라 달라지는 발화 음성의 MFCC(Mel-Frequency Cepstral Coefficient), 포먼트(formant), 음성의 높낮이(pitch) 등을 통해 화자의 연령을 분류하였고, 주로 전통적인 기계학습 분류모델인 SVM이 사용되었다[9].

그러나 사람의 발화에는 음향적 특징뿐만 아니라 문장 구조와 의미와 같은 언어적 특징이 수반되며, 이 역시 화자의 연령을 분류하기 위한 단서가 될 수 있다. 따라서 보다 정확한 화자 연령 분류를 위해, 음향적 특징만을 사용한 연구에서 나아가 언어적 특징을 화자 연령 분류 연구에 활용하는 것이 타당하다.

언어적 특징을 사용한 연구로는 특정 단어의 등장 빈도와 간투어 사용 빈도 등의 특징을 사용해 18세 이하와 70세 이상의 연령대를 분류한 연구가 있다[10]. 또한, 최근에는 BERT (Bidirectional Encoder Representations from Transformer) 모델[11]의 출력 벡터를 언어적 특징으로 사용하는 연구가 제안되었다. 이와 관련하여, 텍스트 데이터를 BERT에 입력으로 넣어 해당 문장의 CLS 토큰의 은닉 벡터를 35세 미만과 35세 이상의 연령대 분류에 사용한 연구등이 있다[12]. 이와 같이 대량의 데이터로 사전 학습된 모델을 사용하면, 입력 데이터로부터 복잡한 관계와 패턴을 자동적으로 파악하고 발화의 맥락 정보를 담은 언어적 특징을 추출할 수 있다는 장점이 있다.

Han 등[6]은 한국어 발화 데이터셋을 사용하여 음향적 특징과 언어적 특징에 기반한 성인과 노인 연령대 분류 성능을 비교하는 실험을 진행하였다. 해당 실험은 분류 성능 개선을 위해 한국어 자연어처리 분야에서 BERT 모델보다 우수한 성능을 보이는 한국어 사전학습 모델인 KoBERT[8]를 사용해 언어적 특징을 추출하였고, 향후 화자 연령 분류 기술을 노년층 대상 서비스에 접목하는 등의 실용적인 기여를 위해 60세 미만의 성인 화자와 60세 이상의 노인 화자를 분류하였다. 또한, 언어적 특징을 사용한 성인과 노인 분류에서 최대 71.4%의 weighted F1 점수를 보였다.

3. 제안 방법

본 논문에서는 화자의 발화 특징을 음향적 특징과 언어적 특징 두 가지로 나누어 추출하고, 이를 사용하여 기계학습 분류기를 통해 화자의 연령층을 성인과 노인 두 그룹으로 분류한다. 본 장에서는 위와 같은 절차를 통해 발화 데이터 기반 화자 연령 분류에서 보다 효과적인 모달리티를 판별하고, 언어적 특징에서 성인과 노인에 대한 변별력을 높이는 방법을 제안한다. 제안하는 방법은 Fig. 1과 같이 openSMILE[7]을 통해 음성 데이터로부터 음향적 특징을 추출하고, KoBERT[8]를 통해 전사문 데이터로부터 언어적 특징을 추출한다. 이어서 각 모달리티의 특징을 따로 혹은 결합해 사용하여 화자 연령 분류 성능을 확인한다.

3.1 음향적 특징

음성 데이터를 이용한 분류 과제에서 적합한 특징을 직접 선정하고 수동으로 추출하는 데에는 많은 시간이 소요될 수 있다. 또한, 가능한 모든 특징을 고려하기 위해 다수의 특징을 사용하게 되면 훈련 데이터셋에 대한 과적합 및 일반화 성능 저하 문제가 발생할 수 있다[13]. 따라서 본 논문에서는, 오픈소스 프레임워크인 openSMILE[7]을 통해 GeMAPS(Geneva Minimalistic Acoustic Parameter Set)와 eGeMAPS(extended Geneva Minimalistic Acoustic Parameter Set)[14]에 해당하는 사전 정의된 음향적 특징 세트를 음성 데이터로부터 추출

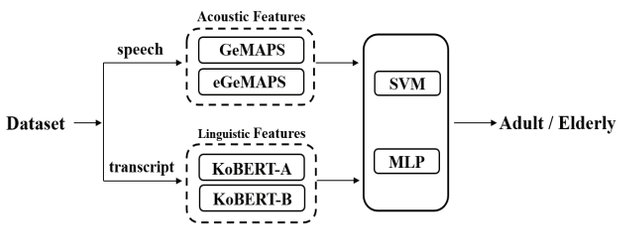

Fig. 1. Framework for Adult and Elderly Classification

하였다. GeMAPS와 eGeMAPS는 발화 음성에서 드러나는 준언어적(paralinguistic) 표현 및 감정 분석 등 다양한 음성 연구 분야에서 활용되는 특징 세트이다. GeMAPS는 기본 주파수(F0), 주파수변동률(Jitter), 진폭변동률(Shimmer)를 포함한 62개의 파라미터로 구성되어 있으며 이러한 파라미터는 연령 분류를 위해 사용될 수 있다[15, 16]. eGeMAPS는 GeMAPS의 파라미터 및 음성의 동적 특성을 반영하는 26개의 파라미터를 합친 총 88개의 파라미터로 구성되어 있다.

3.2 언어적 특징

BERT는 방대한 양의 텍스트 데이터로 사전 훈련된 언어 모델이다[11]. BERT 모델은 문장이 주어지면 이를 토큰 임베딩 표현으로 변환하여 모델의 입력으로 사용한다. 이 때, 각 문장의 토큰 임베딩 표현의 시작을 나타내는 CLS 토큰(Special Classification token)에 대한 BERT 모델의 출력을 해당 문장의 대표 임베딩 결과로써 다양한 하위 작업(downstream task)에 활용할 수 있다.

KoBERT 모델은 BERT 모델의 구조를 기반으로 위키피디아, 뉴스 등에서 수집한 한국어 문장과 단어를 통해 사전학습하여 한국어 자연어처리 분야에서 BERT 모델보다 우수한 성능을 보인다. 일례로, 한국어로 작성된 영화 리뷰를 긍정적 리뷰와 부정적 리뷰로 분류하는 Naver Sentiment Analysis에서 BERT의 다국어 버전인 Multilingual BERT[17]는 87.1%의 Accuracy를 기록한 반면, KoBERT는 그보다 1.9% 높은 89.0%의 Accuracy를 기록하였다[18]. 본 논문에서는 성인과 노인 분류 과제를 위해 학습 데이터셋으로 fine-tuning 된 KoBERT 모델을 사용하여 입력 문장의 CLS 토큰에 대한 임베딩 결과를 언어적 특징으로써 추출하였다.

3.3 분류기

SVM은 주어진 데이터를 분류할 수 있는 최적의 경계선을 찾는 것을 목표로 하는 알고리즘이다. SVM은 이상치에 강하고, 일반화 능력이 높으며 데이터를 고차원 공간으로 투영하여 비선형적인 분류 문제 역시 다룰 수 있다. 본 논문에서는 scikit-learn 라이브러리를 통해 SVM을 구현하였고, 비선형적인 패턴 학습에 장점을 가진 RBF(Radial Basis Function) 커널을 사용하여 성인, 노인 분류를 수행하였다.

MLP(Multilayer Perceptron)는 입력층, 은닉층, 출력층으로 구성된 인공신경망의 한 종류이다. 각 층은 뉴런이라는 단위로 이루어져 있으며, 각 뉴런은 입력값에 가중치를 곱한 후 활성화 함수(Activation Function)를 적용하여 출력값을 계산한다. 이때, 여러 개의 은닉층을 사용함으로써 비선형적인 패턴을 학습할 수 있다. 본 논문에서는 KoBERT 모델의 출력을 입력층과 은닉층의 크기가 각각 768, 출력층의 크기가 1인 MLP에 입력하고 Sigmoid 활성화 함수와 binary cross entropy 손실 함수를 적용하여 성인, 노인 분류를 수행하였다.

3.4 성인과 노인 분류 과정

본 논문에서는 발화 데이터로부터 음향적 특징과 언어적 특징을 추출하고 이를 사용해 성인, 노인에 대한 이진 분류를 수행하였다.

음향적 특징 추출 과정은 다음과 같다. 16kHz로 샘플링 된 wav 파일을 openSMILE에 입력으로 넣어, GeMAPS의 경우 60ms, eGeMAPS의 경우 25ms의 슬라이딩 윈도우를 설정하고, 이를 10ms씩 이동하며 각 시점의 윈도우에서 음향적 특징을 추출한다. 이렇게 추출된 음향적 특징 벡터 전체에 통계 함수를 적용하여 음성 신호 전반에 대한 음향적 특징의 평균, 분산 등을 구한다. 이와 같이 각 음성 샘플마다 62차원(GeMAPS), 88차원(eGeMAPS)의 음향적 특징을 추출한다.

언어적 특징 추출 과정은 다음과 같다. 자유대화 음성 데이터셋의 전사문 데이터를 KoBERT 모델에 입력하여 해당 전사문의 맥락 정보를 가진 벡터 표현을 출력한다. 이 때, 전사문에는 외부 잡음, 간투어, 불명확한 발음에 대한 어노테이션이 포함되어 있는데, 이 중 간투어, 불명확한 발음은 화자의 발화 특징을 반영한다. 이러한 특징 역시 언어 능력의 차이를 나타내는 주요한 지표로 사용될 수 있다. 따라서, 모든 어노테이션을 제거한 전사문을 KoBERT 모델의 입력으로 사용한 경우(KoBERT-A)와 간투어 및 불명확한 발음에 대한 어노테이션을 포함시킨 전사문을 KoBERT 모델의 입력으로 사용한 경우(KoBERT-B)의 언어적 특징을 추출한다.

이렇게 추출된 음향적 특징과 언어적 특징을 따로 혹은 결합해 사용해 SVM 분류기와 MLP 분류기를 학습시켜 성인과 노인 분류를 수행한다. 두 모달리티의 특징을 결합해 사용할 경우, 특징 벡터 단위 결합인 early fusion과 모델의 예측 결과값 단위 결합인 late fusion을 각각 적용하였다.

또한, 클래스 불균형이 있는 데이터셋으로 실험을 진행하였으므로 각 클래스에 대한 분류 성능에 유의미한 차이가 존재할 수 있다. 그러므로, 성인과 노인 각 클래스에 대한 F1 점수를 다운샘플링을 통해 클래스 불균형을 해소한 뒤 같은 방식으로 진행한 실험에서의 성인과 노인 각 클래스에 대한 F1 점수와 비교하였다.

4. 실험 및 결과

4.1 자유대화 데이터셋

본 논문의 실험에서는 AIHub의 자유대화 음성(일반남여) 데이터셋[19]과 자유대화 음성(노인남여) 데이터셋[20]을 사용하였다. 자유대화 음성(일반남여) 데이터셋은 59세 이하의 화자의 발화 음성 및 전사문 데이터로 구성되어 있으며, 자유대화 음성(노인남여) 데이터셋은 60세 이상의 화자의 발화 음성 및 전사문 데이터로 구성되어 있다.

자유대화 음성(일반남여) 데이터셋 및 자유대화 음성(노인남여) 데이터셋은 실내, 실외, 녹음실의 3가지 환경에서 화자의 대화를 녹음하여 얻은 발화 데이터이다. 발화 데이터는 PCM 혹은 WAV 형식의 음성 데이터와 화자의 연령, 녹음 환경, 음성 데이터 전사 결과 등의 정보를 담은 JSON 형식의 메타데이터로 이루어져 있다.

본 논문에서는 이 중 화자에게 별도로 지정된 스크립트가 없는 자유대화에 해당하는 발화 데이터만을 메타데이터를 통해 선별 및 사용하였다. 또한, 화자의 나이가 19~59세인 발화 데이터를 성인 클래스로, 화자의 나이가 60세 이상인 발화 데

Table 1. Statistics of the Dataset

Class	Train		Eval	
	#Hours	#Sents	#Hours	#Sents
Adult	46	21,907	9	4,166
Elderly	15	8,092	7	3,515
Total	61	29,999	16	7,681

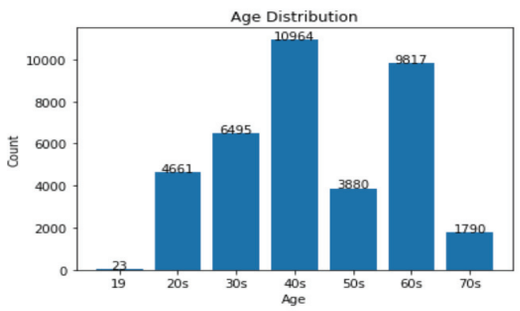

Fig. 2. Age Distribution of the Dataset

이터를 노인 클래스로 할당하였다. 음성 데이터는 약 55시간의 성인 발화 음성과 약 22시간의 노인 발화 음성으로 이루어져있으며 성인의 발화 당 평균 길이는 7.7초, 노인의 발화 당 평균 길이는 6.8초이다. 또한, 전사문 데이터는 26,073개의 성인 발화 문장과 11,607개의 노인 발화 문장으로 이루어져 있다.

성인, 노인 각 클래스에 대한 학습 데이터로 약 46시간, 약 15시간의 음성 데이터와 21,907개, 8092개의 전사문 데이터를 사용하였고, 평가 데이터로 약 9시간, 약 7시간의 음성 데이터와 4166개, 3515개의 전사문 데이터를 사용하였다.

Table 1은 앞서 언급된 바와 같이 실험에 사용된 데이터셋의 데이터 통계를 나타낸다. Train과 Eval은 각각 학습 데이터셋과 평가 데이터셋에 포함된 발화 음성 데이터의 길이와 발화 문장의 개수를 나타낸다.

Fig. 2는 자유대화 음성(일반남여) 데이터셋 및 자유대화 음성(노인남여) 데이터셋의 발화 데이터에 대한 샘플 수 및 화자 연령 분포를 나타낸다.

4.2 실험 환경

실험은 Pytorch 환경에서 진행하였고, KoBERT 모델의 fine-tuning을 위하여 adam optimizer, binary cross entropy loss를 사용하였다. 하이퍼파라미터는 batch size 32, epoch 4, learning rate 5e-5, max sequence length 128로 설정하였다.

4.3 성인과 노인 분류기의 성능

음향적 특징 기반으로 학습된 모델과 언어적 특징 기반으로 학습된 모델, 그리고 음향적 특징과 언어적 특징으로 학습된 모델을 비교하기 위해 평가 데이터셋에 대하여 각 모델의 분류 성능을 확인하였다.

Table 2는 음향적 특징을 기반으로 학습된 모델로 평가 데이터셋에 대해 성인과 노인 분류를 수행하였을 때, 특징과 분류기에 따른 성능을 나타낸다. 평가 지표로는 정밀도(P: Precision), 재현율(R: Recall), 정밀도와 재현율의 조화평균인 F1 점수(F1: F1 Score)를 사용하였다. 또한, 데이터셋의 클래스

Table 2. Evaluation Metrics for Models using Acoustic Features

Feature	Classifier	P	R	F1	(W)F1
GeMAPS	SVM	62.6	60.7	56.7	**55.4**
eGeMAPS	SVM	57.9	56.9	48.3	50.1
GeMAPS	MLP	48.4	51.7	44.7	46.4
eGeMAPS	MLP	44.8	49.1	41.7	43.4

Table 3. Evaluation Metrics for Models using Linguistic Features

Feature	Classifier	P	R	F1	(W)F1
KoBERT-A	SVM	74.3	72.6	70.9	71.5
KoBERT-B	SVM	75.2	73.5	71.9	**72.5**
KoBERT-A	MLP	72.5	70.7	69.9	69.7
KoBERT-B	MLP	74.3	72.5	70.8	71.4

비율을 고려한 가중치를 사용하여 가중 F1((W)F1: Weighted F1 Score) 점수를 함께 확인하였다.

음향적 특징을 사용한 실험에서는 GeMAPS 특징 세트에 기반한 분류 성능과 eGeMAPS 특징 세트에 기반한 분류 성능을 비교하였고, SVM 분류기와 MLP 분류기를 사용하였다.

Table 3는 언어적 특징을 기반으로 학습된 모델로 평가 데이터셋에 대해 성인과 노인 분류를 수행하였을 때, 특징과 분류기에 따른 성능을 나타낸다. 언어적 특징을 사용한 실험에서는 모든 어노테이션을 제거한 발화 전사문을 KoBERT 모델에 입력으로 사용해 추출한 언어적 특징(KoBERT-A)에 기반한 분류 성능과 간투어 및 불명확한 발음에 대한 어노테이션을 포함시킨 전사문을 KoBERT 모델에 입력으로 사용해 추출한 언어적 특징(KoBERT-B)에 기반한 분류 성능을 비교하였고, 음향적 특징 기반 분류와 마찬가지로 SVM 분류기와 MLP 분류기를 사용하였다. 실험 결과에서는, KoBERT-B와 SVM 분류기를 사용해 성인과 노인을 분류한 결과로 72.5%의 Weighted F1 점수를 보였다. 이는 선행 연구[6]에서 KoBERT-B와 MLP 분류기를 사용해 얻은 71.4%의 Weighted F1 점수보다 1.1% 높은 수치이다.

Table 4는 음향과 언어 각 모달리티에서 가장 좋은 분류 성능을 보였던 GeMAPS 및 KoBERT-B와 SVM 분류기를 사용한 성인과 노인 분류 성능을 나타낸다. 음향적 특징과 언어적 특징을 연결하여 하나의 벡터로 만들어 이를 기반으로 학습하는 early fusion 모델의 분류 성능과 각 특징을 학습한 두 모델의 결과값을 종합하여 예측하는 late fusion 모델의 분류 성능을 확인하였다.

Fig. 3은 각 모델의 weighted F1 점수를 나타낸다. 성인 및 노인 연령대 분류 실험 결과에서 음향적 특징 기반 모델은 최대 55.4%의 weighted F1 점수를 보였다. 반면에, 언어적 특징 기반 모델은 최대 72.5%의 weighted F1 점수를 보이며,

Table 4. Evaluation Metrics for Models using Acoustic and Linguistic Features

Feature	Fusion	Classifier	P	R	F1	(W)F1
GeMAPS & KoBERT-B	Early Fusion	SVM	73.4	71.7	69.9	**70.6**
	Late Fusion	SVM	70.4	63.3	56.2	57.7

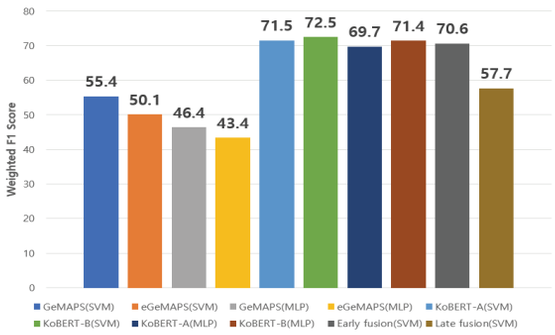

Fig. 3. Weighted F1 Score for Each Model

음향적 특징 기반 모델보다 17.1% 더 높은 성능을 보였다. 간투어 및 불명확한 발음에 대한 어노테이션을 추가로 사용하여 추출된 KoBERT-B 특징을 사용한 모델은 최대 72.5%의 weighted F1 점수를 보여, 어노테이션 정보를 제거하여 추출된 KoBERT-A 특징을 사용한 모델의 성능인 71.5% 보다 1.0%높은 weighted F1 점수를 보였다.

결과적으로, 성인과 노인 연령대 분류에 있어 음향적 특징을 사용하였을 때 혹은 음향적 특징과 언어적 특징을 함께 사용하였을 때보다 언어적 특징을 단일로 사용하였을 때 더 높은 분류 성능을 보임을 확인할 수 있었다. 또한, 언어적 특징을 기반으로 분류하는 데 있어 간투어 및 불명확한 발음 등 발화 오류에 대한 어노테이션이 분류 성능 향상에 도움이 됨을 확인할 수 있었다. 마지막으로, 음성적 특징과 언어적 특징 모두 MLP 분류기보다 SVM 분류기를 통해 분류하였을 때 높은 성능을 보였다.

4.4 성인과 노인 각 클래스에 따른 분류 성능

실험에 사용된 데이터셋에 클래스 불균형이 존재하여 음향적 특징(GeMAPS, eGeMAPS)을 사용한 모델과 언어적 특징(KoBERT-A, KoBERT-B)을 사용한 모델의 성인과 노인 각 클래스에 대한 F1 점수를 비교하였다.

Table 5는 음향적 특징을 기반으로 학습된 모델과 언어적 특징을 기반으로 학습된 모델로 평가 데이터셋에 대하여 성인과 노인을 분류하였을 때의 각 클래스에 대한 F1 점수를 나타낸다. 전반적으로 성인에 대해서는 상대적으로 잘 분류하고 있으나, 노인에 대한 F1 점수가 더 낮은 것을 확인할 수 있다.

각 클래스에 대한 F1 점수의 차이가 클래스 불균형으로 인한 것인지 알아보기 위해 추가 실험을 진행하였다. 추가 실험에서는 다운샘플링을 통해 각 클래스별 학습 데이터의 양을 노인 데이터에 맞춰 약 15시간 길이의 음성 데이터와 8,092개의 전사문 데이터로 설정하였다.

Table 6은 Table 5와 동일한 특징과 분류기로 클래스 불균형이 해소된 데이터셋을 통해 학습하고 평가 데이터셋에 대하여 성인과 노인을 분류한 결과로 나온 F1 점수이다. Fig. 4를 통해 여전히 성인보다 노인에 대한 F1 점수가 낮지만, 클래스 불균형 데이터셋을 사용하였을 경우보다 각 클래스별 F1 점수의 차이가 줄어든 것을 확인할 수 있다.

Table 5. F1 Score for Imbalanced Data

Feature	Classifier	Adult	Elderly
GeMAPS	SVM	70.8	40.0
eGeMAPS	SVM	69.3	27.3
KoBERT-A	SVM	77.9	63.9
KoBERT-B	SVM	**78.5**	**65.2**

Table 6. F1 Score for Balanced Data

Feature	Classifier	Adult	Elderly
GeMAPS	SVM	66.4	57.7
eGeMAPS	SVM	64.5	50.1
KoBERT-A	SVM	77.4	70.2
KoBERT-B	SVM	78.4	71.8

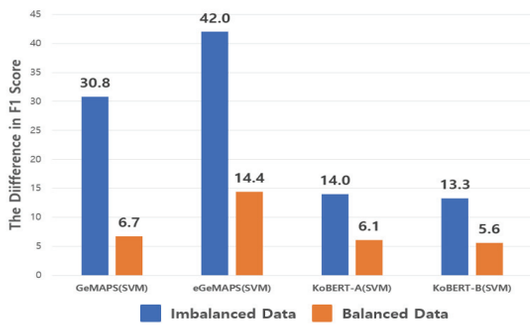

Fig. 4. The Difference in F1 Scores between Classes

5. 결 론

본 논문은 실험을 통해 음향적 특징 기반의 분류 성능과 언어적 특징 기반의 화자 연령 분류 성능을 비교하였다. 분석 결과, 언어적 특징을 사용하였을 때의 분류 성능이 더 높은 것으로 나타났다. 음향적 특징은 음성 신호의 주파수, 세기와 같은 특징을 포함하는데, 음성 신호에는 잡음, 반향, 간섭 등과 같은 외부 요인으로 인한 불확실성이 존재한다. 반면에, 언어적 특징은 문법, 어휘, 문장구조 등과 같은 특징을 포함하며, 이 특징들은 연령에 따라 큰 차이를 보일 수 있다[1]. 이러한 이유로 인해, 음향적 특징을 사용하였을 때보다 언어적 특징을 사용하였을 때 분류 성능이 높게 나온 것으로 보인다. 또한, 발화 오류에 대한 어노테이션 정보를 주었을 때 언어적 특징 기반 분류 성능이 올라가는 것으로 확인되었다. 마지막으로, 성인과 노인 각 클래스에 대한 F1 점수를 살펴본 결과, 각 클래스의 데이터 샘플 수를 동일하게 설정하더라도 노인에 대한 F1 점수가 성인에 비해 낮다는 것을 확인했다. 향후 연구에서는 화자 연령 분류 기술의 실용성을 높이기 위해 연령별 다중분류로 접근할 필요성이 있다. 또한, 한국어 발화 데이터셋에서 노인에 대한 분류 성능을 높이는 방법을 찾고, 음향적 특징 및 언어적 특징을 함께 활용하는 방법에 대한 연구가 진행되기를 기대한다.

References

[1] J. W. Kim and H. H. Kim, "Communicative ability in normal aging: A Review," *Korean Journal of Communication Disorders*, Vol.14, No.4, pp.495-513, 2009.

[2] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, "Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features," *Computer Speech & Language*, Vol.53, pp.181-197, 2019.
[3] M. R. Morales and R. Levitan, "Speech vs. text: A comparative analysis of features for depression detection systems," *2016 IEEE spoken language technology workshop (SLT)*, IEEE. 2016.
[4] I. Vigo, L. Coelho, and S. Reis. "Speech-and language-based classification of alzheimer's disease: A systematic review," *Bioengineering*, Vol.9, No.1, pp.27, 2022.
[5] M. Ehghaghi, F. Rudzicz, and J. Novikova, "Data-driven approach to differentiating between depression and dementia from noisy speech and language data," *arXiv preprint arXiv:2210.03303*, 2022.
[6] S. H. Han, S. H. Dong, and B. O. Kang, "Comparison of classification performance between adult and elderly using acoustic and linguistic features from spontaneous speech," in *Proceedings of the Korea Conference on Software Engineering (KCSE) 2023*, Vol.25, pp.117-118, 2023.
[7] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," *Proceedings of the 18th ACM International Conference on Multimedia*, pp.1459-1462, 2010.
[8] SKTBrain, "KoBERT", github repository, accessed Dec. 16, 2022, [Internet], https://github.com/SKTBrain/KoBERT
[9] F. Rangel, F. Celli, P. Rosso, M. Potthast, B. Stein, and W. Daelemans, "Overview of the 3rd Author Profiling Task at PAN 2015," *Conference and Labs of the Evaluation Forum*, 2015.
[10] A. Liesenfeld, G. Parti, Y. Y. Hsu, and C. R. Huang, "Predicting gender and age categories in English conversations using lexical, non-lexical, and turn-taking features," *arXiv preprint arXiv:2102.13355*, 2021.
[11] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[12] B. P. R. Guda, A. Garimella, and N. Chhaya. "Empathbert: A bert-based framework for demographic-aware empathy prediction," *arXiv preprint arXiv:2102.00272*, 2021.
[13] B. Schuller et al., "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, Vol.1, No.2, pp.119-131, 2010.
[14] F. Eyben et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, Vol.7, No.2, pp.190-202, 2015.
[15] B. Schuller et al., "The INTERSPEECH 2010 paralinguistic challenge," 2010.
[16] F. Burkhardt, M. Brückl, and B. Schuller, "Age classification: Comparison of human vs machine performance in prompted and spontaneous speech," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2021*, pp.35-42, 2021.
[17] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?," *arXiv preprint arXiv:1906.01502*, 2019.
[18] S. Lee, H. Jang, Y. Baik, S. Park, and H. Shin, "Kr-bert: A small-scale korean-specific language model," *arXiv preprint arXiv:2008.03979*, 2020.
[19] "자유대화 음성(일반남여)," AI-Hub, accessed Dec. 16, 2022, https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=109
[20] "자유대화 음성(노인남여)," AI-Hub, accessed Dec. 16, 2022, https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=107

한 승 훈
https://orcid.org/0009-0007-0325-1832
e-mail : tht1102@etri.re.kr
2021년 충남대학교 독어독문학과(학사)
2021년 ~ 현 재 고려대학교 뇌공학과 석사과정
관심분야 : 머신러닝, 헬스케어

강 병 옥
https://orcid.org/0009-0001-8217-720X
e-mail : bokang@etri.re.kr
1997년 포항공과대학교 전기전자공학과(학사)
1999년 포항공과대학교 전기전자공학과(석사)
2017년 충북대학교 전기・전자・정보・컴퓨터학부(박사)
1999년 ~ 2001년 삼성전자 무선사업부
2001년 ~ 현 재 한국전자통신연구원 책임연구원
관심분야 : 인공지능, 머신러닝, 음성인식, 헬스케어

동 성 희
https://orcid.org/0000-0003-4092-506X
e-mail : dsh7560@etri.re.kr
2014년 고려대학교 물리학과(학사)
2019년 고려대학교 뇌공학과(박사)
2019년 ~ 현 재 한국전자통신연구원 선임연구원
관심분야 : 인공지능, 복합지능, 뇌인지과학, 음성처리