2022년도

학사학위논문

음성을 이용한 성별 구별 시스템 개발

Development of a Gender Classification System using Voice Data

2022년 11월 22일

순천향대학교 공과대학

컴퓨터공학과

이수빈

음성을 이용한 성별 구별 시스템 개발

Development of a Gender Classification System using Voice Data

지도교수 남 윤 영

이 논문을 공학사 학위논문으로 제출함

2022년 11월 22일

순천향대학교 공과대학
컴퓨터공학과
이수빈

이 수 빈의 공학사 학위 논문을 인준함

2022년 11월 22일

심 사 위 원 남 윤 영 인

심 사 위 원 이 상 정 인

순천향대학교 공과대학

컴퓨터공학과

초 록

최근에는 목소리를 다른 성별로 변조하여 자신의 신분을 위조하는 범죄가 발생하고 있는데, 이와 같은 방식을 사용하면 발신자의 신분과 의도를 알아채기 어렵다. 이처럼 발신자의 신분을 위조하는 교활한 방법으로 보이스 피싱 사기 범죄 건수는 날마다 증가하는 추세이다. 음성으로 성별을 구분하는 성별 인식 시스템을 구현하여 이러한 보이스 피싱 범죄 절감에 기여하고자 본 연구를 진행하게 되었다. 이 연구의 목적은 선행 연구보다 더 높은 정확도를 얻는 것과 여러 머신러닝 모델을 이용해서 학습해 각 모델의 성능을 측정해서 가장 적합한 모델을 선별하는 것이다.

본 논문에서는 음성 데이터를 입력하면 성별을 판별하는 프로그램을 구현한다. ANN과 Logistic Regression, KNN, SVM-Linear, Random Forest를 이용해 모델링을 하고 혼동 행렬 결과와 정확도, 모델 학습 시간 비교해서 성능을 비교해 보았다. ANN으로 구현한 시스템의 정확도 값이 99%로 가장 높았기 때문에 ANN이 가장 적합한 모델로 선정되었으며, 이는 선행 연구보다 더 높은 수치이다. 이후 ANN을 이용해 모델링한 시스템에 새로운 음성 데이터를 입력해 실험하였다.

주요어 : 머신러닝, 음성 데이터, ANN, 모델링, 정확도

ABSTRACT

Recently, there have been crimes of disgusing one's identity by modifying one's voice to another gender. This method make it difficult to recognize the sender's identity and intention. In this way, the number of voice phishing fraud crimes is increasing every day in a sneaky way that falsifies the identity of the sender. This study was conducted to contribute to these voice phishing crimes by developing a gender recognition system that classifies gender by voice. The purpose of this study is to obtain higher accuracy than previous studies and to measure the performance of each model using several machine learning models to find which model is the most suitable.

In this thesis, we build a program that determines gender when voice data is input. We modeled using ANN, Logical Regression, KNN, SVM-Linear, Random Forest, and NN, and compared the performance with the result of the confusion matrix, accuracy, and model learning time. ANN, whose accuracy value of the system was 99%, was selected as the most suitable model. This value of accuracy is higher than previous studies and the test was implemented by entering new voice data into the system modeled using ANN.

keywords : machine learning, voice data, ANN, modeling, accuracy

차 례

그 림 차 례

표 차 례

수 식 차 례

제 1 장 서 론

1.1. 연구의 필요성

최근 고도화된 범죄 수법으로 보이스 피싱[1]을 구별하기 점점 어려워지고 있다. 최근에는 목소리를 다른 성별로 변조하여 자신의 신분을 위조하는 범죄까지 발생하고 있는데, 이와 같은 방식을 사용하면 발신자의 신분과 의도를 알아채기 어렵다. 이처럼 발신자의 신분을 위조하는 교활한 방법으로 보이스 피싱 사기 범죄 건수는 날마다 증가하는 추세이다. 하지만 나날이 증가하는 보이스 피싱의 심각성과 달리 국내외를 막론하고 범죄 수사에 활용할 수 있는 음성인식에 대한 딥러닝[2] 기술 연구 및 적용은 충분히 이루어지지 않고 있다. 현재 4차 산업 혁명의 주역인 딥러닝과 머신러닝에 대한 연구가 활발히 진행되어 뛰어난 성능의 모델들이 개발되고 있음에도 불구하고, 사진이나 동영상 등에 비해 음성과 관련된 연구 성과는 그다지 많지 않기 때문이다. 이와 같은 배경을 바탕으로 음성으로 성별을 구분하는 인공지능을 구현하여 보이스 피싱 범죄에 기여하고자 본 연구를 진행하게 되었다.

---

1) 보이스 피싱(voice phishing) : 주로 금융 기관이나 유명 전자 상거래 업체를 사칭하여 불법적으로 개인의 금융 정보를 빼내 범죄에 사용하는 범법 행위, 음성(voice)과 개인 정도(private data), 낚시(fishing)를 합성한 용어이다.

2) 딥 러닝(deep learning) : 심층학습이라고도 불리며, 기계학습 기술의 종류 중 하나인 인공신경망을 수많은 계층 형태로 연결한 기법을 뜻하며 현대 인공지능의 기술의 핵심 기술이다.

1.2. 연구의 목적

이 연구의 목적은 여러 머신러닝 모델을 이용해 학습된 각 모델의 성능을 측정함으로써 어느 모델이 가장 적합한지 구하여 최종적으로 음성의 성별을 구분하는 머신러닝 시스템을 구현하는 것이다. 이 과정에서 선행 연구보다 더 높은 정확도를 얻는 것을 목표로 하며, 나아가 실제 활용될 수 있는 분야에 대한 적용 방안도 검토하고자 한다. 제1장 서론을 시작으로 제2장 본론에서는 관련 연구 방법을 기술하고, 제3장에서 프로그램 구현, 제4장에서는 생성한 5가지의 모델에 대한 성능을 평가하고 비교해서 가장 성능이 좋은 모델을 선정하고자 한다. 마지막 제5장에서는 결론 및 향후 과제를 제시한다.

제 2 장 관련 연구

2.1 선행 연구의 음성 분석 방식

선행 연구에서는 주로 RNN이나 LSTM을 기반으로 음성을 분석했다.[1] RNN은 고정 길이 입력이 아닌 임의의 길이를 가진 시퀀스를 다룰 수 있다는 점에서 다루기 어려운 음성인식에서 자주 사용된다. 최근 LSTM도 많은 연구가 진행되고 있는데, RNN에 비해서 더 많은 시간과 데이터를 요구하지만, 자원이 충분히 확보된다면 더 좋은 성능을 보인다. 하지만 LSTM은 역전파 과정에서 발생하는 지연으로 인해 실시간으로 서비스를 제공하는 것이 어렵다는 단점이 존재한다.

한국어는 교착어라는 특성 때문에 타국어보다 음성인식 모형 구현이 어렵다. [3]교착어란 언어 유형학적 분류의 한 갈래로서, 실질 형태소인 어근(語根, root)에 형식 형태소인 접사(接辭, affix)를 붙여 단어를 파생시키거나 문법적 관계를 나타내는 언어를 가리킨다. 예를 들어, '되-'라는 어근은 접사에 의해 '되다', '된다', '됐다' 등 여러 형태로 파생될 수 있다. 교착어는 단어 단위로 훈련이 어렵기 때문에 이러한 중간 단계를 생략하여 학습하는 종단 간 학습이 제안되었다. 종단 간 모형에 베이즈 딥러닝을 결합해서 전보다 더 좋은 성능을 가지는 새로운 모형을 구현하였다. 이 모형을 적용했을 때 대부분 모형에서 더 좋은 성능을 보였고, 특히 CTC[4]에서 가장 성능이 뛰어난 것을 확인할 수 있었다.

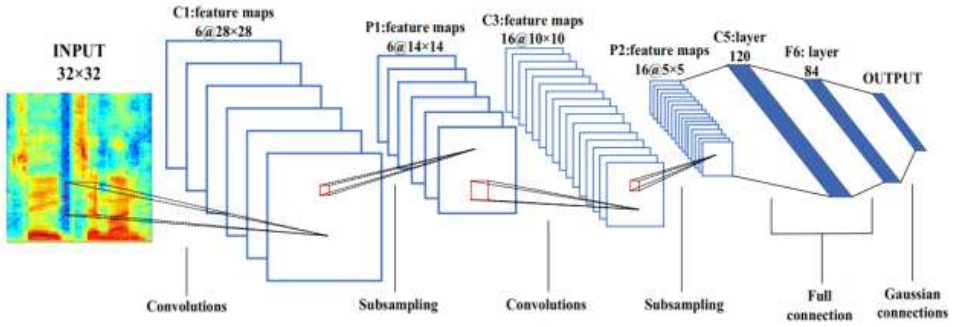
---

3) 교착어. NAVER 지식백과.
https://terms.naver.com/entry.naver?docId=939143&cid=47319&categoryId=47319

4) Connectionist Temporal Classification, 입력 음성 프레임 시퀀스와 타겟 단어

2.2 선행 연구 결과

CNN을 이용해 성별 인식 시스템 개발을 한 선행 연구가 있다. 6개의 레이어로 구성되었고 특징 추출은 MFCC를 사용하여 실행되었다.



5)[그림 1] 선행 연구에서 사용된 CNN 구조

이 연구에서는 흔하지 않은 음성 신호 인식을 위한 CNN 기반의 새로운 접근법을 제시한다. 연령, 억양, 환경이 다른 11명의 화자로부터 녹음된 데이터로 학습을 진행했다. 각 데이터는 38개로 구성되어 있으며 총 데이터의 개수는 418개다. Kaldi 툴킷을 이용해 음성 인식을 수행한다. 다음 단계로 진행되었다.

매개변수: Alpha - 0.333, 편향 = 0.1 & 축 = −1

입력: 샘플 오디오.wav

1단계: PRAAT 프로그램을 사용하여 스펙트로그램 및 파형 생성

/음소 시퀀스 간에 명시적인 조정 정보 없이도 음성인식 모델을 학습할 수 있는 기법

5) Sakshi Dua 외 5명. 2022. “Developing a Speech Recognition System for Recognizing Tonal Speech Signals Using a Convolutional Neural Network”

2단계: LIBROSA Python 라이브러리를 사용하여 MFCC 추출

3단계: CNN, 6-2D Conv를 사용한 기능 학습 그리고 2개의 FC 레이어

4단계: Conv2D와 Dense Layer(256 Units) 사이에 Flatten Layer 포함

5단계: Softmax 활성화 함수에 의한 활성화

6단계: 'Categorical Cross Entropy'에 의한 손실 계산

7단계: diff(li1, li2)로 입력(li1)을 Ground Truth(gt)-li2와 비교

Speakers	Sex	Recognition Accuracy
Speaker 1 (sp1)	F	90.87%
Speaker 2 (sp2)	F	89.871%
Speaker 3 (sp3)	F	89.783%
Speaker 4 (sp4)	M	88.559%
Speaker 5 (sp5)	F	88.370%
Speaker 6 (sp6)	F	89.077%
Speaker 7 (sp7)	M	90.911%
Speaker 8 (sp8)	F	88.942%
Speaker 9 (sp9)	M	86.765%
Speaker 10 (sp10)	F	86.677%
Speaker 11 (sp11)	F	90.87%
Overall accuracy of system	-	**89.153%**

[그림 2] 선행 연구의 정확률

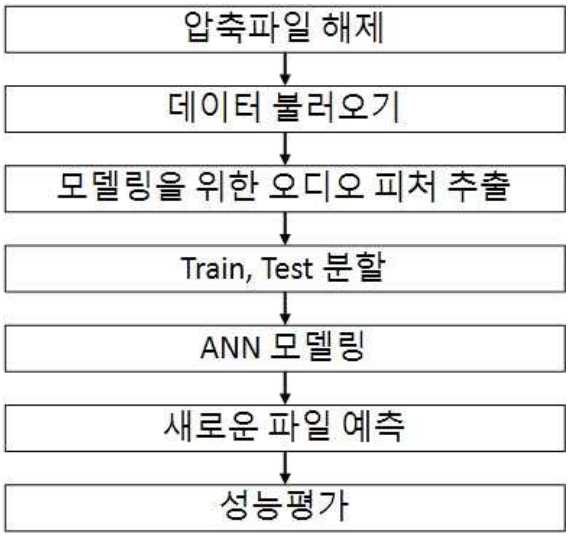
[그림 2]는 해당 선행 연구의 음성 인식 정확률이다. 평균 정확도는 0.89153이다.

제 3 장 프로그램 구현

3.1. 구현 방식

프로그램은 Python으로 개발하였고 케라스와 사이킷런을 이용해 구현하였다. 머신러닝 시스템인 Logistic Regression, KNN, SVM-Linear, Random Forest, NN, ANN을 사용하여 모델링하고 성능을 비교했다. 비교 군내에서 ANN의 성능이 가장 높게 측정되어 ANN을 사용하여 모델을 구현하였다.

[그림 1]은 ANN을 사용해서 구현한 발화자 성별 구별 프로그램 플로우 차트이다.



[그림 3] 프로그램 플로우 차트

3.2 모델 구현

3.2.1 ANN 구현

ANN(Artificial neural network)은 신경 세포인 뉴런을 모방한 모델로 딥러닝의 핵심적인 기술로 볼 수 있으며 음성인식 분야에서 뛰어난 성능을 발휘한다. 이 ANN 알고리즘을 이용해 프로그램을 구현하였다. 먼저 남자와 여자의 음성 데이터를 각각 불러와서 male_df, female_df 데이터 프레임으로 만든다. 그리고 남녀의 데이터를 통합한다.

```
male      1000
female    1000
Name: gender, dtype: int64
```

	filename	gender
1696	./여자/여자6_전라실내/일반남여_일반통합11_F_1528170230_40_전라_실...	female
1414	./여자/여자4_수도권실내2/일반남여_일반통합07_F_1524044702_43_수도...	female
1588	./여자/여자5_경상실내/일반남여_일반통합11_F_1527844391_41_경상_실...	female

[그림 4] 음성 데이터 종류별 개수

그 후, 오디오 파일을 불러와서 모델링을 위해 MFCC(Mel-Frequency Cepstral Coefficients) 피처를 추출한다. MFCC는 음성 정보에서 불필요한 정보를 제거해서 중요한 특질만 가진 피처이다.[5] 음성 데이터를 특징 벡터화해준다고 볼 수 있는데, 음성 분석을 하기 위한 더 정확한 학습이 가능하게 한다. 평균으로 스케

일된 피처를 저장하고 저장된 데이터 셋에서 음성 신호가 가지고 있는 특징을 뽑는다. 각 오디오 파일의 모든 특징을 추출하기 위해 데이터 프레임의 각 행에 대한 루프를 사용하면서 tqdm 파이썬 라이브러리를 사용해 작업의 진행 상황을 추적한다. MFCC 과정을 통해 뽑힌 특징들을 가시화하기 위해 데이터 프레임으로 저장한다.

	feature	class
42	[-301.83716, 96.580025, -27.539726, 28.864597,...	male
718	[-402.9354, 121.86536, 4.500896, 32.377666, 1....	male

[그림 5] MFCC 과정을 통해 추출된 특징

이후 모델링을 위한 X(feature)와 Y(Label) 데이터 분리 작업을 진행한다. 라벨 인코더를 이용해 숫자로 변환하는데 to_categorical 함수를 이용해 원-핫 인코딩을 해준다. 훈련과 테스트 셋은 8:2 비율로 분리한다.

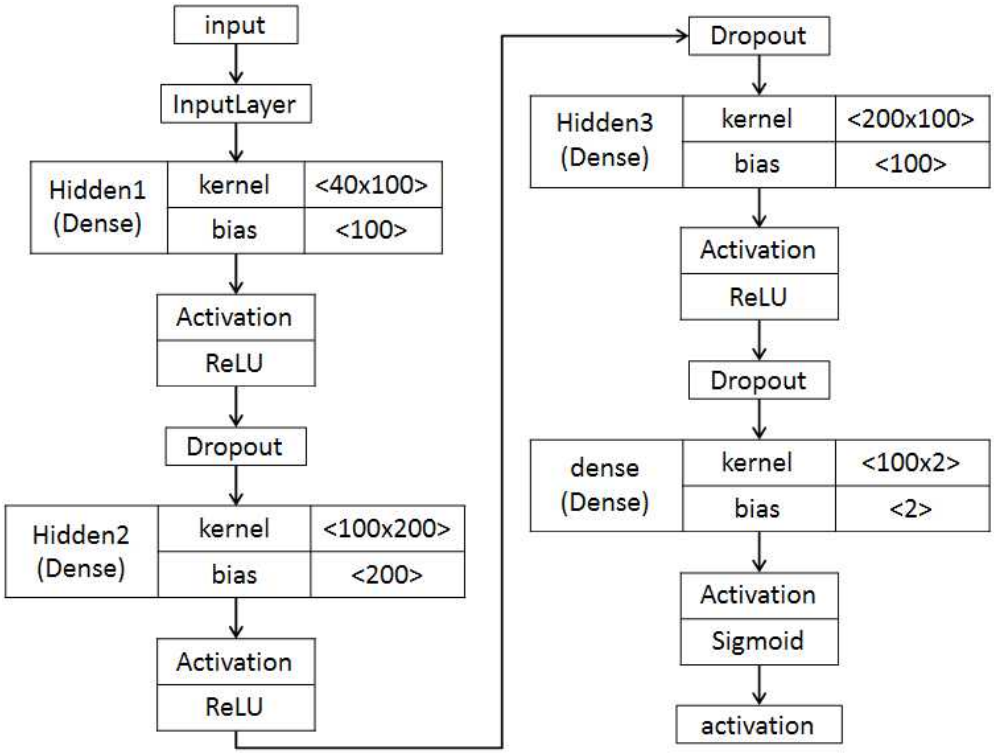
[표 1] 데이터셋 구성

Class	Train	Valid	Test
Male	800	200	200
Female	800	200	200

Trainig Set을 K-fold 방식을 사용하여 데이터를 나누지 않고 test 데이터를 validation 데이터로 사용했다.

ANN 모델링을 하기 위해 케라스 딥러닝 프레임워크를 사용했다. 그 중에서도 케라스의 sequential 함수를 사용했다. 중간 은닉층은 3개가 존재하는데 은닉층 모두 음수를 0으로 만드는 활성화 함수인 relu를 사용하였다. 특정 feature만 과도하게 집중해 학습해서 발생하는 과대적합(Overfitting)을 방지하기 위해 Drop-out을 설정했다. Drop-out Rate는 0.5로 설정하여 뉴런 별로 0.5의 확률로 제거되게 했다. 그리고 2개의 클래스를 분류하는 모델이므로 sigmoid 함수로 최종 레이어를 선정하였다. 마지막 시그모이드 함수는 결과의 임의 값을 [0,1] 사이로 분류해주었다. 남, 여 이진분류이므로 binary_crossentropy를 손실함수로 사용했고 지표는 정확도, 최적화 함수는 adam을 활용했다.

모델 구조는 다음과 같다.



[그림 6] 본 연구의 제안 신경망

epoch는 100으로, 배치 사이즈는 32로 모델을 학습시켰다. 모델 가중치의 내용은 다음과 같다.

[표 2] 모델 가중치

kernel	<100x2>	<200x100>	<100x200>	<40x100>
bias	<2>	<100>	<200>	<100>

이제 새로운 파일을 예측하기 위해 이전과 동일하게 특징 추출 과정을 거쳐 MFCC를 2d 배열로 변환한다. 새로운 파일 예측 후에 성능 평가를 한다.

3.2.2 머신러닝 모델링

Logistic Regression, KNN, SVM-Linear, Random Forest 모델을 모델링했다. Train과 Text 데이터를 분할하고 하이퍼 파라미터 튜닝 함수를 정의하였다. 더 견고하게 모델을 예측하기 위해 K-fold로 데이터를 분할하고 후보군들의 조합 중 최적의 조합을 찾기 위해서 GridSearch를 이용해 하이퍼 파라미터 튜닝을 한다. 각 모델들을 다음과 같이 튜닝했다.

[표 3] Logistic Regression 하이퍼 파라미터 튜닝

model	LogisticRegression()
params	penalty : ['l1', 'l2'], C : [0.1, 1, 10]

Logistic Regression은 penalty, C 2가지를 튜닝했다.
penalty는 어떤 규제를 적용할 것인지에 관한 것으로, l1과 l2 규제를 선택하였다. C는 Cost를 뜻하고 값이 클수록 훈련을 더 복잡하게 하면서 규제가 약해진다. C의 값으로 0.1, 1, 10을 주었다.

[표 4] KNN 하이퍼 파라미터 튜닝

model	KNeighborsClassifier()
params	n_neighbors : [3, 5, 7]

KNN은 검색할 이웃의 수인 n_neighbors을 3, 5, 7로 설정했다.

[표 5] SVM-Linear 하이퍼 파라미터 튜닝

model	SVC(kernel = 'linear')
params	C: [1, 10, 100]

SVM-Linear는 C 값을 1, 10, 100으로 조정했다.

[표 6] Random Forest 하이퍼 파라미터 튜닝

model	RandomForestClassifier()
params	n_estimators': [100, 300, 500], 'max_depth': [3, 5, 7], 'min_samples_split': [5, 7, 9]

RandomForest는 n_estimators, max_depth, min_samples_split 3가지를 튜닝했다. n_estimators는 결정트리의 개수를 지정해주는데 100, 300, 500개로 지정했다. max_depth는 트리의 최대 깊이로 3, 5, 7로 설정했다. min_samples_split은 노드를 분할하기 위해 필요한 최소한의 샘플 데이터 수로 5, 7, 9로 설정했다.
그 후에는 각 모델의 성능을 평가를 위해 혼동 행렬과 정확도, f1 score를 출력하였다.

제 4 장 실험 및 결과

4.1 데이터

[6] AI-Hub의 자유 대화 음성(일반 남, 여)의 데이터를 일부 사용하였다. 총 2,000개의 데이터가 사용되었고, 더 정확한 결과를 위해 여러 지역, 실내/실외, 10~50대로 구성된 발화자들의 데이터를 사용하였다. 한 명당 100개, 남녀 각각 10명의 음성 데이터로 구성되어 있다. 각 음성 데이터는 약 5~10초의 길이를 가지고 있다.

[표 7] 남성 음성 데이터 구성 요소

지역	서울	전라도	경상도_실내		제주도	강원도	충청도	서울_실내	
나이	38	27	33	19	41	43	55	26	17
개수	100	100	100	100	100	100	100	200	100

[표 2]는 남성 음성 데이터의 구성 요소이다. 실내라고 표기된 데이터 외에는 모두 실외에서 녹음되었다. 나이는 10대 2명, 20대 3명, 30대 2명, 40대 2명, 50대 1명으로 구성되어 있고, 지역은 서울 4명, 전라도 1명, 경상도 2명, 제주도 1명, 강원도 1명, 충청도 1명으로 구성되어 있다.

---

6) https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=109

[표 8] 여성 음성 데이터 구성 요소

지역	서울		전라도_실내		경상도_실내		제주도	강원도	충청도	
나이	47	22	31	42	33	52	29	15	36	27
개수	100	100	100	100	100	100	100	100	100	100

[표 3]은 여성 음성 데이터의 구성 요소이다. 나이는 10대 1명, 20대 3명, 30대 3명, 40대 2명, 50대 1명으로 구성되어 있고, 지역은 서울 2명, 전라도 2명, 경상도 2명, 제주도 1명, 강원도 1명, 충청도 2명으로 구성되어 있다.

4.2 실험 결과

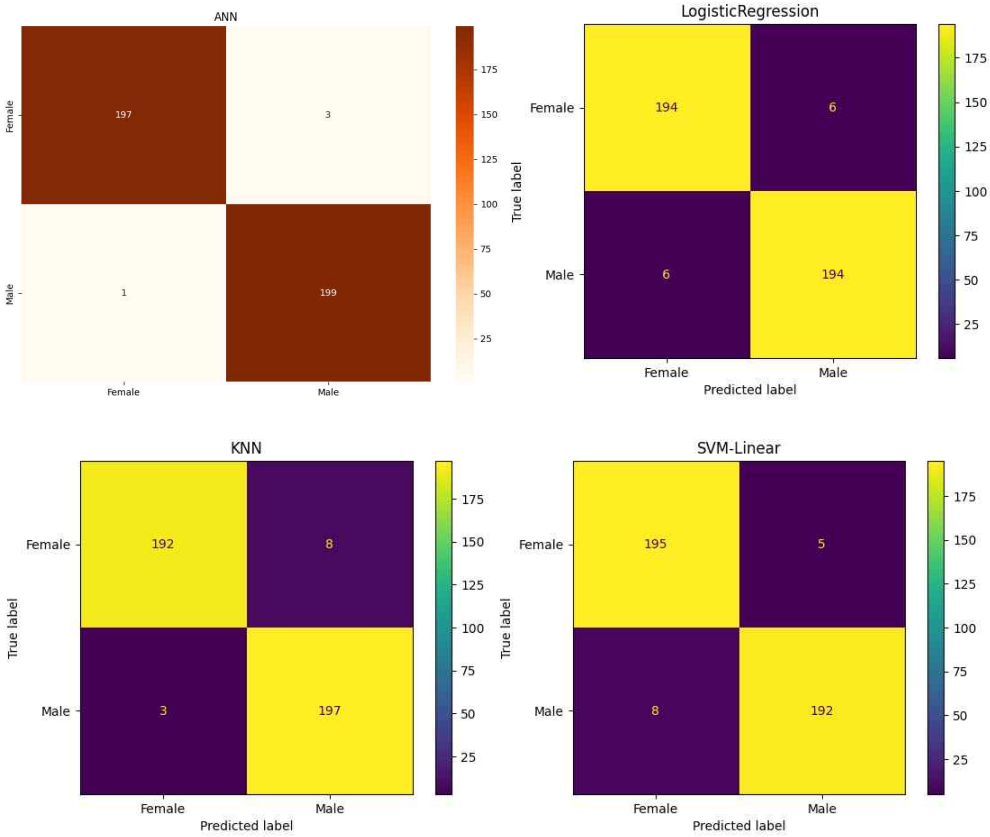
4.2.1 혼동행렬 결과

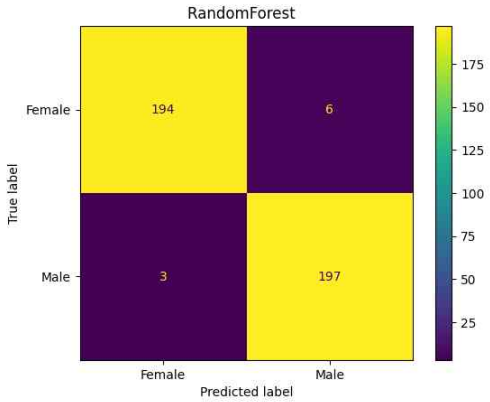
혼동행렬이란 모델의 성능을 평가할 때 사용되는 지표로, 예측값이 실제 관측값을 얼마나 정확히 예측했는지 보여준다.

[표 9] 혼동 행렬(Confusion matrix)

		Actual Class	
		Positive (P)	Negative (N)
Predicted Class	Positive (P)	TP	FN
	Negative (N)	FP	TN

모델들의 혼동행렬은 다음과 같다.


ANN
197
3
1
199
Female
Male
LogisticRegression
194
6
6
194
True label
Predicted label
KNN
192
8
3
197
SVM-Linear
195
5
8
192



[그림 7] 모델의 혼동 행렬

정확도는 모델이 입력된 데이터에 대해 얼마나 정확하게 예측하는지를 나타내며 다음과 같이 구한다.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

정밀도는 모델의 예측값이 얼마나 정확하게 예측됐는가를 나타내는 지표이다.

$$\mathrm{Precision} = \frac{TP}{TP+FP} \quad (2)$$

민감도는 실제 값 중에서 모델이 검출한 실제 값의 비율을 나타낸다.

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

특이도는 현실이 실제로 부정일 때 예측 결과도 부정일 확률이다.

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

[표 10] 혼동행렬 결과

	정확도	정밀도	민감도	특이도	f1-score
ANN	0.99	0.985	0.9949	0.9851	0.9899
Logistic Regression	0.97	0.97	0.97	0.97	0.97
KNN	0.9725	0.96	0.9849	0.961	0.9722
SVM-Linear	0.9675	0.975	0.9606	0.9746	0.9677
Random Forest	0.9775	0.97	0.9848	0.9704	0.9733

[표 2]는 각 모델들의 혼동행렬 결과값이다.

ANN의 정확도는 0.99, 정밀도는 0.985, 민감도는 0.9949, 특이도는 0.9851, f1 score는 0.9899가 나왔다. Logistic Regression의 정확도는 0.97, 정밀도는 0.97, 민감도는 0.97, 특이도는 0.97, f1 score는 0.97이 나왔다. KNN의 정확도는 0.9725, 정밀도는 0.96, 민감도는 0.9846, 특이도는 0.961, f1 score는 0.9722가 나왔다. SVM-Linear의 정확도는 0.9675, 정밀도는 0.975, 민감도는 0.9606, 특이도는 0.9746, f1 score는 0.9677이 나왔다. Random Forest의 정확도는 0.9775, 정밀도는 0.97, 민감도는 0.9848, 특이도는 0.9704, f1 score는 0.9733이 나왔다.

4.2.2 정확도 비교

[표 11] 모델별 정확도

	ANN	Logistic Regression	KNN	SVM-Linear	Random Forest
정확도(%)	99%	97%	97.25%	96.75%	97.75%

정확도는 ANN > Random Forest > KNN > Logistic Regression > SVM-Linear 순으로 높았다.

[표 12] 모델 별 f1 score

	ANN	Logistic Regression	KNN	SVM-Linear	Random Forest
f1 score (%)	98.99%	97%	97.22%	96.77%	97.73%

f1 score란 정밀도와 재현율을 활용한 평가용 지수로 모델의 성능을 측정하는데 사용한다. 정밀도와 재현율이 비슷할수록 f1 score도 높아지므로 f1 score가 높을수록 성능이 좋다.

f1 score는 정확도와 마찬가지로 ANN > Random Forest > KNN > Logistic Regression > SVM-Linear 순으로 높았다.

4.2.3 모델 학습 시간 비교

[표 13] 모델별 학습 시간

	ANN	Logistic Regression	KNN	SVM-Linear	Random Forest
time(sec)	12.06	2.023	0.409	4.563	101.2

학습 시간은 Random Forest > ANN > SVM-Linear > Logistic Regression > KNN 순으로 높았다.
Random Forest가 1분 이상으로 압도적으로 높았고, ANN은 12.06초, SVM-Linear은 4.563초, Logistic Regression은 2.023초, KNN은 0.409초로 학습 시간이 매우 짧았다.

정확도 0.99, f1 score 0.98의 ANN의 성능이 가장 좋다는 결론이 나왔다. 비록 학습 시간이 12.06초로 비교적 높은 수치를 보였지만, 이는 성능에 크게 영향을 미치지 않으며 더 중요한 정확도와 f1 score가 다른 모델보다 훨씬 높았기 때문에 ANN의 성능이 가장 좋다고 볼 수 있다.

4.2.4 테스트 결과

```
filename="여자_test1.wav"

infer_gender(filename)[0]
```

✓ 0.6s

```
'female'
```

[그림 8] 여성 음성 테스트 결과1

20대 여성의 음성을 테스트한 결과이다. female로 옳게 판별했다.

```
filename="남자_23_test1.wav"

infer_gender(filename)[0]
```

✓ 0.1s

```
'male'
```

[그림 9] 남성 음성 테스트 결과1

20대 남성의 음성을 테스트한 결과이다. male로 옳게 판별했다.

이번에는 여성이 남성처럼 보이기 위해 목소리를 의도적으로 낮게 낸 음성 파일을 테스트해봤다.

```
filename="여자_낮게_23_test1.wav"

infer_gender(filename)[0]
```

✓ 0.1s

```
'female'
```

[그림 10] 여성 음성 테스트 결과2

여자라고 잘 판별되었다.

하지만 [그림 6]의 실험을 음역대가 더 낮은 목소리를 가진 여성으로 테스트해보니 잘 구별하지 못하였다.

```
filename="여자_낮게_25_test2.wav"

infer_gender(filename)[0]
```

✓ 0.1s

```
'male'
```

[그림 11] 여성 음성 테스트 결과3

제 5 장 결론 및 향후 과제

머신러닝을 이용한 발화자의 성별 인식을 시스템을 개발하였다. 혼동 행렬의 결과와 학습 시간을 종합해 ANN, Logistic Regression, KNN, SVM-Linear, Random Forest 모델을 연구했는데, 그 중 ANN 모델이 정확도 99%로 구현하고자 하는 시스템 모델로 가장 적합하다는 결론이 나왔다. 선행 연구에서는 약 89%의 정확도를 보였으나, 이 연구를 통해 99%의 정확도를 얻었다. 따라서 본 연구의 목적은 달성하였다고 볼 수 있다.

하지만 의도적으로 성별을 속이려고 한 음성 데이터에 대해서는 알맞은 결과를 내지 못할 때도 있다. 향후 과제로 위와 같은 음성들로 구성된 데이터들을 따로 모아서 학습시켜야 할 필요가 있다. 예를 들어 여성이 의도적으로 낮게 낸 목소리와 남성이 높게 낸 목소리의 데이터를 모아서 어떤 특징이 있는지 분석하고 학습시켜서 제대로 구별할 수 있게 개선해야 할 것이다. 또한 이를 이용해 보이스 피싱 범죄 절감에 기여할 수 있도록 더욱 연구를 해야 할 것이다.

참 고 문 헌

이수지 외 4명. 2019. “딥러닝 모형을 사용한 한국어 음성인식”. The Korean journal of applied statistics 32 (2): 213-227. doi:10.5351.

이영한. 2017. “딥러닝 기반의 음성/오디오 기술”, Speech/Audio Processing based on Deep Learning

Sakshi Dua 외 5명. 2022. “Developing a Speech Recognition System for Recognizing Tonal Speech Signals Using a Convolutional Neural Network”

박형민. 2022. 영상 및 음성 데이터를 이용한 머신러닝. 대한응급의학회 학술대회 초록집, 2022, no.1:157-158

김대진, 김용원, 우성흔. (2019). 음성인식: 서비스 품질 향상을 위한 성별 인식 정확도 개선. 한국품질경영학회 추계학술발표논문집, 2019(0), 253-253.

이후영. (2021). CNN(Convolutional Neural Network) 알고리즘을 활용한 음성신호 중 비음성 구간 탐지 모델 연구. 융합정보논문지, 11(6), 33-39.

감 사 의 글

고등학교를 마치고 대학교에 입학한 지 얼마 안 된 것 같은데 벌써 졸업을 앞두고 있습니다. 4년간의 학부 과정을 거치면서 정말 많은 것을 배웠습니다. 먼저 항상 저를 믿어주시고 응원해주신 가족들에게 감사의 말을 전하고 싶습니다. 본 논문이 완성되기까지 계속 지도해주시고 도움을 주신 남윤영 교수님에게 깊은 감사를 드립니다. 또한, 지난 4년간의 학위 과정에서 항상 열정 있고 헌신적인 지도를 아끼지 않으셨던 천인국 교수님, 이상정 교수님, 이해각 교수님, 하상호 교수님, 홍인식 교수님에게도 감사를 표합니다.

힘든 일이 있어도 위로해주며 앞으로 나아갈 수 있게 도와준 장채림, 손정민, 서영채, 문수정 학생과 많은 조언을 해주신 임주현 선배에게 감사드립니다. 그 외 같이 학부생활을 걸었던 19학번 동기와 선후배님들에게도 감사를 드립니다.

학부 과정을 밟으면서 꿈이 생기기도 하였고, 무언가를 열정적으로 할 때도 있었습니다. 여기서 얻은 값진 경험을 토대로 멋진 사회인이 되도록 노력하겠습니다. 감사합니다.