JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 10, pp. 37-43, October 2021
https://doi.org/10.9708/jksci.2021.26.10.037

Performance Comparison of Korean Dialect Classification Models Based on Acoustic Features

Young Kook Kim*, Myung Ho Kim*

*Student, Dept. of Software, Soongsil University, Seoul, Korea
*Professor, Dept. of Software, Soongsil University, Seoul, Korea

[Abstract]

Using the acoustic features of speech, important social and linguistic information about the speaker can be obtained, and one of the key features is the dialect. A speaker's use of a dialect is a major barrier to interaction with a computer. Dialects can be distinguished at various levels such as phonemes, syllables, words, phrases, and sentences, but it is difficult to distinguish dialects by identifying them one by one. Therefore, in this paper, we propose a lightweight Korean dialect classification model using only MFCC among the features of speech data. We study the optimal method to utilize MFCC features through Korean conversational voice data, and compare the classification performance of five Korean dialects in Gyeonggi/Seoul, Gangwon, Chungcheong, Jeolla, and Gyeongsang in eight machine learning and deep learning classification models. The performance of most classification models was improved by normalizing the MFCC, and the accuracy was improved by 1.07% and F1-score by 2.04% compared to the best performance of the classification model before normalizing the MFCC.

▸ **Key words**: Machine Learning, Deep Learning, MFCC, Dialect Classification, Speech Analysis

[요 약]

말소리의 음향 특징을 이용하여 화자에 대한 중요한 사회, 언어학적 정보를 얻을 수 있는데 그 중 한 가지 핵심 특징은 방언이다. 화자의 방언 사용은 컴퓨터와의 상호작용을 방해하는 주요 요소이다. 방언은 발화의 음소, 음절, 단어, 문장 및 구와 같이 다양한 수준에서 구분할 수 있지만 이를 하나하나 식별하여 방언을 구분하기는 어렵다. 이에 본 논문에서는 음성 데이터의 특성 중 MFCC만 사용하는 경량화된 한국어 방언 분류 모델을 제안한다. 한국인 대화 음성 데이터를 통해 MFCC 특징을 활용하는 최적의 방법을 연구하고, 8가지 머신 러닝 및 딥러닝 분류 모델에서 경기/서울, 강원, 충청, 전라, 경상 5개의 한국어 방언 분류 성능을 비교한다. MFCC를 정규화하는 방법으로 대부분의 분류 모델에서 성능을 향상시켰으며, MFCC를 정규화하기 전 분류 모델의 최고 성능과 비교하여 정확도는 1.07%, F1-score는 2.04% 향상된 성능을 기록하였다.

▸ **주제어**: 머신 러닝, 딥러닝, MFCC, 방언 분류, 성능 비교, 음성 분석

• First Author: Young Kook Kim, Corresponding Author: Myung Ho Kim
*Young Kook Kim (1101978003@soongsil.ac.kr), Dept. of Software, Soongsil University
*Myung Ho Kim (kmh@ssu.ac.kr), Dept. of Software, Soongsil University
• Received: 2021. 09. 15, Revised: 2021. 10. 21, Accepted: 2021. 10. 25.

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

I. Introduction

최근 스마트폰, 인공지능 스피커 및 다양한 가전제품에서 음성 서비스를 제공하기 위한 사람의 말소리를 분석하는 기술이 지속적으로 발전하고 있다. 이런 서비스는 주로 음성 인식을 통해 사용자가 말하는 언어를 컴퓨터가 해석하여 문자 데이터로 처리하는 기술을 이용한다. 이때 사용자가 표준어가 아닌 방언을 사용하는 경우, 컴퓨터가 인간의 언어를 제대로 이해하지 못하여 성능이 저하되는 문제점을 가지고 있다[1, 2].

말소리의 음향적 특성은 화자에 대한 중요한 사회언어학적 및 언어학적 정보를 드러낼 수 있으며, 이러한 특성은 화자의 정체성에 대한 식별을 가능하게 할 수 있다. 말소리(연설)의 한 가지 핵심 특징은 화자의 방언이다. 방언은 하나의 언어가 지역적으로 달리 변화하여 어느 정도 상이한 언어체계를 이룰 때 각각의 지역에 속한 그룹에서 관찰되는 말하기 패턴을 의미한다. 이런 변화는 주로 지리적 위치에서 오지만 사회적 계층, 문화적 배경, 교육 등과 같은 여러 주변 요인으로 인해 발생하기도 한다. 방언은 표준어와 동일한 국가의 언어지만 사람에게도 이해하는 정도와 받아들이는 분위기 등에 차이가 있어 원활한 커뮤니케이션이 방해되기도 한다.

방언 분류 시스템의 구현은 인간과 컴퓨터의 상호작용을 보완하여 콜센터의 방언 사용 고객에 따른 대응, 화자의 출생 식별, 원격 의료 시스템의 활용 등 다양한 분야에 활용될 수 있다. 따라서, 본 논문에서는 한국어의 5개 방언에 대한 음성 특징 기반의 분류 시스템을 제안한다. 음성 및 화자 인식 분야에서 사용하는 오디오의 특징 중 MFCC(Mel-Frequency Cepstral Coefficient)를 사용하여 특징을 벡터화하고, 이를 MinMax Scaling의 방법으로 정규화하여 8가지 분류 모델 SVM(Support Vector Machine), RF(Random Forest), DNN(Deep Neural Network), RNN(Recurrent Neural Network), LSTM (Long-Short Term Memory), Bi-LSTM(Bidir ectional LSTM), GRU(Gated Recurrent Unit), 1D CNN(1D Convolution Network)에 대해 Accuracy, Precision, Recall, F1-score 4가지 평가 지표로 성능을 비교하고 최적의 분류 모델을 제안한다.

본 논문은 다음과 같이 구성된다. 2장에서는 방언 분류에 대한 기존 방법을 설명하고, 3장에서 특징 처리 방법 및 모델 구성을 설명한다. 4장에서는 실험 과정 및 결과를 요약하고, 5장에서 결론 및 미래 연구를 기술한다.

II. Preliminaries

1. Related works

화자 인식은 음성으로부터 화자의 정보를 찾아내는 기술을 의미한다. 크게 화자 식별, 화자 검증, 화자 검출 3가지 종류로 나뉘는데 화자 식별은 주어진 음성으로 다수의 후보 중 한 명의 화자를 찾는 기술이고, 화자 검증은 주어진 음성이 가지고 있는 음성과 동일한지 검증하는 기술이고, 화자 검출은 특정한 화자를 찾아내는 기술이다. 화자 인식을 위해 음성 인식을 통해 얻은 텍스트를 특징으로 사용하기도 하지만 MFCC 음성 특징 또한 주로 사용한다. MFCC는 음성 인식에서 널리 사용되는 음성의 특성으로 음성 데이터 전체를 대상으로 하지 않고, 일정 구간으로 나누어 이 구간에 대한 스펙트럼을 분석하여 추출한 특징이다.

방언을 구분하는 단서는 음소, 음절, 단어, 문장 등 다양한 수준에 존재할 수 있다[3, 4]. 기존 논문에 따르면 사람과 컴퓨터가 방언을 구별하는 다양한 접근 방식이 제안되었다. 일반적으로 방언은 음성 정보 및 음소 접근 방식을 사용하여 분류한다. 최근에는 음성 특징을 단순히 표현하기 위해 GMM(Gaussian Mixture Model)을 활용하여 I-vector를 추출하여 PLDA(Probabilistic Linear Discriminant Analysis, 확률적 선형 판별 분석)을 사용하여 방언 및 화자를 분류한다[5-7]. 이후 I-vector를 DNN에서 추출한 임베딩으로 대체하는 시스템이 등장했다[8]. 단어 수준에서 음성 특징을 활용하는 등 다양한 방법으로 음성 특징을 활용하고 주로 GMM, SVM, RF와 같은 기계 학습의 모델로 분류한다[9, 10].

2. Machine Learning - based approach

음성에서 사람의 성별 및 나이 정보를 분류하기 위한 머신 러닝 기반의 대표적인 연구 모델로 GMM과 SVM이 있다. 초기의 연구는 GMM과 SVM을 단일 모델로 사용하여 분류했지만 GMM을 사용해 슈퍼 벡터를 생성하고 이를 SVM으로 모델링하는 등 둘 이상의 시스템을 융합하여 분류 정확도를 높이기 위해 다양한 방법을 활용한다[11-14].

2.1 SVM(Support Vector Machine)

SVM은 기계 학습의 분야 중 하나로 결정 경계(Decision Boundary), 즉 데이터를 분류하는 선을 정의하는 모델이다. 분류되지 않은 새로운 데이터가 입력되면 결정 경계를 기준으로 어느 쪽에 속하는지 예측한다.

2.2 RF(Random Forest)

랜덤 포레스트는 기계 학습 분야 중 하나로 분류, 회귀 등에 사용되는 앙상블 학습 방법 모델이다. 여러 개의 결정 트리(decision tree)를 만들고 새로운 데이터를 각 트리에 통과시키며, 각 트리가 분류한 결과에서 투표를 시행하여 최종 분류 결과를 결정한다. 많은 수의 결정 트리를 생성함으로 일부 결정 트리가 오버피팅되어도 최종 분류에 영향을 미치지 않는다.

3. Deep Learning - based approach

최근 화자 인식 및 검증 등의 분야에서 딥러닝 기반의 기법들이 적용되면서 단순하게는 신경망의 구조에 따른 분류기에 관한 연구에서 MFCC와 같은 기존의 음성 특징을 입력으로 신경망을 거쳐 새로운 특징 벡터를 임베딩하고 생성하는 음성 특징 추출 연구 및 여러 음성 특징을 동시에 활용하는 복잡한 시스템 등 다양한 연구가 이루어지고 있다[15, 16, 17, 18, 19].

3.1 DNN(Deep Neural Network)

신경망(Neural Network)은 두뇌의 신경세포, 즉 뉴런이 연결된 형태를 모방한 모델이다. 생물학적인 뉴런을 수학적으로 모델링한 것으로 여러 개의 뉴런과 같이 각 노드들은 입력층(Input Layer), 은닉층(Hidden Layer), 출력층(Output Layer)으로 구분되며 각층의 노드들은 상호작용을 통해 입력값을 받고 일정 수준이 넘어서면 활성화되어 출력값을 내보낸다.

3.2 RNN(Recurrent Neural Network)

순환신경망은 입력과 출력을 시퀀스 단위로 처리하는 모델이다. 딥러닝에 있어 가장 기본적인 시퀀스 모델로 은닉층의 노드에서 활성화된 결과값을 출력층으로 보내면서 동시에 다시 은닉층 노드의 다음 계산의 입력으로 보내는 특징을 가지고 있다.

3.3 LSTM(Long-Short Term Memory)

LSTM은 RNN의 한 종류로 기본적인 RNN 구조보다 더 긴 시퀀스를 처리하는데 용이하다. 순환신경망은 시퀀스의 길이가 길어질 때 이전의 결과가 계속 축적되어 먼 거리에 있는 상태가 현재의 결과에 미치는 영향이 미미해지는 단점이 있다. LSTM은 이를 해결한 모델로 입력, 출력, 삭제 게이트라는 3개의 게이트를 통해 과거의 정보 중 어떤 정보를 버릴지 결정하고, 새로운 정보를 업데이트하고, 마지막으로 무엇을 출력할지 정한다. 일반적으로 LSTM은 시퀀스를 단방향으로 처리하는데 양방향으로 처리하는 방법을 Bi-LSTM이라고 한다. Bi-LSTM은 시퀀스에서 이전 데이터와의 관계뿐만 아니라 이후 데이터와의 관계까지 학습하기 때문에 시퀀스 전체를 파악하기에 용이하다.

3.4 GRU(Gated Recurrent Unit)

GRU는 LSTM의 장점은 유지하면서 구조를 간단화한 모델이다. LSTM에 존재하는 출력, 입력, 삭제 게이트를 업데이트, 리셋 게이트 두 가지 게이트로 줄여 LSTM보다 학습 속도를 빠르게 하고 비슷한 성능을 보인다.

3.5 1D CNN(1D Convolution Neural Network)

CNN은 인간의 시신경을 모방하여 만든 딥러닝 구조 중 하나이다. 일반적으로 2차원에서 컨볼루션 연산을 이용하여 이미지의 공간적인 정보를 추출하여 이를 기반으로 이미지 분류 등 이미지 처리에서 좋은 성능을 보인다. 1D CNN은 Recurrent한 유닛을 사용하지 않고 컨볼루션 연산을 이용하여 시퀀스 데이터를 처리하는 모델이다.

III. The Proposed Scheme

1. System Architecture

Fig. 1은 본 논문에서 제안하는 방언 분류 시스템의 구조를 나타낸다. 음성 데이터에서 MFCC 특징을 추출하고 MinMax Scaling 방법으로 정규화한다.

MFCC는 인간의 청각 시스템과 유사하게 저주파수 대역에서 민감하고, 고주파수 대역에서 상대적으로 둔감한 특징을 표현하기 위한 특징으로, DFT(Discreate Fourier Transform, 이산 푸리에 변환) 기반 스펙트럼에서 Mel-scale 필터를 사용하여 얻은 cepstral 계수이다. Fig. 2는 MFCC를 추출하는 과정을 나타낸다. 음성 데이터를 짧은 구간으로 나누고 나누어진 구간에 DFT를 적용하여 주파수 정보를 추출한다. 이때 나누어진 구간을 프레임(frame), 각 프레임에 DFT를 적용한 결과인 주파수 정보

Fig. 1. System Architecture

를 스펙트럼이라고 한다. 다시 각 스펙트럼에 Mel-scale 필터를 사용하고 로그를 취한 다음 DCT(Discreate Cosine Transform, 이산 코사인 변환)를 적용하여 MFCC를 계산한다.

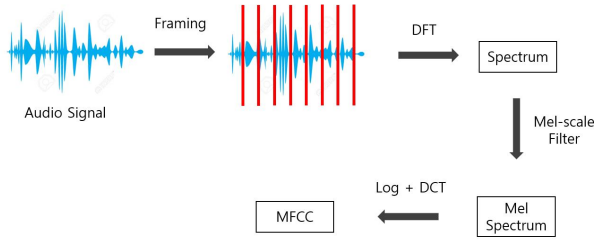

Fig. 2. MFCC Generation Process

MinMax Scaling은 벡터로 표현된 MFCC의 각 차수에 따른 특징값을 [0, 1] 사이의 값으로 변환한다. 식 (1)은 데이터셋 X에서 MinMax Scaling 방법을 의미한다[20].

$$x_i' = \frac{x_i - \min(X)}{\max(X) - \min(X)},\ where\ x_i \in X \qquad (1)$$

예를 들어 세 벡터 x=(1, 4, 3), y=(4, 2, 1), z=(3, 6, 2)가 있을 때 세 벡터의 각 성분끼리 비교하여 최대값과 최소값을 찾아 식(1)에 적용하여 변환한다. 세 벡터의 성분 중 최대값만을 가지는 벡터를 max, 최소값만을 가지는 벡터를 min이라 하면 max=(4, 6, 3), min=(1, 2, 1)이다. 세 벡터 x, y, z에 대해 x를 MinMax Scaling 방법으로 정규화하면 $x' = \left(\frac{1-1}{4-1}, \frac{4-2}{6-2}, \frac{3-1}{3-1}\right) = (0, 0.5, 1)$이다.

위의 방법으로 정규화된 벡터 형태의 특징을 입력으로 하는 분류 모델을 구성하고 이를 통해 5가지 한국어 방언을 분류한다.

2. Classification Model

2.1 Machine Learning Medel

SVM의 모델은 데이터가 어떤 결정 경계를 기준으로 분류되는지에 따라 성능의 차이가 결정된다. 실험에 사용할 SVM 모델은 결정 결계의 마진을 조절하기 위한 파라미터인 c의 값을 10으로 하고, RBF(Radial Bias Function) 커널을 사용한다.

RF 모델은 의사결정나무 모형의 각 마디에서 변수를 선택하는 데 임의성이 도입되기 때문에 결정 트리의 수가 증가할수록 예측 오차가 줄어들며, 오버피팅되지 않는 장점이 있다. 하지만 그만큼 메모리와 훈련 시간이 증가하기 때문에 실험에서는 150개의 결정 트리를 사용한다.

2.2 Deep Learning Model

실험에 사용하는 모든 딥러닝 모델은 같은 분류층을 사용하며 활성화 함수로 softmax를 사용하고, L2 정규화 방법을 통해 오버피팅을 방지하면서 5가지 방언을 분류한다.

DNN 모델의 구성은 5개의 은닉층과 분류층으로 구성하며 각 은닉층별 크기는 각각 256, 256, 128, 128, 64로 구성하며 각 은닉층은 0.3의 확률로 Dropout 과정을 거쳐 오버피팅을 방지한다.

RNN, LSTM, Bi-LSTM, GRU 등 RNN 기반의 모델은 128개의 특징을 가진 네트워크 레이어를 사용하고, Flatten 작업 이후 각각 128, 64, 32 크기의 은닉층으로 구성되며 각 레이어는 0.3의 Dropout을 적용한다.

1D CNN 모델은 128개의 특징을 가진 1D CNN 레이어를 사용하고, 이후 모델의 구성은 RNN 기반의 모델 구성과 같게 한다.

3. Performance Evaluation

학습된 모델은 분류 성능 평가에 일반적으로 사용하는 지표인 Accuracy, Precision, Recall, F1-score를 측정한다. Accuracy는 전체 데이터 중 모델이 바르게 분류한 비율을 의미하고, precision은 모델이 참이라 분류한 것 중 실제값이 참인 데이터의 비율, Recall은 실제값이 참인 데이터 중 모델이 참이라고 분류한 비율을 의미한다. F1-score는 데이터 간 불균형에 대해 보정된 성능을 보여주는 지표로 Precision과 Recall의 조화평균을 의미한다. 실제 데이터 라벨링과 모델의 분류 결과에 따라 Table 1과 같이 표현할 때 각 지표는 다음 식 (2)~(5)와 같이 표현한다[21].

Table 1. Table of Confusion

		Actual	
		True	False
Predicted	True	True Positive (TP)	False Positive (FP)
	False	False Negative (FN)	True Negative (TN)

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (2)$$

$$\mathrm{Precision} = \frac{TP}{TP + FP} \qquad (3)$$

$$Recall = \frac{TP}{TP + FN} \qquad (4)$$

$$F1 - score = 2 \times \frac{\mathrm{Precision} \times Recall}{\mathrm{Precision} + Recall} \qquad (5)$$

본 논문에서는 5가지 클래스에 대한 분류 작업이므로 각 클래스에 해당하는 데이터의 개수에 가중치를 주어 평균을 구하는 방법으로 4가지 지표를 측정한다.

IV. Experiments

1. Dataset

본 논문에서는 AI Hub에서 제공하는 한국인 대화 음성 데이터 세트를 사용하여 모델 학습과 테스트를 수행한다 [22]. 한국인 대화 음성 데이터 세트는 한국인의 일상 대화를 인식하고 음성을 문자로 실시간 변환하는 인공지능 기술 개발을 위해 춘천 MBC와 EBS의 방송콘텐츠에서 음원을 추출하고 화자의 성별, 나이, 방언 등의 정보를 라벨링한 데이터 세트이다. 실험을 위해 전체 데이터 세트에서 제주 지역의 방언을 제외하고 서울/경기, 강원, 충청, 경상, 전라로 구분된 5개의 방언 음성 데이터 약 1만 1천 개를 임의로 추출하여 7:3의 비율로 훈련 데이터와 테스트 데이터로 나누어 사용하였다. 빠르고 간단한 모델 구성을 위해 음성 데이터에서 MFCC 특징만을 추출하여 방언 분류를 위한 입력 데이터로 사용하여 실험한다.

2. Comparison of MFCC Extration Method

MFCC 특징 추출 방법에 따른 성능을 측정한다. 음성 데이터에서 4초간 추출한 13차 MFCC 특징값 전체와 평균값을 사용하여 LSTM 모델에서 방언 분류 성능을 비교한다. 4초간 MFCC를 이용하여 훈련한 경우 정확도는 45%이고, 평균 MFCC를 이용하여 훈련한 경우 59%의 정확도를 기록하였다. Fig. 3은 방언별 MFCC 특징에 따른 Precision과 Recall 값을 나타낸 그래프이다. 두 가지 특징을 사용한 분류 모델 모두 충청과 전라 지역의 방언 분류에서 성능이 부족하지만, 평균 MFCC를 이용한 경우 정확도 및 각 방언에 대한 대부분의 Precision과 Recall에서 훨씬 좋은 성능을 기록하여 본 논문에서는 평균 MFCC의 값을 입력으로 하여 실험을 진행하였다.

3. Classification Model Comparison

Table 2는 평균 MFCC를 이용하여 분류 모델을 학습하고 성능을 비교한 결과이다. 8가지 분류 모델 중 기계 학습의 방법 중 RF가 정확도 63.93%, F1-score 61.39%를 기록하고, Precision과 Recall 지표 모두 가장 높은 성능을 기록하였다. 다른 단일 모델들과 달리 RF가 여러 결정트리의 앙상블 형태의 모델이기 때문에 가장 성능이 좋은

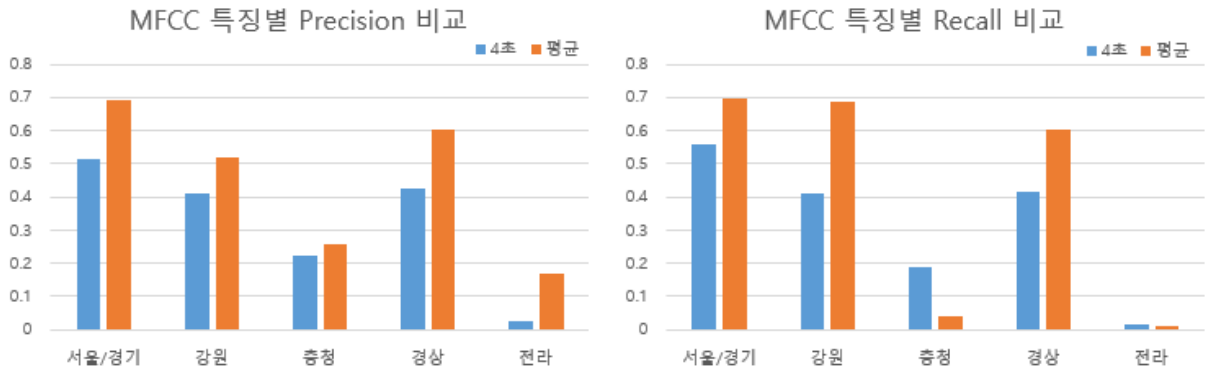

Fig. 3. Precision, Recall Graph according to MFCC Extration Method

Table 2. Table of Classification Model Performance

	Accuracy	Precision	Recall	F1-score
SVM	60.08	55.25	60.08	56.80
RF	**63.93**	**63.17**	**63.93**	**61.39**
DNN	61.21	60.42	61.21	59.09
RNN	59.48	56.46	59.48	56.50
LSTM	59.63	56.84	59.63	57.10
Bi-LSTM	62.94	57.11	62.94	59.65
GRU	61.72	56.29	61.72	58.42
1D-CNN	61.51	60.07	61.51	58.80

Table 3. Table of Classification Model Performance with Scaled MFCC

	Accuracy	Precision	Recall	F1-score
SVM	64.02	62.69	64.02	61.73
RF	64.50	63.97	64.50	62.02
DNN	**65.00**	**64.49**	**65.00**	**63.43**
RNN	63.66	58.05	63.66	60.29
LSTM	64.56	63.39	64.56	61.42
Bi-LSTM	61.84	58.47	61.84	58.47
GRU	62.62	56.99	62.62	59.28
1D-CNN	64.23	58.80	64.23	60.84

것으로 판단된다. 딥러닝 학습 방법 중에서는 Bi-LSTM, 1D CNN과 같이 시퀀스의 양방향으로 훈련되는 모델들이 시퀀스의 단방향 정보로 훈련되는 RNN, LSTM 모델보다 준수한 성능을 보인다.

4. Classification Model Comparison With Scaled MFCC

Table 3은 평균 MFCC에 MinMax Scaling 방법을 통해 정규화하여 분류 모델을 학습하여 측정한 성능을 비교한 결과이다. Bi-LSTM을 제외한 나머지 7가지 분류 모델 모두에서 향상된 성능을 기록하였다. 그중 딥러닝 방법의 DNN이 65% 정확도, 63.43%의 F1-score로 가장 높은 성능을 기록하였다. 4.3의 실험과 비교하여 시퀀스의 방향성에 따라 훈련되는 모델들보다 전체 데이터에서 특징을 추출하여 훈련되는 SVM, DNN 모델의 성능이 크게 향상되었다. MFCC 특징의 차수가 13으로 작고, 평균값을 사용하므로 상대적으로 매우 작은 크기의 입력값이 정규화를 통해 DNN 모델이 훈련하기에 최적화된 것으로 보인다. Fig. 4는 DNN으로 측정한 방언별 Precision, Recall, F1-score를 나타내는 그래프이다. 이전의 실험 결과인 Fig. 2와 비교하여 상대적으로 부족한 성능을 기록한 충청, 전라 지역 방언에서도 향상된 성능을 기록한다.

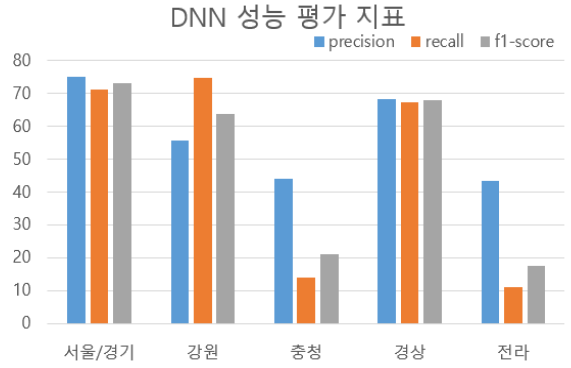

Fig. 4. DNN Evaluation Graph with Scaled MFCC

V. Conclusions

음성 인식 및 화자 인식에서 사용자의 방언 사용은 성능을 떨어뜨리는 주요 원인이다. 방언 분류 시스템을 구현하여 음성 기반 시스템의 성능을 향상시키기 위해 본 논문에서는 한국어 음성 데이터에서 MFCC 특징을 추출하여 입력으로 하여 5가지 방언을 분류하는 모델을 제안한다. MFCC 특징을 효율적으로 사용하기 위해 MFCC의 특징을 시간 단위가 아닌 평균값을 사용할 것을 제안한다. 나아가 MinMax Scaling의 방법으로 MFCC 특징을 정규화하여 입력으로 사용하는 DNN 모델에서 65%의 정확도, 63.43%의 F1-score의 성능을 보였다. MFCC 특징을 정규화하지 않고 훈련했을 때 가장 좋은 성능을 보인 RF 모델보다 정확도는 1.07%, F1-score는 2.04% 향상시켰다.

본 논문은 한국어 음성 데이터에서 4초 동안의 MFCC 특징의 평균값을 정규화하여 추출하고, 이 특징을 입력으로 하는 방언을 분류하는 최적의 모델을 제안하며 다른 모델들과의 성능을 비교한다. 제안하는 모델은 기존 연구들과 달리 MFCC 특징만 입력으로 사용하여 작은 수의 파라미터를 사용하는 경량 모델이다. 성능 대비 연산량을 줄이고 처리 시간은 빠른 모델이다. 향후 한국어 음성 데이터의 개수를 늘리고 MFCC 특징과 함께 다른 오디오 특징을 활용하는 연구와 특정 지역 방언 분류 정확도가 상대적으로 부족한 점을 개선하는 연구가 필요하다.

REFERENCES

[1] Park Jeon-gyu, “Deep Learning-based Speech Recognition Technology”, http://www.itdaily.kr/news/articleView.html?idxno=76405

[2] S. S. Jo and Y. G. Kim, "AI (Artificial Intelligence) Voice Assistant Evolving to Platform", IITP, pp. 1-25, Feb. 2017

[3] Rongqing Huang and John HL Hansen, “Dialect/accent classification via boosted word modeling”, In IEEE International

Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings (ICASSP'05), volume 1, pages I-585. IEEE, 2005

[4] Thomas Purnell, William Idsardi, and John Baugh. "Perceptual and phonetic experiments on american english dialect identification", Journal of language and social psychology, 18(1):10-30, 1999.

[5] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P, "Front-end factor analysis for speaker verification", IEEE Trans. Audio Speech Lang. Process. 19, pp. 788-798, August 2010, DOI: 10.1109/TASL.2010.2064307

[6] Dehak, N., Torres-Carrasquillo, P., Reynolds, D., and Dehak, R, "Language recognition via ivectors and dimensionality reduction" in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH (San Francisco, CA), pp. 857-860, August 2011.

[7] Song, Y., Jiang, B., Bao, Y., Wei, S., and Dai, L.-R, "I-vector representation based on bottleneck features for language identification", Electron. Lett. 49, pp. 1569-1570, 2013, DOI: 10.1049/el.2013.1721

[8] Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., and Khudanpur, S, "Spoken language recognition using x-vectors," in Proceedings of Odyssey 2018 The Speaker and Language Recognition Workshop, pp. 105-111, 2018, DOI: 10.21437/Odyssey.2018-15

[9] C. Themistocleous, "Dialect Classification From a Single Sonorant Sound Using Deep Neural Networks" frontiers om Communication, November 2019. DOI: 10.3389/fcomm.2019.00064

[10] Nagaratna B. Chittaragi; Shashidhar G. Koolagudi, "Acoustic features based word level dialect classification using SVM and ensemble methods" IEEE Trans, In 2017 Tenth International Conference on Contemporary Computing (IC3), pp. 1-6, August 2017, DOI: 10.1109/IC3.2017.8284315.

[11] Li, Ming, Chi-Sang Jung, and Kyu J. Han. "Combining five acoustic level modeling methods for automatic speaker age and gender recognition." Eleventh Annual Conference of the International Speech Communication Association. pp. 2526-2829, 2010.

[12] Reynolds, Douglas A., Thomas F. Quatieri, and Robert B. Dunn. "Speaker verification using adapted Gaussian mixture models." Digital signal processing 10.1-3, pp. 19-41, 2000.

[13] Li, Ming, et al. "Spoken language identification using score vector modeling and support vector machine." Eighth Annual Conference of the International Speech Communication Association. pp. 350-353, 2007.

[14] Stolcke, Andreas, et al. "Speaker recognition with session variability normalization based on MLLR adaptation transforms." IEEE Transactions on Audio, Speech, and Language Processing 15.7, pp. 1987-1998, 2007.

[15] Qawaqneh, Zakariya, Arafat Abu Mallouh, and Buket D. Barkana. "Deep neural network framework and transformed MFCCs for speaker's age and gender classification." Knowledge-Based Systems 115, pp. 5-14, 2017.

[16] Mallouh, Arafat Abu, Zakariya Qawaqneh, and Buket D. Barkana. "New transformed features generated by deep bottleneck extractor and a GMM-UBM classifier for speaker age and gender classification." Neural Computing and Applications 30.8, pp. 2581-2593, 2018.

[17] Ghahremani, P., Nidadavolu, P. S., Chen, N., Villalba, J., Povey, D., Khudanpur, S., & Dehak, N. "End-to-end Deep Neural Network Age Estimation." In INTERSPEECH, pp. 277-281, December 2018.

[18] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. "X-vectors: Robust dnn embeddings for speaker recognition.", In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 5329-5333, April 2018.

[19] Hourri, Soufiane, and Jamal Kharroubi. "A deep learning approach for speaker recognition." International Journal of Speech Technology 23.1, pp. 123-131, 2020.

[20] S. Gopal Krishna Patro, Kishore Kumar Sahu, "Normalization: A Preprocessing Stage" arXiv preprint arXiv:1503.06462 (2015).

[21] Goutte, Cyril, and Eric Gaussier. "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation." European conference on information retrieval. Springer, Berlin, Heidelberg, pp. 345-359, 2005, DOI: 10.1007/978-3-540-31865-1_25.

[22] AI Hub, Korean Conversation Voice, https://aihub.or.kr/aidata/7968.

Authors

Young Kook Kim received the B.S. degree in Mathematics from Soongsil University (SSU), Korea, in 2017. He is currently a M.S. Student in the Department of Software, Soongsil University (SSU).

His research interests are Deep Learning, Machine Learning, Recommendation System, NLP and Big data analysis.

Myung Ho Kim received the B.S. in Department of Computer Science and Engineering from Soongsil University, Korea, in 1989. M.S. and Ph.D. degreesin Department of Computer Engineering from Postech University, Korea, in 1991 and 1995, respectively. He is currently a professor in the Dept. of Software, Soongsil University. He is interested in Machine Learning, Deep Learning and Block chain.