Multimodal dialect speech recognition

Yeonjae Kim
Department of Software,
Sungkynkwan University
Suwon, Korea
semote2@gmail.com

Seyeon Park
Department of Software,
Sungkynkwan University
Suwon, Korea
sytrocl00@gmail.com

Sechang Lim
Department of Software,
Sungkynkwan University
Suwon, Korea
rhkrqnwk98@gmail.com

Myeongwon Jung
Department of Math/Software,
Sungkynkwan University
Suwon, Korea
jmw1790@gmail.com

Sangmin Han
Department of Math/Software,
Sungkynkwan University
Suwon, Korea
smhan213@gmail.com

ABSTRACT

81% of people in Korea live outside of Seoul, and most of them use dialect. Many people can converse with each other even though they use dialects, but machines can't. For dialect speech recognition, it is necessary to solve the problem of classifying which dialect the corresponding speech is. In this study, the multimodal method is used to improve classification accuracy by utilizing various characteristics of speech. We transformed .wav file to MFCC, Spectrogram, and Chromagram. ResNet18, ResNet18 version 2, and LSTM were used to measure the accuracy of each data. MFCC and Spectrogram showed the best test accuracy in LSTM with 0.97 and 0.93, respectively, and Chromagram showed the best performance with 0.83 in ResNet18. As a result of creating a multimodal model by combining each network, the final accuracy was 0.98 and the best performance was achieved.

Keywords

Dialect Classification, MFCC, Spectrogram, Chromagram, Deep Learning, ResNet18, LSTM, Multimodal

1. INTRODUCTION

81% of people in Korea live outside of Seoul, and most of them use dialect.[1] In a situation where the population density in the metropolitan area is getting higher and higher, as voice recognition technology is developed mainly in standard languages, dialects are gradually neglected. Since dialects are unique to Korea and represent the culture of each region, it is necessary to continuously preserve them.

Many people can converse with each other even though they use dialects, but machines can't. For example, a virtual assistant like Siri hard to recognize dialects in comparison with standard language.

For dialect speech recognition, it is necessary to solve the problem of classifying which dialect the corresponding speech is. A lot of research has been done on the problem of classifying English in countries with different English pronunciations or classifying Arab dialects. On the other hand, there have not been many studies on the classification of Korean dialects by voice recognition.

There are several methods of vectorizing a voice file for voice recognition. Among them, the representative pre-processing method is to make MFCC. It is created by processing the .wav file to be the same as that perceived by humans. The maximum accuracy in the Korean dialect classification dictionary study, which was conducted a dialect classification study using only MFCC, was 65% [1].

In this study, the multimodal method is used to improve classification accuracy by utilizing various characteristics of speech based on the fact that dialects are recognized through various characteristics of speech, such as changes in pronunciation strength and prosody.

ResNet18 and LSTM are used as neural networks, and MFCC, Spectrogram, and Chromagram are used. We measure the dialect classification performance of a single neural network model using single data and compare and analyze the performance of a multimodal model that trains three data simultaneously using ResNet18 and a multimodal model that combines a neural network with the best accuracy for each data.

2. Related works

2.1 Korean Dialect Classification

There was a study that compared and analyzed the accuracy of MFCC using several machine learning models for the classification of Korean dialect speech data [2]. In this study, the audio file was cut in 4-second increments, converted to MFCC, and then normalized by applying min-max scaling. The classification performance is measured by putting them as inputs to SVM, Random Forest (RF), DNN, RNN, LSTM, Bi-LSTM, GRU, and 1D CNN, respectively. At this time, for the normalized MFCC, the accuracy of the RNN was the lowest at 61.84%, and the accuracy of the DNN was the highest at 65%. In this study, in addition to MFCC, deep learning is applied to Chromagram and Spectrogram to compare and analyze the accuracy.

2.2 Foreign Dialect Classification

Difficulty in speech recognition due to the diversity of dialects is not just a problem for Korean. There are ten major dialects in China [3]. The differences between the dialects are quite large, and it is

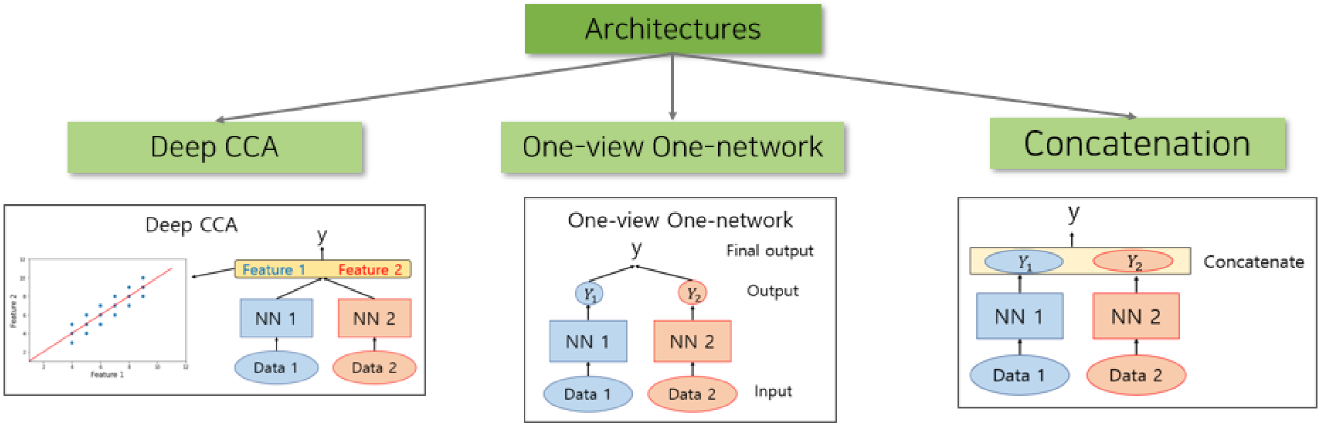


Figure 1. Multimodal Architectures

difficult to distinguish the dialects due to the lack of commentary resources. Therefore, research on automatic speech recognition capable of distinguishing dialects even with a small amount of resources is being conducted. In [2], a method for distinguishing Chinese dialects through transfer learning and data augmentation was studied. They classified speech data with insufficient annotation resources by using a GAN (Generative Adversarial Networks)-based data augmentation model in a ResNet model that learned speech with relatively rich annotation resources.

In this study, since sufficient data is used, the multimodal model is used to improve accuracy by simultaneously learning voice data preprocessed in various ways instead of increasing the model performance by generating data using GAN. In this case, we will focus on using ResNet, which has been used in previous studies.

2.3 Features of Speech Audio for Accent Recognition

This work extracts 5 features as MFCC, Spectrogram, Chromagram, Spectral Centroid, and Spectral Roll-off from speech data. [4] Based on each feature, they inspect accents and classify dialects. According to the study results, except for the Spectral Centroid, Test Loss was all-around 1%. Based on this research, we chose MFCC, Spectrogram and Chromagram to use in our study.

3. Background

3.1 Multimodal

Multimodal learning is a method that is learning different feature dimension data with several model architectures simultaneously as Figure 1. The natural progression of deep-learning research points to problems involving larger and more complex multimodal data [5]. Such multimodal data sets consist of data from different sensors observing common phenomena, and the goal is to use the data in a complementary manner toward learning a complex task. There are some architectures in multimodal. First, Deep CCA uses two steps. In the first step, it extracts the same dimension output vectors from data 1 and data 2 each. And use loss function for correlation between two output vectors. After training, we can get well correlated two output vectors. In the second step, using these vectors for inputs of FC layers to get the final output y. Second, One-view One-network is that we get final output y_i from each network_i and get real final output y by a weighted sum of y_i. Third, Concatenation is when we get the same dimension output vectors and concatenate these. And then train concatenated vectors on FC layers to predict a final output. Since multimodal is heavy due to using several data and several networks simultaneously and Colab performance was not very good, we need a light architecture. So, we choose the third architecture for our model.

3.2 Transform

In this section, we exploit some transform methods for audio data.

3.2.1 Spectrogram

Sound waves are expressed as changes in amplitude over time. But Sound waves cannot identify frequencies, which is one of the important factors in analyzing sound. Therefore, the Spectrogram has been devised. The Spectrogram is expressed as a change in amplitude along the time axis and the frequency axis. It can be obtained through Short-time Fourier Transform. Simply, Short-time Fourier Transform is a technique that obtains frequency information. When it is applied to voice data, the output value is expressed in complex numbers, so take an absolute value on it to extract magnitude. Then take a log on it to convert magnitude to decibels [6].

3.2.2 MFCC

MFCC is a technique of extracting the characteristics of sound, not the entire input sound, but by analyzing the spectrum for this section by dividing it by a short time equation. The human auditory organ is more sensitive in the low-frequency band than in the high-frequency band. Mel Scale is a representation of the relationship between physical frequency and actual human perceived frequency reflecting these characteristics of a person. Mel Spectrum was derived by applying Filter Bank based on Mel Scale to Spectrum. In this way, MFCC can be extracted by performing a Cepstral analysis on the Mel Spectrum finally obtained by reflecting the characteristics of the human hearing organ [7].

3.2.3 Chromagram

Chromagram is one of the conversion methods that convey a spectrum into an octave of 12 scales. The original audio data is represented in a continuous format, but Chromagram converts it into a discrete format [8].

4. Dataset

The dataset used in our project was brought from AIHub [9]. It consists of dialect voice data from five regions: Gangwon, Jeolla, Gyeongsang, Chungcheong, and Jeju. We explain the statistics and preprocessing method for our dataset below.

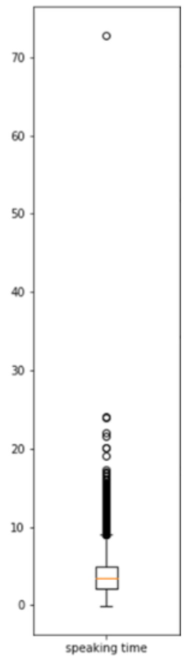


Figure 2. Speaking time Box plot

4.1 Statistics

We were able to get the metadata of our dialect voice data, including gender, birthplace, residential area, educational background, and age about the dataset. To understand the characteristics of the dataset, we extract some statistical information from our dataset.

First of all, we analyzed gender distribution. Interestingly, the number of female speakers was four times bigger than that of males. This distribution was observed equally for each region as well as for the whole.

Second, we analyzed age distribution. We divide age into six age groups: 10-19, 20-29, 30-39, 40-49, 50-59, 60-. Looking at the overall regional statistics, most of them were in their 20s. In addition, the 20s dominated other age range in each region's age distribution.

Finally, we analyzed the length of utterance distribution. It was found that the minimum value of the utterance length was 0, the maximum value was 72.78, the average value was 3.43, and the median was 3.43. Figure 2 shows the distribution of the utterance lengths of our dataset.

4.2 Preprocessing

We use the Librosa library to preprocess data [10]. It is a kind of audio preprocessing library that provides various transformation functions such as MFCC, Chromagram, and Spectrogram. Before converting our original audio data to preprocessed format, we first split, cropped, and padded our data. In the beginning, we split our data into units of length that were too short. It cannot contain the different characteristics of each dialect speaker. So, we split our data into 4-second to 6-second lengths based on the distribution of utterance length computed in the section above. Next, we cropped and padded the split into 5-second lengths to equalize the length of all data into the same length. We also normalized our audio data to fit the data range from -1 to 1. Finally, we transformed our preprocessed data into three formats: MFCC, Chromagram, and Spectrogram using the librosa library. This is shown in Figure 3. The total number of the preprocessed datasets (MFCC, Chromagram, Spectrogram) is 12793, and by region, they are Jeolla 2717, Chungcheong 2946, Gangwon 2310, Gyeongsang 2476, and Jeju 2344. This is shown in Table 1.

Table 1. Size of Train and Test data set

	Jeolla	**Chung Cheong**	**Gang won**	**Gyeong sang**	**Jeju**
Train	2173	2356	1848	1980	1875
Test	544	590	462	496	469

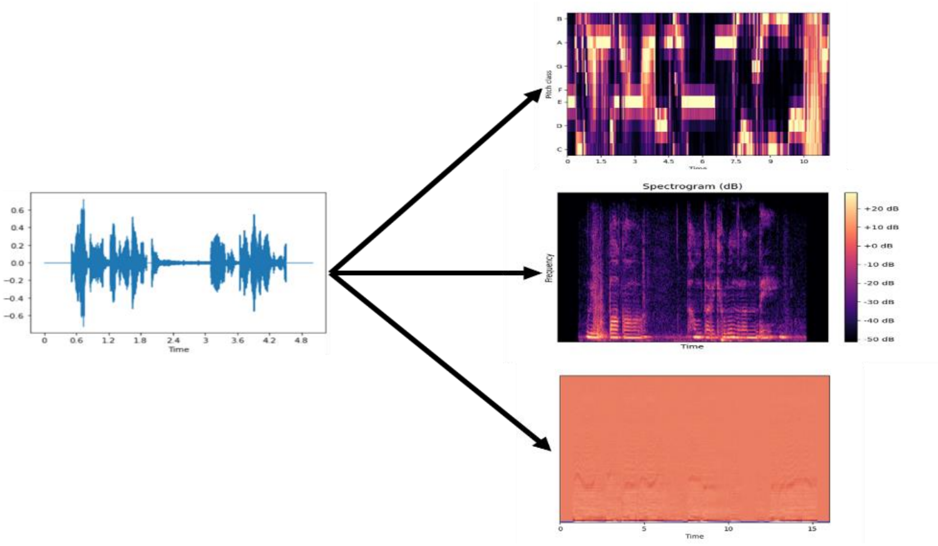


Figure 3. Preprocessing wave file

5. Model Architecture

5.1 ResNet18

ResNet18 followed the existing ResNet18 model [11]. A block layer consisting of a convolutional layer, batch normalization, ReLU, convolutional layer, and batch normalization was stacked and the residual value was added every time it passed through the block layer. 3 block layers that pass 64 3x3 filters, 1 down-sampling layer, 3 block layers that pass 128 3x3 filters, 1 down-sampling layer, and 256-dimensional vectors are extracted using global average pooling layer. Finally, through the FC layer, a 256-dimensional output vector is extracted when used for multimodal, and a 5-dimensional output vector is extracted when used alone. Adam was used as the optimization function and cross-entropy was used as the loss function.

5.2 ResNet18 version 2

ResNet18 version 2 is structurally the same as ResNet18, and only the filter size is configured differently. This is because the shape of the data used in this study is different from the input shape handled by Original ResNet18. Since the shape of the MFCC is 100x501, a 6x7 layer is used instead of a 7x7 filter in the first convolutional layer. A 3x4 filter is also used in the first down-sampling layer. Since the Spectrogram is 201x501, a 4x3 filter is used only in the second down-sampling layer. Since the chromatogram is 12x501, a 6x7 layer is used instead of a 7x7 filter in the first convolutional layer. And a 2x3 filter is used in the second down-sampling layer.

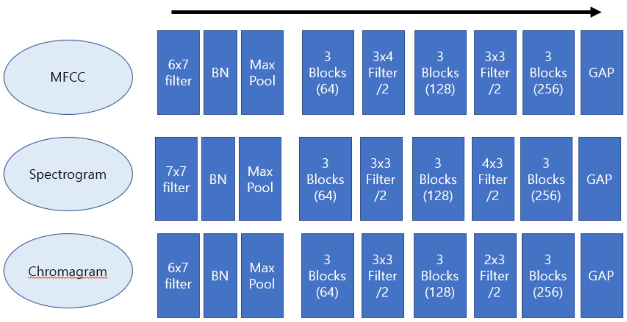


Figure 4. ResNet18 version 2

5.3 LSTM(Long-Short Term Memory)

LSTM is a model suitable for time series data with a continuous order [12]. In this study, the filter channel size of the hidden layer is fixed to 64, and the outputs from each hidden layer through 501 sequences are finally passed to the FC layer. When a multimodal model is entered into a submodule, the dimension of the final output vector is 256, and when it is used as a single model, it is 5 dimensions.

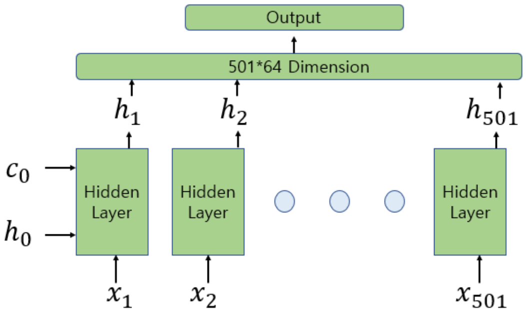


Figure 5. LSTM model

5.4 Multimodal

The Multimodal concatenates three 256-dimensional output vectors obtained by passing three data types through a user-specified submodule, then passes through the FC layer to reduce it to 256 dimensions and passes through the FC layer again to obtain the final output vector. Since three types of data are bundled in a tuple format and received as input at once and passed through a sub neural network at the same time, it can be trained using three types of data at once.

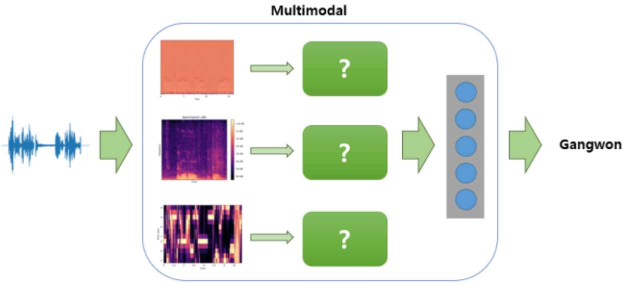


Figure 6. Multimodal model

6. Experiment and Analysis

In this section, we experimented with various models (LSTM, ResNet18, ResNet18 v2) to find out which one is the best model for each type of data. So, we trained each model respectively, and combine the best performing models to create a final multimodal model. We used the Adam optimizer cross-entropy loss function because these are the popular optimizer and loss function for the classification problem.

6.1 Evaluation Method

We used an accuracy and confusion matrix to evaluate our model. In particular, the confusion matrix is very useful to our project. Because we classify five regional confusion matrix is a good method for comfort-label classification.

6.2 Best Hyper-parameter Setting

We used K-Fold validation to find the best hyper-parameter for each model. K-Fold validation is a widely used method of hyper-parameter tuning. We used K-fold (K=5) validation. The hyper-parameter candidates were learning rate and epoch. The candidates of learning rates were 0.001, 0.0005and 0.0001 and the candidates of the epoch were 15 and 30. The hyper-parameters of each model and each data type are shown in Table 2.

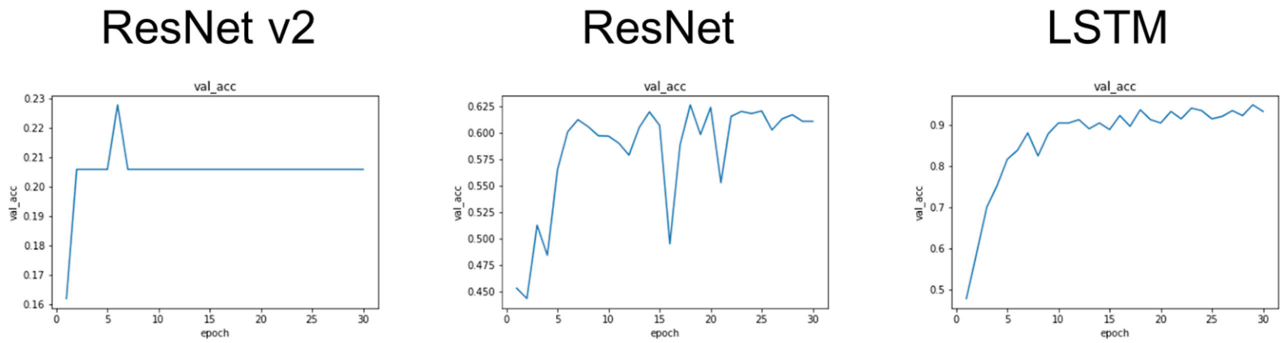


Figure 7. Test accuracy of each neural networks for Spectrogram

Table 2. Cross-validation

Data Type & Models		**LSTM**	**ResNet**	**ResNet v2**
MFCC	Epoch	30	30	30
	Learning rate	0.001	0.0005	0.0001
Chromagram	Epoch	30	30	30
	Learning rate	0.0005	0.0001	0.0001
Spectrogram	Epoch	30	30	30
	Learning rate	0.0001	0.0001	0.0001

6.3 Final model

We experimented to know which model is most suitable for each data type. We trained each model with the best hyper-parameters that were found in Table 2. And the result was quite interesting. Each model shows a different performance for each data type. For example, in the Spectrogram, LSTM shows the best performance while ResNet v2 shows poor performance. On the other hand, in the Chromagram, ResNet shows the best performance among the three models. The test accuracy graph of each model in Spectrogram is shown in Figure 7. Also, the test accuracy of each model for each data type is shown in Table 3. LSTM shows the best performance in MFCC, LSTM shows the best performance in Spectrogram and ResNet shows the best performance in Chromagram. Therefore, we use ResNet in Chromagram and LSTM in the rest to construct our final multimodal model.

We trained our final multimodal model (LSTM LSTM ResNet) with the best hyper-parameters of the final model. The best learning rate of the final model is 0.0001 and the best epoch of the final model is 30. Train loss and test accuracy are shown in Figure 8. It shows high performance than the unimodal model. The confusion matrix about the final model is shown in Figure 9. Also, we trained the multimodal model with only ResNet because we were curious about the difference between the multimodal model with the models that performed best in each data type and the multimodal model constructed regardless of the performance of each data type. In addition, we were curious about the difference between the multimodal model and the unimodal model. Table 4 shows all of the test accuracy and precision, recall, and f1 scores of the final model and benchmark models.

Table 3. Single model Test Accuracy

Data Type & Models	**LSTM**	**ResNet**	**ResNet v2**
MFCC	0.9675	0.6	0.17
Chromagram	0.5	0.825	0.1975
Spectrogram	0.93	0.61	0.205

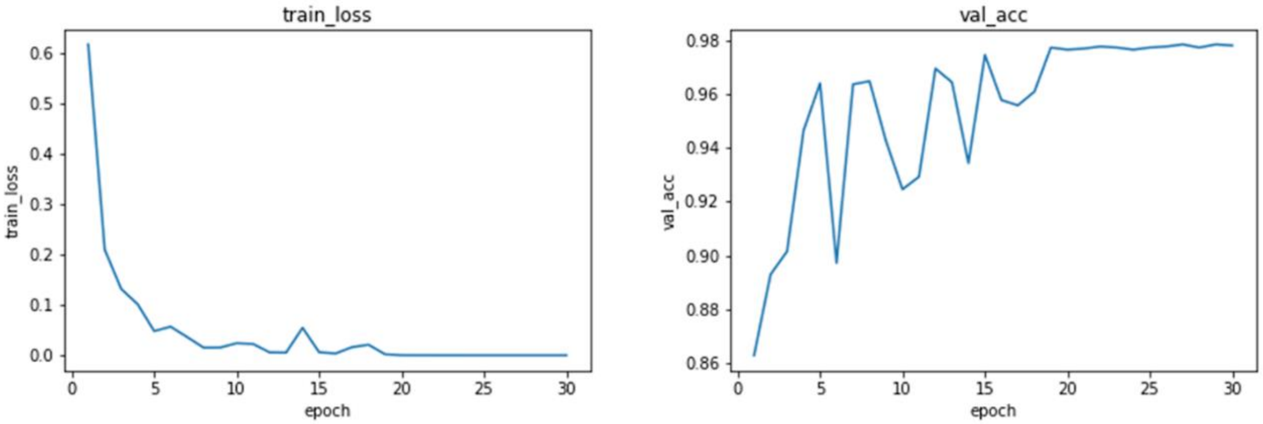
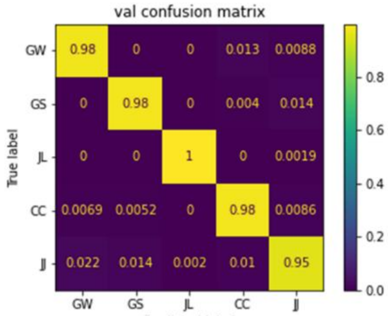


Figure 8. Train loss and Test accruacy of Final Multimodal model



	GW	GS	JL	CC	JJ
GW	0.98	0	0	0.013	0.0088
GS	0	0.98	0	0.004	0.014
JL	0	0	1	0	0.0019
CC	0.0069	0.0052	0	0.98	0.0086
JJ	0.022	0.014	0.002	0.01	0.95

Figure 9. Confusion matrix of Final Multimodal model

6.4 Analysis

From the result the Table 4, we discovered that our final multimodal model was better than the benchmark models. It performs better than any unimodal model and even multimodal model with only ResNet. One more interesting thing is that the multimodal model with only ResNet overwhelms other benchmarks that are trained with only one data type. Individual ResNet performs not well in each data type. For example, in MFCC it was very low accuracy and F1 score. But the Multimodal model with only ResNet performs very well. In this phenomenon, we could see that a multimodal model combined with weak models is better than a single strong model (ex. LSTM with MFCC). And we could see that making a multimodal model with the best-performed model in each data is the best choice for the formal model.

7. Conclusion

In this project, we make a multimodal model with three best-performed models for three data types (MFCC, Chromagram, Spectrogram) to classify Korean dialect. The accuracy of the final multimodal model was 0.978116 and F1 score was 0.977858, and the value overwhelms other benchmarks. It shows much better performance than the previous works about Korean dialect classification.

However, we did experiment with only three base models for multimodal (LSTM ResNet ResNet v2). In the future, we hope to do more experiments with other models like the transformer model or VGG, etc.

8. REFERENCES

[1] 이태영, 2011, 전라북도 방언 연구

[2] Young Kook Kim, Myung Ho Kim, 2021, Performance Comparison of Korean Dialect Classification Models Based on Acoustic Feature

[3] Fan Xu, Yangjie Dan, Keyu Yan, Yong Ma, and Mingwen Wang, 2021, Low-Resource Language Discrimination toward Chinese Dialects with Transfer Learning and Data Augmentation

[4] Rishabh Upadhyay, Simon Lui, 2018, Features of Speech Audio for Accent Recognition

[5] Dhanesh Ramachandram, Graham W. Taylor, 2017, Deep Multimodal Learning: A Survey on Recent Advances and Trends

[6] Wyse, L. 2017. Audio Spectrogram Representations for Processing with Convolutional Neural Networks, arXiv:1706.09559, DOI=https://doi.org/10.48550/arXiv.1706.09559

[7] Beth, Logan. 2020. Mel Frequency Cepstral Coefficients for Music Modeling, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.9216, Last Accessed 19 May. 2022.

[8] Meinard Müller, 2015, Fundamentals of Music Processing

[9] AIHub, https://aihub.or.kr/

[10] Librosa, https://librosa.org/

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, 2015, Residual Learning for Image Recognition

[12] Sepp Hochreiter, Jurgen Schmidhuber, 1997, LONG SHORT-TERM MEMORY

Table 4. Evaluation metrics

Models & Metrics	accuracy	precision	recall	F1 score
Multimodal (LSTM LSTM ResNet)	**0.978116**	**0.977775**	**0.977941**	**0.977858**
Multimodal (Only ResNet)	0.969236	0.968753	0.971017	0.969884
LSTM (Chromagram)	0.504493	0.511109	0.500062	0.505525
LSTM (MFCC)	0.968000	0.966519	0.967647	0.967149
LSTM (Spectrogram)	0.872215	0.870695	0.871233	0.870964
ResNet (Chromagram)	0.831473	0.842913	0.823304	0.832933
ResNet (MFCC)	0.355998	0.242559	0.395329	0.300650
ResNet (Spectrogram)	0.610981	0.463008	0.577014	0.513762
ResNet v2 (Chromagram)	0.197544	0.039508	0.199999	0.065983
ResNet v2 (MFCC)	0.182000	0.036399	0.199999	0.061590
ResNet v2 (Spectrogram)	0.211019	0.042203	0.199999	0.069699