Acoustic Feature-Based Recognition of Palestinian Regional Accents – Ramallah, Palestine

Faculty of Engineering and Technology, Birzeit University
Department of Electrical and Computer Engineering
Dunia Jaser-1201345, Ayah Saad-1191334, Aseel Shadid-1192463

Abstract

Palestinian Accent Recognition is the procedure of determining the speaker's regional accent by analyzing several characteristics in the related audio waveform. This project proposes machine learning models to detect Palestinian accents using acoustic waves. The models are trained using a dataset consisting of short speech segments. The dataset has four accents: Hebron, Jerusalem, Nablus, and Ramallah. Mel-Frequency Cepstrum Coefficients (MFCCs) are employed to extract features from the input acoustic signal. The collected features correspond to the matrix of the speaker's characteristics, which is employed for automated recognition using Support Vector Machine (SVM), and K-Nearest Neighbors (KNN).

1. Introduction

Speech serves as the primary means of communication among people. Over the past few decades, significant research has been devoted to automatic speech recognition (ASR) to enable machines to comprehend human speech, facilitating tasks ranging from simple automation to complex human-machine interactions. When machines convert human speech into signal form, a wealth of information becomes available, including the speaker's gender, accent, language, age, and emotional state.

Recently, the identification of dialects and accents has become as important as general speech recognition tasks. Identifying dialects and accents can reveal a speaker's nationality and cultural background. It is important to note that accents and dialects, although often used interchangeably, are different. Accents refer to pronunciation variations, whereas dialects encompass grammatical, lexical, and phonological differences [1].

Dialect Identification (DID) is a specialized case of Language Identification (LID) and is generally more challenging. DID plays a critical role in various applications, including ASR, intelligent speech-interactive devices like Siri and Google Home, text-to-speech (TTS) synthesis, and machine translation (MT) [2].

Recent research in speaker dialect identification is divided into two main approaches: acoustic and phono-tactic, each employing various machine learning methods. This project aims to develop and evaluate a simple acoustic system for recognizing regional Palestinian accents, specifically Jerusalem, Nablus, Hebron, and Ramallah. The system will classify short speech segments into one of these four accents using acoustic features extracted from the speech.

2. Background/Related Work:

Identifying Arabic dialects has proven to be challenging due to the huge linguistic differences between different places. Different machine learning techniques have been utilized to address this issue, with each utilizing distinct feature extraction methodologies and classification algorithms.

El-Haj et al. (2022) conducted a comprehensive study on Arabic dialect identification using multiple machine learning methods. They explored several classifiers, including Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and deep learning approaches, to identify different Arabic dialects. Their study highlighted the effectiveness of feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCCs) and i-vectors in capturing the unique characteristics of each dialect. The results demonstrated that SVM and deep learning models provided high accuracy rates, underscoring the potential of these techniques in dialect identification tasks [3].

Previous works have also emphasized the importance of feature selection and dimensionality reduction in enhancing model performance. For instance, the use of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) has been shown to improve classification accuracy by reducing noise and focusing on the most relevant features. These methodologies contribute to a more robust and efficient dialect identification system [4].

Overall, the integration of advanced feature extraction methods with powerful machine learning algorithms forms the foundation of current research in Arabic dialect identification. This project enhances existing methods by implementing and evaluating the effectiveness of Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) classifiers. The objective is to improve and streamline the process of identifying dialects.

3. Methodology

This section details the methodology employed to develop and evaluate a Palestinian regional accent recognition system. The methodology encompasses data preparation, feature extraction, model training, and evaluation. Using the most recent methods available in the field of accent and dialect recognition, we aim to establish a baseline system using acoustic systems, using Mel Frequency Cepstral Coefficients (MFCC) and modeling techniques like k-Nearest Neighbors(KNN) and Support Vector Machines (SVM) for accent and dialect recognition.

3.1. Data Preparation

The provided datasets were downloaded and organized into training and testing sets. Each dataset contained .wav files corresponding to different Palestinian accents: Hebron, Jerusalem, Nablus, and Ramallah_Reef. The audio files were systematically loaded and labels were assigned based on the accent category.

3.2. Extract Features

The librosa library was utilized to extract acoustic features from the audio files, such as Mel-Frequency Cepstral Coefficients (MFCCs), where MFCC is a type that represents the Cepstral features based on Mel filter banks. The Mel scale, a perceptual scale of pitches judged by listeners to be equal in distance from one another, is designed to capture the way

humans perceive the frequency of sounds, which differs from a linear relationship. The formula for converting a frequency f in Hertz to its corresponding Mel scale value M is given by:

$$M(f) = 2595 \times log_{10}(1 + \frac{f}{700}) \quad (1)$$

In audio processing, this transformation is important because it ensures that the frequency components of signals are represented in a way that is more in line with human auditory perception, especially in speech recognition. This conversion is crucial for creating features like the Mel-Frequency Cepstral Coefficients (MFCCs), which are widely used in various applications, including accent and speech recognition systems.

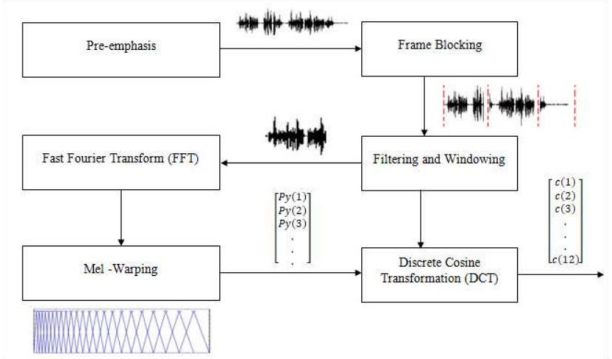

Figure 1: The process of extracting MFCC coefficients [3]

3.3. Data Normalization and Feature Selection

To standardize the features, normalization was conducted by removing the mean and scaling to unit variance, ensuring that each feature contributed equally to the model's learning process. Additionally, the top 30 features most relevant to the target variable were selected using the SelectKBest method with an ANOVA F-test. This step reduced dimensionality and enhanced the model's performance by focusing on the most significant features.

3.4. Proposed Models

3.4.1. Support Vector Machine (SVM) with Grid Search

Different techniques were used to solve the Arabic dialect identification problem. The previously extracted features (such as MFCC) were leveraged as input in the proposed technique.

A Support Vector Machine (SVM) classifier, a powerful supervised learning model used for classification tasks, was utilized. The SVM works by finding the optimal hyperplane that maximizes the margin between different classes in the feature space. This hyperplane is determined by support vectors, which are the data points nearest to the hyperplane and most critical in defining its position.
We trained a Support Vector Machine (SVM) classifier using Grid Search with Cross-Validation to find the best hyper-parameters. The parameters optimized included the penalty parameter C and the kernel type. Cross-Validation was used to prevent overfitting and ensure model performance across different data subsets. This enhanced the SVM's generalization ability to new data.

3.4.2. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) classifier is a simple yet effective algorithm that works by classifying a data point based on the majority class among its k-nearest neighbors in the feature space. We utilized a Grid Search with Cross-Validation approach to tune the hyper-parameters of the KNN classifier, specifically focusing on the number of neighbors (k) and the weight function used in prediction. This method allowed us to identify the optimal configuration for the KNN model, enhancing its predictive performance on the Arabic dialect identification task.

3.5. Model Evaluation

The model was evaluated using a classification report, which provided precision, recall, and F1-score for each class. This report was essential in understanding the model's performance in distinguishing between different accents. And to visualize the performance of the classifier, we plotted a confusion matrix. The matrix illustrated the true positive rates and the misclassification rates, offering insights into which accents were often confused with each other.

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

Where:

TP: is the number of true positives.
FP: is the number of false positives.

$$\boldsymbol{Recall} = \frac{TP}{FN+TP} \quad (3)$$

Where:

FN: is the number of false negatives.

$$\boldsymbol{F1-Score} = 2 \cdot (\frac{Precision \cdot Recall}{Precision+Recall}) \quad (4)$$

Table 1: **Confusion Matrix**

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Where:

- TP: is the number of true positives.
- FP: is the number of false positives.
- FN: is the number of false negatives.
- TN: is the number of true negatives.

3.6. Testing with New Data

The model was further evaluated on a new set of data to assess its generalization ability. The model's predictions on the new data were compared to the true labels to evaluate its accuracy and overall performance.

This comprehensive methodology ensured the systematic development, training, and evaluation of our accent recognition system.

4. Experiments and Results

4.1.1. Dataset

The dataset used for this project consists of speech recordings from four Palestinian regional accents: Hebron, Jerusalem, Nablus, and Ramallah_Reef. These recordings were provided in `.wav` format, with each file representing a short speech segment. The dataset was carefully curated to include a balanced representation of each accent, ensuring that the models could learn and generalize effectively.

The dataset was organized into two separate folders: one for training the models and another for testing the models. The

training folder contains approximately 70% of the data, while the testing folder contains the remaining 30%. This split ensures that the models are trained on a substantial amount of data while still being evaluated on unseen data to assess their generalization capabilities.

4.1.2. Model Training and Evaluation

4.1.2.1 Support Vector Machine (SVM)

The SVM model was trained using Grid Search with Cross-Validation to find the best hyper-parameters. The optimal parameters were found to be `C = 0.05` and `kernel = 'linear'`. The best cross-validation score achieved was 0.74. The model's performance was then evaluated on the test set.

Table 2: SVM Classification Report on training data

	Precision	Recall	F1-Score
Hebron	1.00	0.67	0.80
Jerusalem	0.67	0.67	0.67
Nablus	1.00	1.00	1.00
Ramallah_Reef	0.50	0.67	0.57
Accuracy			0.75
Macro Avg	0.79	0.75	0.76
WeightedAvg	0.79	0.75	0.76

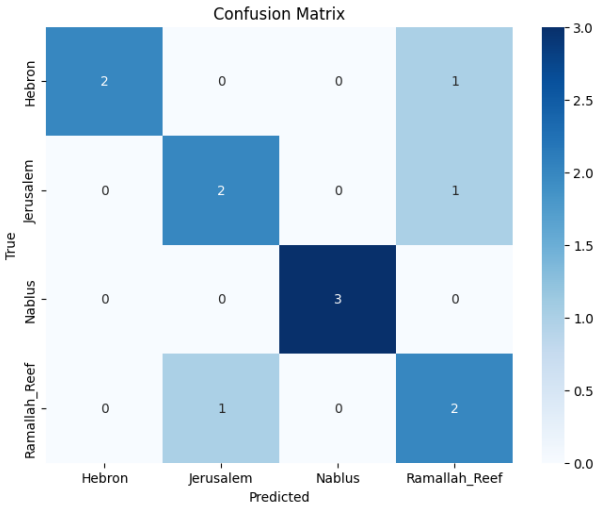

Figure 2: SVM Confusion Matrix on Training Data

- **Hebron:** Out of 3 samples, 2 were correctly predicted as Hebron, while 1 was incorrectly predicted as Ramallah_Reef.
- **Jerusalem:** Out of 3 samples, 2 were correctly predicted as Jerusalem, while 1 was incorrectly predicted as Ramallah_Reef.
- **Nablus:** All 3 samples were correctly predicted as Nablus.
- **Ramallah_Reef**: Out of 3 samples, 2 were correctly predicted as Ramallah_Reef, while 1 was incorrectly predicted as Jerusalem.

The overall accuracy score achieved by the SVM model was 75%.

Table 3: SVM Classification Report on Testing Data

	Precision	Recall	F1-Score
Hebron	1.00	0.60	0.75
Jerusalem	0.67	0.67	0.67
Nablus	1.00	1.00	1.00
Ramallah_Reef	0.50	0.67	0.57
Accuracy			0.75
Macro Avg	0.79	0.75	0.76
WeightedAvg	0.79	0.75	0.76

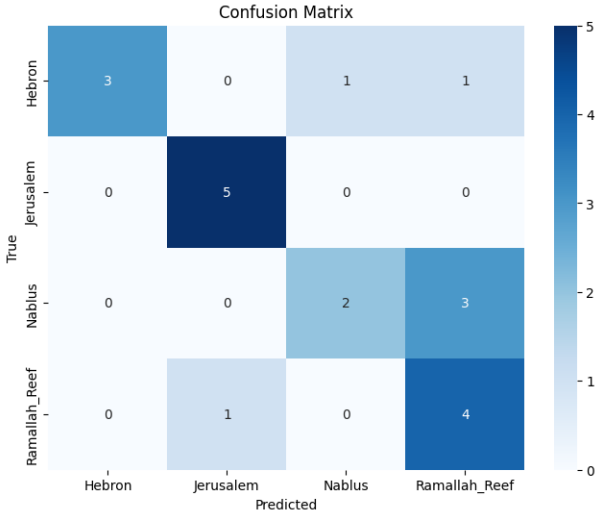

Figure 3: SVM Confusion Matrix on Testing Data

- **Hebron:** Out of 5 samples, 3 were correctly predicted as Hebron, while 1 was incorrectly predicted as Nablus, and 1 as Ramallah_Reef.
- **Jerusalem:** All 5 samples were correctly predicted as Jerusalem.
- **Nablus:** Out of 5 samples, 2 were correctly predicted as Nablus, while 3 were incorrectly predicted as Ramallah_Reef.
- **Ramallah_Reef:** Out of 5 samples, 4 were correctly predicted as Ramallah_Reef, while 1 was incorrectly predicted as Jerusalem.

The overall accuracy score achieved by the SVM model on the testing data was 70%.

4.1.2.2 K Nearest Neighbor (KNN)

The KNN model was trained using Grid Search with Cross-Validation. The optimal parameters were found to be `n_neighbors = 4` and `weights = 'distance'`. The best cross-validation score achieved was 0.71. The model's performance was then evaluated on the test set.

Table 4: KNN Classification Report on Training Data

	Precision	Recall	F1-Score
Hebron	1.00	0.33	0.50
Jerusalem	0.40	0.67	0.50
Nablus	0.60	1.00	0.75

Ramallah_Reef	0.00	0.01	0.00
Accuracy			0.50
Macro Avg	0.50	0.50	0.44
WeightedAvg	0.50	0.50	0.44

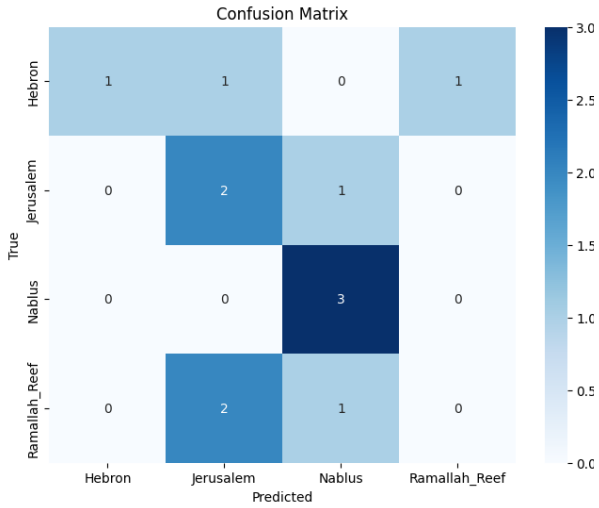

Figure 4: KNN Confusion Matrix on Training Data

- **Hebron:** Out of 3 samples, 1 was correctly predicted as Hebron, while 1 was incorrectly predicted as Jerusalem, and 1 as Ramallah_Reef.
- **Jerusalem:** Out of 3 samples, 2 were correctly predicted as Jerusalem, while 1 was incorrectly predicted as Ramallah_Reef.
- **Nablus:** All 3 samples were correctly predicted as Nablus.
- **Ramallah_Reef:** Out of 3 samples, 2 were incorrectly predicted as Jerusalem, and 1 as Nablus.

The overall accuracy score achieved by the KNN model on the training data was 0.50.

5. Conclusion and future work

In this project, we developed and evaluated a Palestinian regional accent recognition system using advanced machine learning techniques. The MFCC features were extracted from the signals. A Support Vector Machine (SVM) classifier and K-Nearest Neighbor (KNN) were used, and there results were compared. The results obtained reveal that the identification accuracy using SVM is much better compared to the KNN. The SVM demonstrated high accuracy on training data with an overall accuracy of 75%, indicating good discrimination between classes. On testing data, it achieved an overall accuracy of 70%, maintaining consistent performance despite variations in precision and recall across different classes. However, the KNN model achieved a lower average accuracy of 50% on the training data, indicating that it was less effective in distinguishing between the different accents compared to the SVM.

However, there are several areas for future improvement, including exploring additional acoustic features, implementing deep learning models, expanding the dataset to include more samples and additional regional accents, employing data augmentation techniques, developing a real-time accent recognition application, and integrating additional modalities such as visual lip movement or text transcriptions.

6. Partner's participation tasks

- **Ayah Saad:** Data preprocessing and feature extraction and report writing.
- **Dunia Jaser:** Model training and hyper-parameter tuning and report writing.
- **Aseel Shadid:** Evaluation and report writing.

7. References

[1] H. Jiang, "Speech recognition based on deep learning: A systematic review," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 106-115, Jan. 2016. [Online]. Available: https://ieeexplore.ieee.org/document/7343088

[2] A. Dosovitskiy, J. Tobias Springenberg, and T. Brox, "Learning to Generate Chairs with Convolutional Neural Networks," arXiv preprint arXiv:1509.06928, 2015. [Online]. Available: https://arxiv.org/abs/1509.06928

[3] M. El-Haj, D. Rayson, R. Moore, and A. Barras, "Arabic Dialect Identification Using Different Machine Learning Methods," in *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France, Jun. 2022. [Online]. Available: https://www.researchgate.net/publication/361159745_Arabic_Dialect_Identification_Using_Different_Machine_Learning_Methods.

[4] I. Jolliffe, Principal Component Analysis, 2nd ed. New York, NY, USA: Springer, 2002.

8. Appendix

The code was executed in a Google Colab environment to ensure reproducibility and ease of access. You can access the full implementation of our code using the following Google Colab link:
https://colab.research.google.com/drive/1QJIuU-HIFqOKYcqooMyxmnMuBMQKFgww?usp=sharing