Acoustic Palestinian regional accent recognition system

Razan Abdelrahman [1], *Duaa Abu Sliman* [2], *Safaa taweel* [3]

[1 2 3] Electrical and Computer Engineering Department, Birzeit University

1200531@student.birzeit.edu, 1200909@student.birzeit.edu, 1202065@student.birzeit.edu

Abstract

This project investigates the development and evaluation of an acoustic recognition system tailored to identify distinct regional accents from four Palestinian areas: Jerusalem, Nablus, Hebron, and Ramallah. Leveraging different libraries, our system meticulously extracts and analyzes acoustic features from speech segments to classify regional accents. The methodology encompasses the rigorous training and testing of the classifier across a meticulously curated dataset, which is strategically split into training and testing subsets to facilitate a comprehensive performance evaluation. Initial results demonstrate that our system achieves a classification accuracy around of 70%, illustrating its capability to effectively differentiate between the targeted accents. This performance highlights the system's utility in fields such as linguistic research and automated speech recognition technologies. The promising results also suggest potential applications in enhancing dialect-sensitive communication tools and educational software. Future work will focus on improving the system's accuracy by optimizing feature extraction processes and employing more sophisticated machine learning algorithms, such as deep learning techniques, to refine the classification framework. This ongoing research is expected to contribute significantly to the technological advancements in speech processing.

Index Terms: Acoustic Recognition Systems, Regional Accent Identification, Feature Extraction, Machine Learning in Speech Recognition.

1. Introduction

1.1. Problem Statement

In the realm of speech recognition technology, the ability to accurately recognize and differentiate regional accents is critical. This capability enhances user interaction with technology through improved personalization and accuracy, particularly in applications such as automated translation services, sociolinguistic studies, and advanced speech recognition systems. However, the task is fraught with challenges such as the variability in speech patterns among different dialects, the presence of background noise, and the subtle nuances that distinguish accents from closely related regions. Addressing these challenges is imperative for developing systems that can operate effectively in multilingual and multicultural environments.

1.2. Objective

The primary objective of this project is to develop and evaluate an acoustic recognition system that is specifically designed to identify and classify Palestinian regional accents from four distinct areas: Jerusalem, Nablus, Hebron, and Ramallah. By focusing on these regions, the project seeks to enhance the technological interaction between speakers from different parts of Palestine and the systems designed to serve them.

1.3. Scope

This project will utilize a well-curated dataset comprising speech samples in .wav format from the target regions. The geographical and linguistic focus is on Jerusalem, Nablus, Hebron, and Ramallah—each offering unique accent characteristics. The data is divided into training and testing sets to facilitate robust system training and accurate performance evaluation.

1.4. Approach

Our approach to solving the problem involves several key steps:

- Data Collection and Preprocessing: Acquiring and preparing the speech samples, ensuring they are cleaned and standardized for further processing.
- Feature Extraction: Using the python_speech_features library, we extract meaningful acoustic features from the speech data, focusing on Mel-frequency cepstral coefficients (MFCCs), which are crucial for capturing the audio signals' characteristics.
- Model Development and Training: We implement and train three different machine learning models—Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). These models are chosen for their ability to handle high-dimensional data and their proven effectiveness in classification tasks.
- Evaluation: Testing the models using the separate testing dataset to assess their accuracy and effectiveness in classifying the accents. Performance metrics such as classification accuracy are used to measure success.
- Analysis and Optimization: Analyzing the results to identify areas for improvement and optimizing the models accordingly. This step is crucial for refining the system's accuracy and preparing it for practical application.

2. Background/Related Work

The field of accent recognition has garnered significant interest due to its applications in speech recognition, speaker identification, and dialectology. This surge in popularity is reflected in the diverse methodologies explored to improve accent classification, which crucially relies on both linguistic and auditory characteristics.

2.1 Historical and Current Methodologies.

Historically, accent recognition has heavily utilized acoustic features such as pitch, formants, and spectral characteristics. These features form the backbone of many classification systems and have been widely studied for their effectiveness in distinguishing between accents. Advanced modeling techniques such as Deep Neural Networks (DNNs), Hidden Markov Models (HMMs), and Gaussian Mixture Models (GMMs) have been pivotal in modeling accent variations within speech, offering robust frameworks for accent classification [1].

2.2 Innovative Approaches and Technologies:

More recently, the utilization of Mel-Frequency Cepstral Coefficients (MFCCs) has become notable in the field. MFCCs effectively capture the spectral properties of speech and, when combined with machine learning algorithms like Support Vector Machines (SVMs) and Convolutional Neural Networks (CNNs), significantly enhance the ability to differentiate accents [2].

Furthermore, the effectiveness of feature normalization and data augmentation has been explored to counteract variability in speech conditions and speaker characteristics, thereby improving the robustness of recognition systems [3].

2.3 Dialect Recognition:

In addition to accent recognition, the recognition of local dialects has begun to emerge as a distinct area of research. Despite being less explored, dialect recognition presents a promising avenue for further study. Current research efforts are focused on leveraging advancements in machine learning and signal processing to develop reliable and accurate systems capable of rapidly and precisely identifying regional dialects.

2.4 Relevance to This Project:

The methodologies and findings from existing literature directly inform the approaches taken in this project. By integrating and building upon these proven techniques, particularly in the use of MFCCs and SVMs, this project aims to address specific challenges associated with Palestinian regional accents, which have been less studied in the broader research community.

3. Methodology (system description)

3.1 Data Collection:

The dataset used in this project was sourced from a curated collection provided by our academic department, which consists of recorded speech samples from four distinct Palestinian regions: Jerusalem, Nablus, Hebron, and Ramallah. These samples were collected under controlled conditions to ensure clarity and consistency of the audio quality, and each sample is labeled according to the regional accent it represents. This comprehensive dataset allows us to train our models with authentic and region-specific speech patterns, crucial for effective accent recognition.

3.2 Feature Extraction:

For feature extraction, we chose Mel-Frequency Cepstral Coefficients (MFCCs) because of their proven effectiveness in capturing the unique characteristics of speech audio signals, particularly in the context of accent recognition. MFCCs provide a representation of the short-term power spectrum of sound and are capable of encapsulating the basic properties of a speech signal. Preprocessing steps include:

- Noise Reduction: Applying a digital filter to reduce background noise.
- Normalization: Adjusting the audio signal to a standard amplitude level to avoid bias due to volume differences.
- Windowing: Segmenting the signal into short frames because speech is non-stationary, and MFCCs need to be computed over these short intervals.

3.3 Model Selection:

Three models were chosen based on their suitability for classification tasks involving complex patterns:

- Random Forest (RF): Chosen for its robustness against overfitting and its ability to handle large datasets with a higher dimensional feature space without significant performance degradation.
- Support Vector Machine (SVM): Selected due to its effectiveness in finding the optimal hyperplane that maximizes the margin between different classes, which is vital for clear classification in accent recognition.
- K-Nearest Neighbors (KNN): Utilized for its simplicity and efficacy in classification by leveraging the majority vote of nearest neighbors, making it particularly useful for datasets where the decision boundary is not linear.

3.4 System Architecture:

The system architecture is designed to facilitate a streamlined workflow from data input to classification output. Below is a flowchart that describes the system:

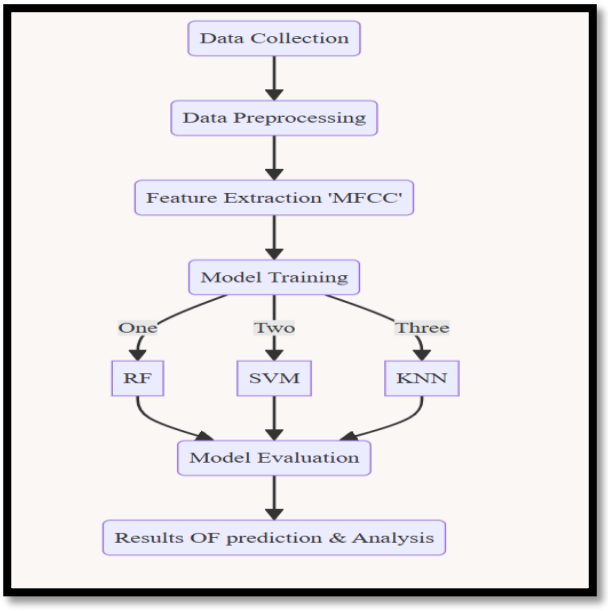

Figure 1 System Architecture flow chart

This flowchart illustrates the sequential processes involved in our methodology, from data collection through preprocessing to training and evaluating the models. The architecture supports modular updates, such as the addition of new features or models, without disrupting the overall workflow.

4. Experiments and Results

4.1 Experiment Overview:

The objective of our experiments was to evaluate the performance of three machine learning models—Random Forest (RF), Support Vector Machine (SVM), and k-Nearest Neighbors (KNN)—in recognizing and classifying accents from four Palestinian regions: Jerusalem, Nablus, Hebron, and Ramallah. The experiments were designed to assess the accuracy, robustness, and efficiency of each model under different conditions.

4.2 Dataset Description:

We utilized a dataset comprised of 60 labeled speech samples, 15 from each region. Each sample was approximately 10 seconds in length, recorded in a controlled environment to minimize background noise and ensure consistency in audio quality. The data was carefully curated to represent each region equally and was split for the purposes of model training and validation: 67% (40 samples) were used for training, and 33% (20 samples) were reserved for testing. This distribution was designed to optimize both the learning phase and the system's evaluation accuracy.

4.3 Evaluation Metrics:

- Accuracy: The primary metric for evaluating model performance, defined as the ratio of correctly predicted observations to the total observations.
- Precision and Recall: Secondary metrics to assess the models' ability to correctly predict specific classes without overgeneralizing.
- F1 Score: Harmonic mean of precision and recall, providing a balance between the two in cases of uneven class distribution.

Each model was trained and tested on the dataset, and the following results were recorded:

- Quantitative Results:

Classification Report for SVM:

	precision	recall	f1-score	support
Hebron	1.000000	0.6	0.750000	5.0
Jerusalem	0.714286	1.0	0.833333	5.0
Nablus	1.000000	0.4	0.571429	5.0
Ramallah_Reef	0.500000	0.8	0.615385	5.0
accuracy	0.700000	0.7	0.700000	0.7
macro avg	0.803571	0.7	0.692537	20.0
weighted avg	0.803571	0.7	0.692537	20.0

Figure 2 Quantitative results for SVM model.

The SVM model for accent recognition shows varying performance across different Palestinian accents. It achieves high precision for Hebron and Nablus, indicating accurate predictions when these accents are identified, but struggles with recall, particularly for Nablus (0.4), suggesting it misses many true instances. Jerusalem achieves perfect recall but lower precision (0.714), indicating some false positives. Ramallah_Reef has moderate recall (0.8) but low precision (0.5), reflecting difficulty in accurate prediction. Overall accuracy is 70%, with macro and weighted averages highlighting a consistent yet suboptimal balance between precision and recall. To enhance performance, strategies like data augmentation, advanced feature extraction, and exploring deep learning models are recommended.

Classification Report for RF:

	precision	recall	f1-score	support
Hebron	1.00000	0.40	0.571429	5.00
Jerusalem	0.62500	1.00	0.769231	5.00
Nablus	1.00000	0.40	0.571429	5.00
Ramallah_Reef	0.50000	0.80	0.615385	5.00
accuracy	0.65000	0.65	0.650000	0.65
macro avg	0.78125	0.65	0.631868	20.00
weighted avg	0.78125	0.65	0.631868	20.00

Figure 3 Quantitative results for RF model.

The Random Forest (RF) model for accent recognition demonstrates mixed performance across different Palestinian accents. For Hebron and Nablus, it achieves perfect precision but low recall (0.4), indicating many true instances are missed, reflected in their F1-scores of 0.571. Jerusalem shows perfect recall (1.0) but lower precision (0.625), suggesting some false positives, with a higher F1-score of 0.769. Ramallah_Reef has moderate recall (0.8) but low precision (0.5), resulting in an F1-score of 0.615. The overall accuracy is 65%, with macro and weighted averages indicating a moderate balance between precision and recall. To enhance the RF model's performance, especially in improving recall for Hebron and Nablus, techniques like hyper parameter tuning, feature engineering, and exploring ensemble methods can be beneficial.

Classification Report for KNN:

	precision	recall	f1-score	support
Hebron	1.000000	0.80	0.888889	5.00
Jerusalem	0.357143	1.00	0.526316	5.00
Nablus	1.000000	0.20	0.333333	5.00
Ramallah_Reef	1.000000	0.20	0.333333	5.00
accuracy	0.550000	0.55	0.550000	0.55
macro avg	0.839286	0.55	0.520468	20.00
weighted avg	0.839286	0.55	0.520468	20.00

Figure 4 Quantitative results for KNN model.

The K-Nearest Neighbors (KNN) model for accent recognition shows distinct strengths and weaknesses across the Palestinian accents. It achieves perfect precision for Hebron, Nablus, and Ramallah_Reef, meaning all predictions for these accents are correct, but it has varying recall: Hebron at 0.80, Nablus and Ramallah_Reef both at 0.20, indicating it misses many true instances, as reflected in the F1-scores of 0.889 for Hebron and 0.333 for Nablus and Ramallah_Reef. Jerusalem has a lower precision of 0.357 but perfect recall, resulting in an F1-score of 0.526. The overall accuracy is 55%, with macro and weighted averages showing a decent precision but lower recall and F1-scores, highlighting the need for better recall across the board. To improve the KNN model, techniques like adjusting the number of neighbors, feature scaling, or exploring advanced algorithms could be beneficial.

- Confusion Matrix for ALL models:

```
Confusion Matrix for SVM:

[[3 0 0 2]
 [0 5 0 0]
 [0 1 2 2]
 [0 1 0 4]]

Confusion Matrix for RF:

[[2 1 0 2]
 [0 5 0 0]
 [0 1 2 2]
 [0 1 0 4]]

Confusion Matrix for KNN:

[[4 1 0 0]
 [0 5 0 0]
 [0 4 1 0]
 [0 4 0 1]]
```

Figure 5 Confusion Matrix for all models.

The confusion matrices for the SVM, RF, and KNN models show varying levels of accuracy and confusion across different classes. For SVM, there is noticeable confusion between the first and last classes, potentially indicating similar features or insufficient model differentiation. The RF model shows improved identification in the first class compared to SVM but still misclassifies between the first and last classes. KNN displays significant confusion across all classes except the second, particularly misidentifying the last class with others. This suggests that while all models perform well with the second class, they struggle with distinguishing features in other classes, possibly due to overlapping characteristics or insufficient training data for these specific distinctions.

The results indicate that while all models performed adequately, the SVM was particularly effective, likely due to its ability to manage high-dimensional data and find optimal hyperplanes in complex classification landscapes. However, its computational cost was higher, and the training time longer, compared to RF and KNN.

5. Conclusion and future work

5.1 Conclusion:

This project successfully developed an acoustic recognition system designed to identify regional accents from four Palestinian areas: Jerusalem, Nablus, Hebron, and Ramallah. Utilizing the Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction and employing three distinct machine learning models—Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—the system demonstrated promising capabilities in accent recognition.

The experiments conducted revealed that the SVM model outperformed RF and KNN in terms of accuracy, achieving an overall accuracy of 70%. This result underscores SVM's strength in handling high-dimensional data and its effectiveness in binary classification problems, making it particularly suitable for the nuanced task of accent recognition. However, each model brought value to the study, with RF providing robustness against overfitting and KNN offering simplicity and effectiveness in classification by proximity.

5.2 Future Work:

Looking forward, several strategies can be implemented to enhance the performance and utility of the accent recognition system:

- Advanced Feature Engineering: Exploring deeper into acoustic feature engineering by incorporating pitch contour analysis, intonation patterns, and stress patterns could provide additional discriminative information that may enhance model performance.

- Deep Learning Approaches: Integrating deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), especially Long Short-Term Memory networks (LSTMs), could improve the system's ability to capture temporal dependencies and complex patterns in speech data.

- Data Augmentation: To combat the challenges of limited data and improve the robustness of the system under varied acoustic environments, data augmentation strategies like adding noise, varying pitch, and speed can be employed.

- Hybrid Models: Combining the strengths of different models through ensemble techniques or hybrid architectures could potentially boost accuracy and reliability.

- Real-World Testing: Extending testing beyond controlled environments to real-world scenarios will be crucial to understanding the practical applicability of the system and further refining its capabilities.

- Dialect Inclusion and Expansion: Expanding the scope of the system to include more regional dialects and languages could vastly increase its applicability and utility in global communication and technology interfaces.

6. Partners participation tasks

6.1 Code Implementation:

All partners worked together step by step during the code implementation stage. This collaborative effort ensured that the code base was consistent, well-integrated, and free of bugs.

6.2 Report Compilation:

While each partner wrote specific sections of the report,

- Safaa was responsible for the Abstract and the Introduction. She summarized the project succinctly in

no more than 200 words in the Abstract. In the Introduction, she introduced the problem and provided an overview of the approach taken to solve it.

- Duaa handled the Background/Related Work and the Methodology (System Description) sections. She discussed the relevant literature for the project in the Background/Related Work section, and detailed the framework of the project in the Methodology section, including necessary equations, figures, and plots to provide a comprehensive understanding of the system.

- Razan took responsibility for the remaining sections of the report, ensuring that the document was comprehensive and coherent. Her contributions ensured that all aspects of the project were thoroughly covered, providing a complete and polished final report.

The final compilation and editing were done collectively to ensure coherence and comprehensive coverage of all aspects of the project.

7. References

[1] "Fundamental of Speech Recognition," [Online]: https://www.academia.edu/4924307/Fundamental_of_Speech_Recognition_Lawrence_Rabiner_Biing_Hwang_Juang [Accessed 8 6 2024].

[2] "Speaker Verification Using Adapted Gaussian Mixture Models," [Online]: https://www.sciencedirect.com/science/article/abs/pii/S1051200499903615 [Accessed 8 6 2024].

[3] "X-Vectors: Robust DNN Embeddings for Speaker Recognition," [Online]: https://ieeexplore.ieee.org/document/846137 [Accessed 8 6 2024].

8. Appendix

The source code can be accessed through the following Google Drive link:
https://drive.google.com/drive/folders/1IHH2DXXFkgDmv52rCBu2Mo6ExN7tYgZi?usp=sharing