Text-Independent Speaker Identification Using Arabic Phonemes

Samiha R. Alarjani [1,*], Imran Rao [2], Iram Fatima [3], and Hafiz Farooq Ahmad [1]

[1] Computer Science Department, College of Computer Sciences and Information Technology (CCSIT), King Faisal University, P.O. Box 400, Al-Ahsa 31982, Saudi Arabia
[2] Blue Brackets
[3] Independent Researcher
Email: 222401728@student.kfu.edu.sa, Same7a.567@gmail.com (S.R.A.); imranrao@gmail.com (I.R.); iram.fa@gmail.com (I.F.); hfahmad@kfu.edu.sa (H.F.A.)
*Corresponding author

***Abstract*—Speaker identification has become a fundamental aspect of digital speech processing and user authentication. However, challenges persist, including speaking in noisy environments, variations in speakers' emotions, and voice length. Despite the improvements in speech processing techniques, further enhancements are needed, especially in text-independent Arabic speaker identification. Therefore, this research aims to investigate the use of phonemes for speaker identification, which has not been investigated thoroughly in the Arabic language. This study involves using three machine learning models, namely the Gaussian Mixture Model (GMM), Support Vector Machine (SVM), and Gaussian Mixture Model-Universal Background Model (GMM-UBM). In addition, this research employs Mel-Frequency Cepstral Coefficients (MFCC) features and specially constructed datasets containing short Arabic phonemes gathered from 104 speakers. The machine learning models are evaluated using popular metrics such as accuracy, precision, recall, and F1-score. The results show that the GMM model performs better in identifying speakers based on 20 MFCCs, while the SVM model performs better for 40 MFCCs, with accuracies of 96.2% and 95.7%, respectively. The results of the analysis indicate that there are three minimum Arabic phonemes required to identify speakers accurately. This research provides good insights into Arabic speaker identification using short utterances, which can help in the future development of reliable speaker identification systems using short audio samples.**

***Keywords*—speaker identification, Arabic, phonemes, Mel-Frequency Cepstral Coefficients (MFCC), machine learning, Gaussian Mixture Model (GMM), Support Vector Machine (SVM), Gaussian Mixture Model-Universal Background Model (GMM-UBM)**

I. Introduction

Speaker identification has drawn the attention of researchers due to its significance in various domains, particularly cybersecurity and voice biometrics [1]. This technology involves using a machine to examine speech characteristics such as pitch, tone, cadence, and pronunciation to verify the identity of an unknown speaker [2]. Its applications range from voice-controlled technologies like Google Assistant to healthcare, digital assistants, voice-activated writing, and biometric authentication. Despite the advancements in speaker identification, the majority of these systems have primarily focused on English and other widely spoken European languages [3]. There remains a significant gap in research and development for Arabic speaker identification. This study aims to address this gap by proposing and evaluating the performance of three machine learning models for identifying Arabic speakers based on phonemes.

In any language, a phoneme is the smallest unit of sound that can distinguish one word from another [4]. The Arabic language has many different phonemes that contribute to the unique sounds of words. In our work, using phonemes for speaker identification can offer several advantages, including minimizing the computational complexity of the identification task compared to using complete words or sentences. Additionally, phonemes are more robust in noisy environments or with variations in pronunciation than full sentences. Furthermore, focusing on phonetic characteristics can help capture unique aspects of a speaker's voice that are consistent across different speech utterances. This approach is suitable for a wide range of applications in speech technology and biometric systems.

Arabic is one of the oldest and most significant languages globally, spoken by over 422 million people across 25 countries [5], making it one of the five most spoken languages worldwide. Its significance extends beyond the Middle East, impacting numerous languages, such as Urdu, which exhibits several phonetic and linguistic characteristics with Arabic. The study of Arabic phonetics not only aids in developing robust speaker identification systems but also contributes to the broader understanding of language evolution and phonetic diversity. Arabic has a complex phonetic structure, with unique sounds and pronunciation rules, which presents distinct challenges and opportunities for linguistic research and technological development. Moreover, the emphasis on Arabic speaker identification addresses a critical gap in

Manuscript received August 25, 2024; revised October 21, 2024; accepted November 15, 2024; published March 6, 2025.

current research, as most existing systems primarily focus on English and other widely spoken languages. Enhancing speaker identification systems for Arabic can significantly impact cybersecurity, access control, and voice-activated technologies in Arabic-speaking regions. This will provide a more inclusive approach to global technological advancements.

Arabic speaker identification remains a challenging area of research due to its unique linguistic characteristics and the complexity of phonetic analysis. Unlike languages with simpler phonetic structures, Arabic, with its complex phonetic structures, presents a wide range of phonemes and varying dialects across different countries, posing significant challenges for accurate speaker identification systems. In addition, current research often relies on complete sentences; however, this can be computationally intensive and less effective in real-world scenarios where short speech samples are most commonly used. Furthermore, environmental factors, speaker health, emotional state, microphone quality, and background noise can adversely affect system performance [6].

Identifying Arabic speakers involves several technical issues:

- Phonetic Diversity: Arabic exhibits a wide range of phonemes that vary significantly across different dialects.
- Environmental Factors: Background noise and recording conditions can introduce variability.
- Speaker Variability: Speech characteristics can be influenced by health, emotional state, and microphone quality.

To address these challenges, we suggest conducting a study on text-independent speaker identification for the Arabic language. The study will utilize three machine learning models: generative, discriminative, and hybrid models. This research aims to make a significant contribution to the field by proposing a robust framework for Arabic speaker identification. The framework utilizes Mel-frequency cepstral coefficients (MFCCs) for feature extraction and incorporates three classification algorithms: Gaussian Mixture Models (GMM), Support Vector Machines (SVM), and GMM-Universal Background Model (GMM-UBM).This study is a revised and expanded version of a paper entitled "Arabic Speaker Identification Based on Phonemes," presented at Proceedings of ARSSS International Conference [7]. This study provides more investigation of the Arabic phonemes for speaker identification tasks. As compared to the conference paper, this study provides a comprehensive analysis of the literature. Additionally, it examines the effect of using three and four phonemes on the accuracy of the classification models. Moreover, it provides a comprehensive analysis of all the results and suggests the best classification model for speaker identification.

This research aims to investigate the effectiveness of using short speech utterances, such as Arabic phonemes, to identify speakers. In addition, the study evaluates the effectiveness of the aforementioned machine learning models in different recording conditions and offers insights into the top-performing models for Arabic speaker identification.

To overcome these challenges, we investigated three types of machine-learning classification models:

- Generative Models (GMM): These models estimate the probability distribution of speech features for each speaker.
- Discriminative Models (SVM): These models find the optimal boundary between different speakers' feature distributions.
- Hybrid Models (GMM-UBM): These models combine the strengths of generative and discriminative approaches.

These algorithms were specifically selected because of their effectiveness in previous speaker identification studies such as [8–10] of various languages. In addition, these algorithms are adaptable to Arabic phonetic complexities. Moreover, these algorithms can handle the complex distributions of Arabic speech data. Therefore, these classifications help to ensure robust performance across several Arabic dialects.

We comprehensively compared these models using real-world datasets to identify the most effective approach for Arabic speaker identification. The models were assessed based on their capacity to recognize speakers from a variety of unknown speech utterances recorded under different conditions. The major contributions of this paper are:

- Dataset Construction: We constructed four subsets of a dataset comprising Arabic phonemes spoken by 104 speakers. Each subset increases in complexity, with one to four phonemes per record.
- Model Evaluation: We evaluated the performance of three machine-learning models, GMM (generative), SVM (discriminative), and GMM-UBM (hybrid), using 20 and 40 MFCC feature sets.
- Model Recommendations: We recommended the most appropriate machine learning algorithms for speaker identification based on various MFCC feature sets.
- Phoneme Analysis: We investigated the smallest number of phonemes needed to accurately recognize an Arabic speaker.

The rest of the paper is organized as follows: Section II reviews related work, Section III explains the materials and methods, Section IV evaluates the performance and discusses the results, and Section V concludes this paper with the findings.

II. Literature Review

This section provides an overview of audio feature extraction methods for speaker identification using machine learning. It also includes a review of recently proposed systems for Arabic speaker identification. Analyzing audio feature extraction methods is crucial for speaker identification tasks as it directly influences the system's accuracy and reliability. Hence, by evaluating various feature extraction techniques, researchers can find out the features that are good enough to capture the fundamental and unique characteristics of the speaker to

ensure perfect modeling of unique voice attributes. In addition, being updated on recent advancements in the field of speaker identification is essential for researchers as it provides a clear understanding of the emerging trends and new techniques and methods to identify potential improvements and the current challenges in speaker identification.

A. Feature Extraction Methods

According to Tirumala [2], in any speaker identification system, the feature extraction methods can be classified into Mel Frequency Cepstral Coefficients (MFCC)-based methods and non-MFCC-based methods. Moreover, the researchers reported that the speaker identification systems that apply MFCC-based methods such as Pure MFCC, MFCC Fusion, and Linear Prediction achieve better results as compared to the systems that depend on non-MFCC-based methods. However, the performance of the MFCC-based speaker identification systems can be degraded, particularly when the speech signal is contaminated with background noise.

1) MFCC-based methods

Chowdhury *et al.* [11] addressed the problem of speaker identification using degraded speech signals by combining two feature extraction methods: MFCC and Linear Predictive Coding (LPC). In addition, they utilized a 1D Triplet Convolutional Neural Network (1D-Triplet-CNN) in order to combine both of these feature extraction methods in a novel way which enhanced the performance of the speaker recognition process in some challenging conditions. Also, Ruirui *et al.* [12] proposed a text-independent speaker identification system that can learn from limited training data by utilizing few-shot learning. Furthermore, an adversarial learning process is performed to improve the robustness and generalization of the speaker identification model. The proposed system achieved significant improvements in performance as compared to some baseline methods. Moreover, Ashar *et al.* [13] proposed a novel text-independent hybrid architecture based on CNN and MFCC methods that can be used to identify a speaker even in a noisy environment. The system improved the performance with an accuracy of 87.5%. Furthermore, Samia *et al.* [14] proposed a text-independent speaker recognition system that is robust to degradation effects such as reverberation and environmental noise. The MFCCs method along with the spectrum and log-spectrum were all used for the process of feature extraction. After that, the Long-Short Term Memory Recurrent Neural Network (LSTM-RNN) classifier is used as a classification technique to complete the task of speaker recognition. The overall recognition rate of the proposed system reached about 95.33% when using the MFCC method, and 98.7% when using the log-spectrum method. Hourri *et al.* [15] proposed a novel speaker recognition method that is based on Deep Neural Networks (DNNs). The main goal of this system is to enhance the extracted MFCC feature vectors in an unsupervised manner. These enhanced features are denoted as Deep Speaker Features (DeepSFs). In addition, this proposed method outperformed the i-vector/PLDA and some baseline systems in noisy and clean environments. Al-Qaderi *et al.* [16] proposed a speaker identification architecture that contains a generative model called the Gaussian Mixture Model (GMM) and a discriminative model called the Support Vector Machine (SVM). The proposed architecture is composed of two stages: the first model exploits the prosodic features. In addition, the second classifier contains two kinds of short-term features which are MFCC and Gammatone Frequency Cepstral Coefficients (GFCC).

2) Non-MFCC-based methods

Zhang *et al.* [17] proposed some weighting algorithms for calculating the statistics of Baum-Welch during i-vector extraction. The weights are incorporated for the computation of mean vectors, covariance matrices, and posterior probabilities of Gaussian Mixture Models (GMM) which generate new updating rules. The conducted experiments demonstrated that the proposed technique has significantly enhanced the performance of speaker recognition systems that are based on the i-vector method and operate in noisy conditions. As compared to the GMM i-vector baseline, the equal error rate of this algorithm has reduced from 5.75 to 4.72. Shihab *et al.* [18] proposed a speaker identification that is based on a hybrid Gated Recurrent Unit (GRU) and CNN feature extraction method. The GRU and CNN model is used to process the spectrogram, and then transform some part of it to optimize the extraction of the best feature vector. Finally, the most significant features are combined and selected by using a statistical method.

Jia *et al.* [19] proposed a self-organizing feature map SOM (AC-SOM)-based adaptive clustering algorithm. This algorithm is used to set the number of neurons that are present in the competition layer based on the number of speakers that the system has to recognize until the number of clusters matches the number of speakers. Experiments have demonstrated that the suggested method can greatly improve both training and recognition speed without having a major impact on recognition rate.

Dhakal *et al.* [20] presented a new pipelined real-time speaker identification architecture that improves the performance of speaker recognition by using hybrid feature extraction techniques which include CNN, Gabor Filter (GF), and statistical parameters. For the classification process, standard classifiers such as Random Forest (RF), Support Vector Machine (SVM), and DNN were investigated. Feng *et al.* [21] proposed a DNN model based on a 2-D CNN and a Gated Recurrent Unit (GRU) for speaker identification. In the proposed model, the convolutional layer was used to extract the voiceprint feature and reduce dimensionality, enabling faster computations at the GRU layer. Chauhan *et al.* [22] proposed an interesting optimization technique for feature extraction that utilizes various dimensionality reduction methods. The optimization is performed using Genetic Algorithms (GA) and the Marine Predator Algorithm (MPA). Their approach provided significant enhancement in the performance of speaker recognition across multiple classification methods.

B. Recent Arabic Speaker Identification Systems

Over the past few years, many Arabic speaker identification systems have been proposed; however, most of these systems are text-dependent-based systems. Shahin *et al.* [23] proposed a novel cascaded Gaussian Mixture Model-Deep Neural Network (GMM-DNN) classifier. The proposed text-independent speaker identification system improved the performance at several emotions. The experiments demonstrated that the proposed system outperformed the other conventional classifiers, such as Support Vector Machine (SVM) and Multilayer Perceptron (MLP). However, the performance of the proposed classifier needs further enhancements, particularly in noisy and hostile speaking environments. Also, the author in [24] proposed a speech-based authentication system that can recognize the identity of Arabic speakers based on the Arabic word (شكراً). The MFCC method was used to extract biometric features, and the classification process involved feature selection and Radial Basis Function Neural Network (RBFNN). The proposed system achieved 97.5% accuracy during the speaker identification phase. Meftah *et al.* [25] proposed a speaker recognition system taking into consideration the following three states: neutral, emotional, and without considering the state for both Arabic and English languages. The proposed system has high overall accuracy, as it can recognize speakers regardless of their language. However, it was found that type of emotion affects the accuracy of the system. Rawia *et al.* [3] proposed a speaker identification system consisting of two phases: the training phase and the testing phase. The features were extracted by mixing the MFCC, voice frequency, and voice fundamental frequency. In addition, many classification techniques such as Sequential Minimum Optimization, K-nearest neighbors and Logistic Model Tree were used during the classification process. The K-nearest neighbors method outperformed the other methods as it the highest precision which is approximately 94.8%. In reference [26], the author proposed a text-dependent speaker recognition system which can identify a speaker based on the shortest utterance, vowels, and consonants of Arabic language. The experiments demonstrated that for any Arabic text, the utterance that includes the combination of (ع ل م) will produce a high-performance speaker identification system. Roumiassa *et al.* [26] presented a text-dependent speaker identification system for both Arabic and Berber languages. They utilized various feature extraction and reduction methods including MFCC, principal component analysis (PCA), and LPC. The experiments showed that the system achieved a high recognition rate for speaker verification and identification. Our proposed model differs from the above systems. In Ref. [23], the performance of the proposed classifier needs further enhancements, especially in noisy and angry talking environments whereas our classification model performs well even in noisy environments. In addition, our proposed model doesn't require the speaker to speak a specific word to be identified as in [26, 27]. Most previous research studies have focused on methods utilizing complete sentences with limited research studies on the efficacy of phoneme-based speaker identification. Recent studies such as [3] and [22] have addressed the challenges of achieving high accuracy with long speech samples. However, this study aims to fill a crucial gap in the literature by investigating the performance of speaker identification using Arabic phonemes. Therefore, this research is expected to develop more accurate voice-based technologies and optimize user authentication processes in Arabic-speaking countries.

Our identification system can reliably recognize speakers regardless of the phonemes they speak. Moreover, the performance of the proposed system in [25] is affected by the emotions of the speakers. Furthermore, our model doesn't require extensive computations for feature extraction as in [3]. Instead, we only need to compute MFCC to reliably identify the speakers.

III. Materials and Methods

In the field of machine learning, supervised learning can be thought of as an approach where the algorithm learns and gains insights from labelled data. This allows the algorithm to make predictions for new and unseen data [28]. This methodology has been employed in this research to facilitate the ability of the algorithm to learn from the provided dataset. In this study, we evaluate and examine several machine learning models to determine which one performs the best for reliable speaker identification based on Arabic phonemes. The upcoming subsections will explain the details of the evaluation process.

A. Dataset Construction

The main aim of this study is to perform an auditory classification of Arabic speakers; therefore, the system being evaluated is required to learn and acquire knowledge from audio data. Hence, four different versions of an audio dataset are constructed of 104 speakers who speak Arabic phonemes[1]. Each of these versions of the dataset contains a varying number of phonemes per audio record of each speaker. This means that each dataset contains 104 different classes that represent 104 adult male Arabic speakers aged between 7 and 40 years. Each version of the dataset contains 104 folders representing 104 different speakers. The first version of the dataset contains approximately 29 audio files per speaker. In addition, the second version of the dataset contains 14 audio files per speaker, whereas the third version contains 11 audio files per speaker and the fourth version contains 7 audio files per speaker.

In order to construct these four different versions of the original dataset, a number of steps were followed as shown in Fig. 1. These steps began by collecting audio records of different speakers speaking Arabic alphabets/ phonemes regardless of the presence of background noise. The audio data of Arabic phonemes was collected from speakers

[1] https://drive.google.com/file/d/1yir4MLmTDlUgFMg8614mdOON5igylISm/view?usp=sharing

among different Arabic countries. In fact, there are 29 Arabic alphabets including Hamza (ء) [29]. In our case, each speaker recorded his voice speaking all Arabic alphabets in a single .wav file. After that, we followed a systematic and organized procedure of preparing the audio files, such as splitting the audio files to contain one, two, three, or four phonemes per record using an online tool that doesn't degrade the quality of the audio file. Then we converted all the resulting files to .wav extension and organized the datasets to have the same structure so that we have all the resulting datasets with the same settings.

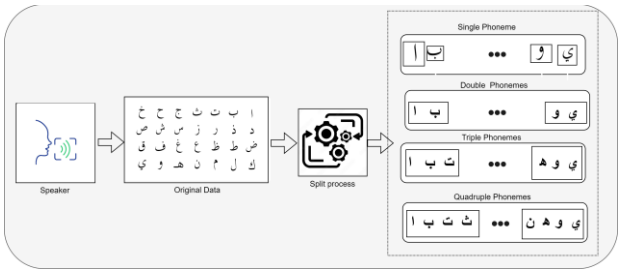

Fig. 1. Construction of first, second, third, and fourth versions of the dataset.

For the construction of our initial dataset version, we segmented each speaker's .wav file into subfiles, with each containing one alphabet or phoneme. For splitting the files, we used an online tool [30] to perfectly split the files without affecting the quality of the original file. As a result, there are approximately 29 files per speaker as presented in Fig. 2 where $N = 29$ in this version. However, to improve the accuracy of the identification process, we thought of increasing the number of alphabets/phonemes per record. As for the second version of our dataset, for every speaker, we segmented the whole .wav file into subfiles, with each containing two alphabets or phonemes. As a result, we have approximately 14 files per speaker as shown in Fig. 2 where $N = 14$ in this version. Moreover, the third version of the dataset consists of audio files containing three alphabets. For every speaker, we segmented the whole .wav file into subfiles, with each containing three alphabets or phonemes. As a result, we have approximately 11 files per speaker as illustrated in Fig. 2 where $N = 11$ in this version. Furthermore, the fourth version of the dataset consists of audio files containing four alphabets. For every speaker, we divided the whole .wav file into subfiles, with each containing four alphabets or phonemes. As a result, we have approximately 7 files per speaker as presented in Fig. 2 where $N = 7$ in this version. The data that support the findings of this study are available from the corresponding author, S.R.A., upon reasonable request.

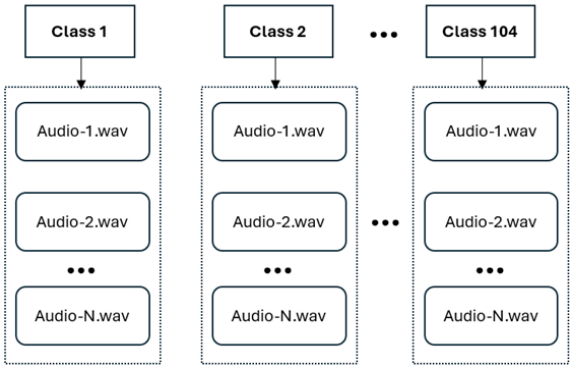

Fig. 2. Structure of the dataset.

B. System Design and Development

The technological architecture of our system comprises both powerful hardware and software components. Regarding the hardware components, we utilize a high-performance computer equipped with large storage components to manage the datasets efficiently. Our software architecture is based on Windows 11 operating system, and Python is the primary programming language used for developing machine learning algorithms for speech data processing. This setup allows for robust speaker identification performance and efficiency.

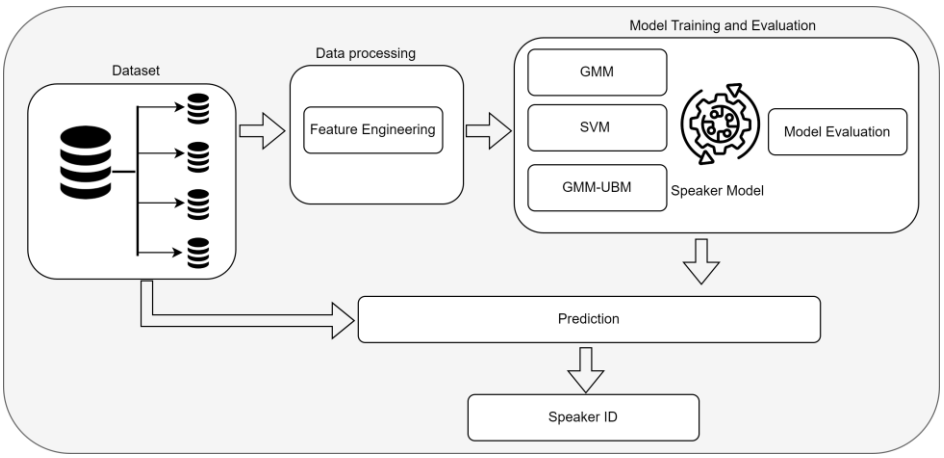

Fig. 3. Architecture of the speaker identification system.

Furthermore, the proposed speaker identification system follows a general framework which consists of two main phases, namely, the training/enrolment phase and the testing/identification phase. During the enrolment phase, the classification system undergoes training using provided labelled speech examples. In contrast, the testing

phase involves presenting an unknown speech sample to the classification system, which then identifies or verifies the identity of the speaker. For the speaker identification system to perform the above-mentioned phases, the system is made up of two main components: feature extraction (or front-end processing) and speaker modeling (or back-end processing), as illustrated in Fig. 3. The following subsection will provide a detailed explanation of each component of the system.

1) Feature extraction

In any speaker identification system, feature extraction plays a crucial role as it is required to transform the input speech signal into a relevant feature vector to be used for the modelling process. There are two main categories of feature extraction techniques: modelling human voice production and modelling peripheral auditory hearing. In the former, linear prediction cepstral coefficients (LPCC) are commonly used, while in the latter, Mel frequency cepstral coefficients (MFCC) and relative spectral perceptual linear predictive coefficients (RASTA-PLP) are popularly used [31]. MFCC is widely adopted as an outstanding speech feature. In our study, we employed MFCC and Delta-Mel Frequency Cepstral Coefficients (ΔMFCC) for their common usage and effectiveness in speech processing tasks.

a) Mel Frequency Cepstral Coefficients (MFCC)

In fact, human perception of sound frequency doesn't align with a linear scale. However, the Mel scale, which is designed to approximate human auditory perception, exhibits linearity up to 1 KHz and becomes logarithmic beyond that [32]. The process of MFCC is like the process of Real Cepstral Coefficients (RCC) in which the power spectrum undergoes filtering through a non-uniform space of frequency band. Therefore, the logarithm of the signal is calculated. Moreover, these frequency filters which are known as Mel frequencies, mirror the logarithmic distribution which reflects the human ear's hearing characteristics. The relationship between the logarithmic frequency distribution and the actual frequency is explained in the below equation:

$$\text{Mel(f)} = \frac{1000}{\ln\left(1+\frac{10}{7}\right)} \times \ln\left(1+\frac{f}{700}\right) \quad (1)$$

As illustrated in the above equation, f is calculated in Hz. In order to derive the Mel Spectral coefficients as per Eq. (1), the following steps are undertaken: First, perform the Fourier transform on each window frame of the signal. Second, map the powers of the spectrum onto the Mel scale by using triangular overlapping windows in each frame. Third, calculate the logarithms of the powers at each of the Mel frequencies. Fourth, compute the discrete cosine transform (DCT) of the list of Mel log powers. Finally, the MFCCs are represented by the amplitudes of the resulting spectrum.

b) Delta-Mel Frequency Cepstral Coefficients (ΔMFCC)

In fact, the human ear is responsive to both the static and dynamic attributes of a signal and the MFCC features primarily capture the static characteristics of the audio signal. To address this issue, the first-order derivative of the MFCC, which is known as ΔMFCC, is added to the traditional MFCC to include the dynamic information in every frame. Therefore, this combination improves the reliability and robustness of MFCC [33]. For further enhancements of identification accuracy, the extraction of second and third-order derivatives can also be employed to capture additional characteristics of the signal. The equation of ΔMFCC is presented below.

$$\Delta\text{MFCC}(m,n) = \frac{1}{\sqrt{\sum_{-k}^{k} i^2}} \sum_{i=-k}^{k} (\text{i} \times MFCC(m,n+i)) \quad (2)$$

2) Classification methods

In the classification phase of the system, a classification model is built and created using the training data and the required machine learning algorithms. Subsequently, the resulting model is utilized to predict and identify the speakers' identities in unknown utterances. This implementation of various machine learning algorithms contributes to the process of performance evaluation and comparison of automatic speaker identification systems. The machine learning algorithms used for classification process in this study speaker identification can be divided into three categories: Generative models, such as Gaussian Mixture Model (GMM), Discriminative Models, such as Support Vector Machine (SVM), and Hybrid models, such as (GMM-UBM) [34]. To cover all the above-mentioned types of classification algorithms, each of these models has been employed and evaluated in the proposed system. The following subsections explain each of these models in detail.

a) Gaussian Mixture Model (GMM)

In general, GMM can be viewed as a machine learning method for modelling multivariate probability distributions which performs well, specifically at handling arbitrary distributions. In addition, it stands out as one of the most common and popular techniques employed for the classification process in speaker identification systems. Furthermore, for speaker "s", the GMM representing the distribution of feature vectors is constructed as a weighted linear combination of M unimodal Gaussian densities [35], where each Gaussian density is characterized by mean vectors (μ_i^s) and a covariance matrix ($\sum_i^s$). These parameters together form a speaker model which is represented by the notation $\{ p_i^s, \mu_i^s \sum_i^s \}_{i=1}^M$. Here, p_i^s corresponds to the mixture weights with respect to the stochastic constraint $\sum_{i=1}^M p_i^s$ =1. When considering a feature vector x, the density mixture for the speakers is then determined as:

$$P(\text{x}|y_s) = \sum_{i=1}^{M} p_i^s b_i^s(x) \quad (3)$$

where,

$$b_i^s(x) = \frac{1}{(2\pi)^{D/2} |\sum_i^s|^{1/2}} e^{\left(-\frac{1}{2}(x-\mu_i^s)^t (\sum_i^s)^{-1}(x-\mu_i^s)\right)} \quad (4)$$

and where D refers to the dimensions of the feature vector.

b) Support Vector Machine (SVM)

Over the past years, SVM has been widely implemented as a primary algorithm for developing speaker identification systems and classifying the speakers in

unknown utterances [36]. During the training process of the model, SVM constructs and builds a binary class classifier based on the convex optimization theory. In addition, it is required to use an optimum hyperplane which expands a decision boundary between the two target classes [37]; however, the recent developments of SVM implement a structured risk minimization model as shown in the following equation.

$$\min_{w,\aleph_k} j_1(w, \aleph_k) = \frac{1}{2} w^{Д}\, w + c\ \sum_{i=1}^{N} \aleph_k \tag{5}$$

The linear method of SVM works by projecting an input feature vector x into a function $f(x)$ as shown in the following:

$$f(x) = sign[\sum_{i=1}^{N} a_i k(x; x_i) + \beta] \tag{6}$$

where x_i represents the support vectors, β refers to a bias, a_i represents the weights and N is the total number of the support vectors of SVM.

c) Gaussian Mixture Model-Universal Background Model (GMM-UBM)

In fact, UBM in the speaker identification system is an important Gaussian Mixture Model that is trained to represent and model the speaker-independent distribution of features [38]. In order to determine the parameters mean, variance, and weight of the UBM model, the Maximum Likelihood Linear Regression (MLLR) algorithm [39] is utilized to combine the data from all utterances of the speakers. After that, the hypothesized speaker-specific model is formed by adapting the UBM parameters using the speaker's training speech samples and employing Bayesian Maximum A Posteriori (MAP) adaptation instead of MLLR training. To identify a speaker from a given set of speakers $S = \{1, 2, ..., s\}$, each speaker is characterized by a model derived from the UBM model through MAP adaptation. In addition, the MAP adapted models for S speakers are represented as $\alpha 1, \alpha 2, \ldots, \alpha s$. Also, the speaker model that maximizes the a posteriori probability for a given sequence of speech utterances can be expressed as:

$$\hat{S} = \arg\_max_{1 \le k \le s} \frac{p(X\,|\alpha_k) p(\alpha_{UBM})}{p(X\,|\alpha_{UBM})} \tag{7}$$

3) Evaluation metrics

The evaluation metrics are utilized to assess the effectiveness of trained models and their accuracy on a specific dataset. In fact, the most widely used performance measures in the speaker identification field are accuracy, precision, recall, and F-measure [34]. In this study, we evaluated the trained ML models using metrics such as accuracy, precision, recall, and f1-score. Furthermore, we conducted a comparative analysis of results from the three trained classification models to identify the most effective model for our dataset. In the below subsections, we briefly explain and elaborate on each metric and its corresponding equation.

- Accuracy

Accuracy is considered as a widely adopted performance measure which is simply used to describe the count of accurately classified instances by a specific classification algorithm [40]. The mathematical representation and equation of the accuracy is illustrated by the following equation:

$$\text{Accuracy} = \frac{\text{Number of Correctly Predicted Instances}}{\text{Total Number of Instances}} \tag{8}$$

By measuring the systems' accuracy, researchers and developers can then gain insights into the system's overall performance which enables them to fine-tune the used algorithms and address potential weaknesses of the system.

- Precision

Precision refers to the calculation of the ratio of true positive instances to the sum of both true positive and false positive instances [41]. The mathematical formula that represents the precision is shown in the following equation:

$$\text{Precision} = \frac{TP}{(TP+FP)} \tag{9}$$

By prioritizing precision metrics, developers can gain insights into the system's ability to precisely classify and identify speakers, especially in scenarios where false positives can lead to serious effects, such as in security or authentication-related scenarios.

- Recall

In fact, recall is also known as True Positive Rate (TPR) or Sensitivity. It measures the ability of the ML classifier to correctly predict and identify positive instances[41]. This metric indicates how well the model captures and predicts all the actual positive cases. The mathematical formula of the recall is represented by the following equation:

$$\text{Recall} = \frac{TP}{(FN+TP)} \tag{10}$$

Recall is calculated to evaluate the system's proficiency in identifying all speakers who should be identified as positive regardless of any false negatives. In addition, recall has been proved to be crucial, especially in scenarios where missing a positive identification could lead to significant impacts, specifically in security or forensic contexts.

- F1-Score

The F1-score metric refers to the calculation of the harmonic mean of precision and recall which provides a balance between both of these metrics[41]. In fact, a higher value of F1-score indicates a better balance between precision and recall, representing the model's overall performance in predicting positive instances while minimizing both false positives and false negatives. Moreover, the mathematical formula of the F1-Score is calculated using the following equation:

$$\text{F1} - \text{Score} = 2\,\frac{(\text{Precesion} \times \text{Recall})}{(\text{Precesion} + \text{Recall})} \tag{11}$$

As a matter of fact, measuring the F1-score is important to ensure the reliability and accuracy of the speaker identification system in real-world applications and across various scenarios and datasets.

II. Results and Discussion

The details of the experiments performed using the three classification methods mentioned in the previous section are presented in the following subsequent sections. It begins by explaining the experimental configuration, followed by an illustration of the experiment's outcomes and results of each of the trained ML models.

A. Experimental Setup

In our experiments, we divided the data as follows: about 70% of the data for training and 30% for testing. In addition, we test each of the classification models, GMM, SVM, and GMM-UBM on the same dataset using the same training and testing data. This uniform setting allows us to fairly compare the performance and results of each model in our experiment. The details of the experiment of each classification model are explained in the following subsections.

For the proposed architecture, we trained our three models which are GMM, SVM, and GMM-UBM on the same dataset (first, second, third, fourth versions) of the original Arabic dataset. We set the number of mixtures of the GMM model and GMM-UBM model to 16 for all the GMM-related experiments. In addition, we set the Regularization Parameter C of the SVM model to 1 for all the SVM-related experiments. Moreover, for each of the three classification models, we conducted model training twice, once on a 20 MFCC feature set and another time using a 40 MFCC feature set by combining 20 MFCCs with 20 delta MFCCs. Therefore, for every model GMM, SVM, and GMM-UBM, we performed eight experiments which are One-Letter-20MFCCs, One-Letter-40MFCCs, Two-Letters-20MFCCs, Two-Letters-40MFCCs, Three-Letters-20MFCCs, Three-Letters-40MFCCs, Four-Letters-20MFCCs, and Four-Letters-40MFCCs.

B. Experimental Results

In our speaker identification experiments, we started by conducting experiments related to the 20 MFCCs of our three classification models: GMM, SVM, and GMM-UBM. After training these models, we evaluated their performance with several evaluation measurements which are accuracy, precision, recall, and F1-score. Moreover, Fig. 4 presents the accuracy results of the 20 MFCCs-based experiments.

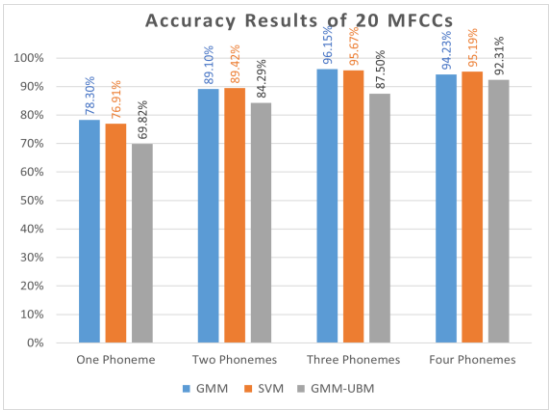

Fig. 4. Accuracy results of 20 MFCCs.

In addition, Fig. 5 demonstrates the precision results of the 20 MFCCs-based experiments. Furthermore, Fig. 6 illustrates the recall results of the 20 MFCCs-based experiments. Additionally, Fig. 7 presents the F1-score results of the 20 MFCCs-based experiments.

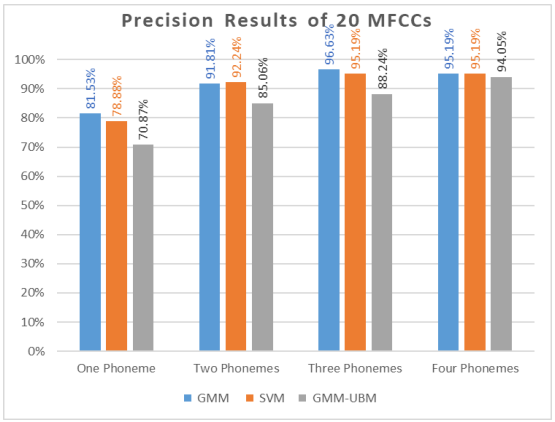

Fig. 5. Precision results of 20 MFCCs.

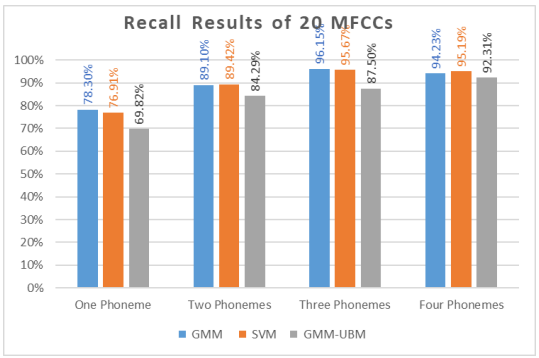

Fig. 6. Recall results of 20 MFCCs.

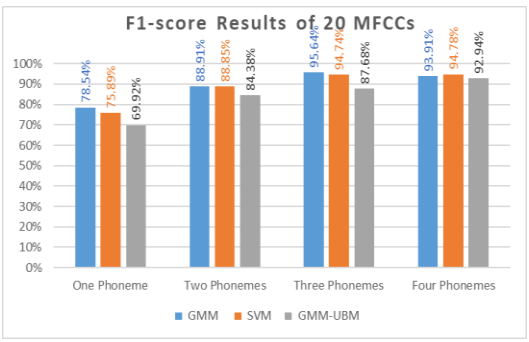

Fig. 7. F1-score results of 20 MFCCs.

The results of one phoneme-based experiment demonstrate that the GMM model achieved the highest accuracy of 78.30%, followed closely by the SVM model with an accuracy of 76.91%. In addition, GMM-UBM-based model achieved the lowest accuracy of 69.82%. Similarly, for two and three phonemes-based experiments, both GMM and SVM models performed comparably well; nevertheless, the SVM model slightly outperformed the GMM in some cases. However, during the four phonemes-based experiments, all models achieved high values of accuracy. However, the SVM and GMM-UBM models perform better than GMM. Moreover, across all

experiments of different numbers of phonemes, the GMM model consistently achieved higher precision, recall, and F1-score as compared to SVM and GMM-UBM. This not only implies higher performance and accuracy, but also indicates that the GMM model maintained a better balance between precision and recall across all experiments. However, the performance of the SVM and GMM-UBM models varied slightly depending on the number of phonemes per record. Overall, experiments with SVM showed robust performance, closely matching the accuracy of the GMM model, especially for experiments involving two or more phonemes. On the other hand, GMM-UBM showed competitive performance, especially in cases with fewer phonemes. This suggests that its accuracy decreased as the task's complexity increased.

After conducting experiments using 20 MFCCs for evaluating the performance of speaker identification using the three ML models, we proceeded to evaluate the performance of these models on a higher-dimensional feature set consisting of 40 MFCCs. The results of these experiments are summarized in Fig. 8, Fig. 9, Fig. 10, and Fig. 11.

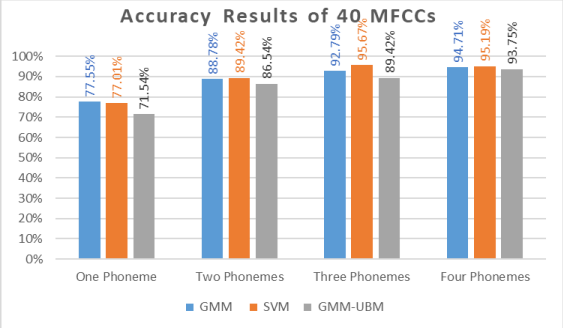

Fig. 8. Accuracy results of 40 MFCCs.

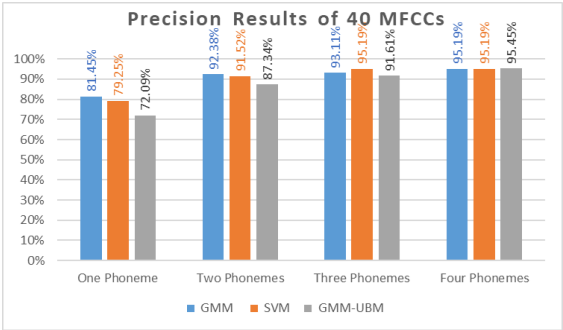

Fig. 9. Precision results of 40 MFCCs.

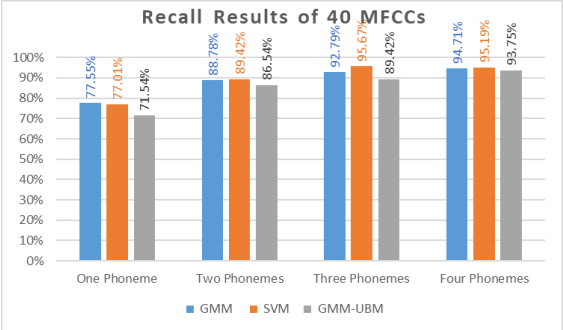

Fig. 10. Recall results of 40 MFCCs.

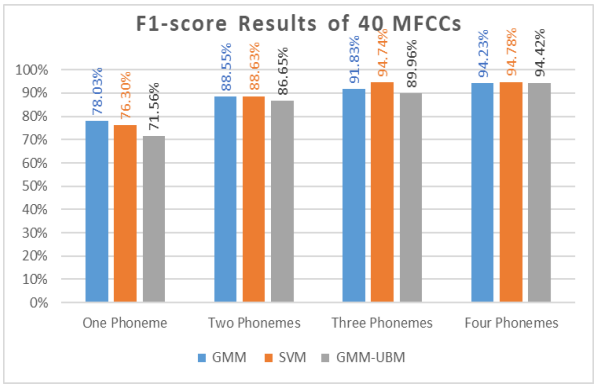

Fig. 11. F1-score results of 40 MFCCs.

As shown in the above figures, the SVM model consistently performs slightly better than the GMM model or same in some cases across all phoneme categories from the accuracy point of view. However, the GMM-UBM model exhibits the lowest accuracy as compared to the other two models. This implies that incorporating the universal background model with the GMM model might not have significant improvements on the performance in this scenario. Similarly, when considering the precision, we observe a similar pattern where the SVM model tends to achieve higher values of precision as compared to the GMM model across all different numbers of phonemes. However, the GMM-UBM model exhibits the lowest precision as compared to the other two models other which indicates a higher rate of false positives. Moreover, in terms of recall, the SVM model also consistently outperforms the GMM model. In addition, the GMM-UBM model shows lower recall values as compared to the other two models, suggesting a higher rate of false negatives. Furthermore, for F1-score evaluation metric, the SVM model generally achieves competitive F1-scores as compared to the GMM model across all phoneme categories. In addition, even though the GMM-UBM model shows improvements over the GMM model in some cases, it is not better than the SVM model in terms of F1-score.

A summary of the numerical results from the experiments based on 20 MFCCs is outlined in Table I. In addition, the outcomes of the experiments using 40 MFCCs are presented in Table II. Examination of the tables reveals that the GMM model outperformed the SVM and GMM-UBM models for the 20 MFCCs dataset. On the other hand, for the 40 MFCCs feature set, the SVM model exhibited superior performance across all speaker identification cases and tasks.

C. Discussion

As shown in Table I and Table II, the minimum number of Arabic phonemes required for accurate and reliable Arabic speaker identification is three. Table I illustrates that for the 20 MFCC feature set, the GMM-based model exhibits remarkable performance with an accuracy of 96.2%, precision of 96.6%, recall of 96.2%, and F1-score of 95.7%. One reason for this success could be the suitability of GMM for a lower number of MFCCs, which is 20 in this case, as it is considered a statistical model. On

the other hand, Table II demonstrates that the SVM-based model is better for the 40 MFCC feature set as it achieved an accuracy of 95.67%, precision of 95.19%, recall of 95.67%, and F1-score of 94.74%. The strength of SVM models, especially in higher-dimensional feature spaces (such as 40 MFCC in this case), lies in their ability to capture complex decision boundaries. This makes them advantageous for handling complex patterns in speaker audio data. Furthermore, the consistent and linear correlation of the evaluation metrics across different cases, including varying recording conditions, underscores and ensures the robustness of the trained ML models for speaker identification tasks.

TABLE I. RESULTS OF 20 MFCC FEATURE SET

Classification Method	Generative Models (GMM)				Discriminative Models (SVM)				Hybrid Models (GMM-UBM)			
# of letters per record	1	2	3	4	1	2	3	4	1	2	3	4
Accuracy	78.3%	89.1%	**96.2%**	94.2%	76.9%	89.4%	95.7%	95.1%	69.8%	84.3%	87.5%	92.3%
Precision	81.5%	91.8%	**96.6%**	95.1%	78.9%	92.2%	95.1%	95.1%	70.9%	85.1%	88.2%	94.1%
Recall	78.3%	89.1%	**96.2%**	94.2%	76.9%	89.4%	95.7%	95.1%	69.8%	84.3%	87.5%	92.3%
F1-score	78.5%	88.9%	**95.7%**	93.9%	75.9%	88.9%	94.7%	94.8%	69.9%	84.4%	87.7%	92.9%

TABLE II. RESULTS OF 40 MFCC FEATURE SET

Classification Method	Generative Models (GMM)				Discriminative Models (SVM)				Hybrid Models (GMM-UBM)			
# of letters per record	1	2	3	4	1	2	3	4	1	2	3	4
Accuracy	77.6%	88.8%	92.8%	94.7%	77%	89.4%	**95.7%**	95.2%	71.5%	86.5%	89.4%	93.8%
Precision	81.5%	92.4%	93.1%	95.2%	79.3%	91.5%	**95.2%**	95.2%	72.1%	87.3%	91.6%	95.5%
Recall	77.6%	88.8%	92.8%	94.7%	77%	89.4%	**95.7%**	95.2%	71.5%	86.5%	89.4%	93.8%
F1-score	78%	88.6%	91.8%	94.2%	76.3%	88.6%	**94.7%**	94.8%	71.6%	86.7%	90%	94.4%

Even though all three models provide good results and can be used for speaker identification using 20 MFCCs, the GMM model outperforms the other two models with consistently high accuracy and robust precision-recall performance across all experiments based on varying number of phonemes. Nevertheless, in some scenarios where there are specific requirements and computational limitations, both SVM and GMM-UBM models could be used as alternatives, especially in cases involving multiple phonemes. Moreover, additional experimentation and fine-tuning might be required to optimize the performance of each model, specifically for certain use cases and datasets. Furthermore, the SVM model shows slightly better performance as compared to the GMM model for speaker identification using 40 MFCCs. In addition, the GMM-UBM model does not provide significant enhancements of the performance over the GMM model in the context of speaker identification.

Upon evaluating the performance of trained ML models for speaker identification against existing research [8–10, 42, 43], we observed notable similarities and differences. Our results show that GMM-based models outperformed SVM-based models when using a 20 MFCC feature set as it achieved an accuracy of 96.2%. This aligns with the findings from previous studies where GMMs proved to be effective in similar contexts. However, for the 40 MFCC feature set, our SVM-based models achieved the highest accuracy of 95.67% which proved the robustness of SVMs in handling higher-dimensional feature spaces. Furthermore, while our findings prove the effectiveness of these classification models in Arabic speaker identification, further analysis is required to understand the behavior and complexity of the models and then optimize them for use in a wider range of scenarios.

The studies mentioned above share similarities such as using GMM or SVM as classification models in addition to utilizing MFCCs features for speaker identification tasks. When reviewing the literature, we observed a significant gap in research focused on phoneme-based speaker identification for the Arabic language. When comparing our proposed approach to those discussed in [44, 45], our approach demonstrates lower computational complexity. In addition, the proposed study achieved notable accuracies without the need for removing background noise which proves the robustness and feasibility of our methodology in real-world scenarios. Furthermore, our research addresses challenges encountered in everyday environments, ensuring the applicability and reliability of our speaker identification system. Therefore, the contribution of our research highlights the effectiveness of our feature extraction and modelling methods in extracting unique speaker characteristics, regardless of environmental variability, which help in developing more robust speaker identification systems. Moreover, this study can also help in related domains, such as speech recognition and natural language processing, which rely on similar audio data analysis principles.

Furthermore, when comparing our results with recent Arabic speaker identification research [3], our study shows better performance. Previous research recommended the use of KNN model in addition to complete sentences to reliably identify speakers, which achieved a precision of 94.8%. However, our findings emphasized that using short utterances or even Arabic phonemes (3 phonemes) instead of full sentences, which makes the system more complex, is sufficient for accurate speaker identification. Furthermore, our study proves that GMM or SVM models outperform KNN based on our experiments. Details of the KNN model's performance are presented in Table III.

One may argue that we should try and compare our results with more advanced classification schemes such as High-level CNNs [46] or the transformer model [47]. The justification is that a 20 or 40 feature set is not a very complex or large problem set, so using the advanced schemes will be overkill.

TABLE III. RESULTS OF APPLYING KNN USING ARABIC PHONEMES

Classification Method	K-Nearest Neighbors (KNN) using 20 MFCCs				K-Nearest Neighbors (KNN) using 40 MFCCs			
# of letters per record	1	2	3	4	1	2	3	4
Accuracy	62.19%	76.92%	87.98%	92.79%	61.65%	76.92%	87.5 %	91.82%
Precision	70.62%	79.62%	86.92%	91.83%	72.13%	80.66%	87.72%	91.57%
Recall	62.19%	76.92%	87.98%	92.79%	61.65%	76.92%	87.5 %	91.83%
F1-score	60.51%	75.01%	85.65%	91.28%	60.06%	75.09%	85.65%	90.42%

III. CONCLUSION

Speaker identification is an important authentication technique. This research investigates the performance of Arabic phonemes for speaker identification. The research provides a comprehensive comparison of the performance of three machine learning models used for Arabic speaker identification based on phonemes. A crucial part of this study involves investigating the feasibility of using short audio samples such as Arabic phonemes or alphabets to accurately identify the speakers, regardless of the presence of background noise. The research findings show that it is necessary to have at least three Arabic phonemes for reliable speaker identification. The analysis of the results emphasizes the robustness of using Arabic phonemes for speaker identification. The findings of the analysis provide valuable insights into phoneme-based speaker identification. The examination of all the speaker identification models proves that the GMM models perform better for identifying the speakers based on 20 MFCC feature set with an accuracy of 96.2%. On the other hand, accuracy of 95.67% was achieved using SVM-based models for 40 MFCC feature set. Moreover, the GMM-UBM model does not offer significant improvements of the performance over the GMM model in this context. This conflicts with the findings of previous research which recommended the use of KNN model along with complete sentences for reliable speaker identification. In contrast, our study proves that GMM and SVM models outperform KNN model based on experimental results. Even though our trained models achieved good results, this can be improved in the future by expanding models on words or sentences. In the future, we plan to consider adding words or sentences from each speaker to our dataset. Furthermore, we plan to expand our dataset by adding more speakers from more Arabic countries with varying dialects. This research can help improve user authentication in various domains such as customer services, and voice-activated devices.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Conceptualization: H.F.A., I.R., I.F. and S.R.A; methodology, S.R.A, I.R., I.F., H.F.A; software, S.R.A.; validation, I.R., I.F. and H.F.A.; formal analysis, H.F.A and S.R.A.; investigation, I.R., I.F. S.R.A; resources, H.F.A.; data curation, S.R.A.; writing—original draft preparation, S.R.A.; writing—review and editing, I.R., I.F. and H.F.A.; visualization, S.R.A.; project administration, H.F.A; funding acquisition, S.R.A. All authors had approved the final version.

FUNDING

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Al-Ahsa, Saudi Arabia [Grant KFU242471].

ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Al-Ahsa, Saudi Arabia for funding this research [Grant KFU242471].

REFERENCES

[1] A. Phipps, K. Ouazzane, and V. Vassilev, "Enhancing cyber security using audio techniques: A public key infrastructure for sound," in *Proc. 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, Guangzhou, China: IEEE, Dec. 2020, pp. 1428–1436. doi: 10.1109/TrustCom50675.2020.00192

[2] S. S. Tirumala, "Speaker identification features extraction methods: A systematic review," *Expert Syst. Appl.*, 2017.

[3] R. A. Mohammed, N. F. Hassan, and A. E. Ali, "Arabic speaker identification system using multi features," *Eng. Technol. J.*, vol. 38, no. 5, 2020.

[4] S. Kinkiri, B. Barakat, and S. Keates, "Phonemes: An explanatory study applied to identify a speaker," in *Machine Learning, Image Processing, Network Security and Data Sciences*, A. Bhattacharjee, S. Kr. Borgohain, B. Soni, G. Verma, and X.-Z. Gao, Eds. Singapore: Springer Singapore, 2020, vol. 1241, pp. 58–68. doi: 10.1007/978-981-15-6318-8_6

[5] Resala Academy. (Jul. 28, 2024). 10 facts about the Arabic language that will surprise you. [Online]. Available: https://resala-academy.com/10-facts-about-the-arabic-language-that-will-surprise-you/

[6] G. S. Morrison *et al.*, "Interpol survey of the use of speaker identification by law enforcement agencies," *Forensic Sci. Int.*, vol. 263, pp. 92–100, Jun. 2016, doi: 10.1016/j.forsciint.2016.03.044

[7] World Research Library: Home. (2024). [Online]. Available: https://worldresearchlibrary.org/proceeding.php?pid=6683

[8] S. S. Nidhyananthan and R. S. S. Kumari, "Language and text-independent speaker identification system using GMM," *WSEAS Transactions on Signal Processing*, vol. 9, no. 4, 2013.

[9] M. S. Sinith and A. Salim, "A novel method for text-independent speaker identification using MFCC and GMM," in *Proc. 2010 International Conference on Audio, Language and Image Processing*, 2010.

[10] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Comput. Speech Lang.*, vol. 20, no. 2–3, pp. 210–229, Apr. 2006, doi: 10.1016/j.csl.2005.06.003

[11] A. Chowdhury and A. Ross, "Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio

signals," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 1616–1629, 2020, doi: 10.1109/TIFS.2019.2941773.
[12] R. Li, J.-Y. Jiang, J. Liu, C.-C. Hsieh, and W. Wang, "Automatic speaker recognition with limited data," in *Proc. the 13th International Conference on Web Search and Data Mining*, Houston TX USA: ACM, Jan. 2020, pp. 340–348. doi: 10.1145/3336191.3371802
[13] A. Ashar, M. S. Bhatti, and U. Mushtaq, "Speaker identification using a hybrid CNN-MFCC approach," in *Proc. 2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, Karachi, Pakistan: IEEE, Mar. 2020, pp. 1–4. doi: 10.1109/ICETST49965.2020.9080730
[14] S. A. El-Moneim, M. A. Nassar, M. I. Dessouky, N. A. Ismail, A. S. El-Fishawy, and F. E. A. El-Samie, "Text-independent speaker recognition using LSTM-RNN and speech enhancement," *Multimed. Tools Appl.*, vol. 79, no. 33–34, pp. 24013–24028, Sep. 2020. doi: 10.1007/s11042-019-08293-7
[15] S. Hourri and J. Kharroubi, "A deep learning approach for speaker recognition," *Int. J. Speech Technol.*, vol. 23, no. 1, pp. 123–131, Mar. 2020. doi: 10.1007/s10772-019-09665-y
[16] M. Al-Qaderi, E. Lahamer, and A. Rad, "A two-level speaker identification system via fusion of heterogeneous classifiers and complementary feature cooperation," *Sensors*, 2021.
[17] X. Zhang, X. Zou, M. Sun, T. F. Zheng, C. Jia, and Y. Wang, "Noise robust speaker recognition based on adaptive frame weighting in GMM for I-vector extraction," *IEEE Access*, 2017.
[18] S. H. Shihab, S. Aditya, J. H. Setu, K. M. Imtiaz-Ud-Din, and I. A. Efat, "A Hybrid GRU-CNN Feature Extraction Technique for Speaker Identification," in *Proc. 2020 23rd International Conference on Computer and Information Technology (ICCIT)*, 2020.
[19] Y. Jia *et al.*, "Speaker recognition based on characteristic spectrograms and an improved self-organizing feature map neural network," *Complex Intell. Syst.*, vol. 7, no. 4, pp. 1749–1757, Aug. 2021. doi: 10.1007/s40747-020-00172-1
[20] P. Dhakal, P. Damacharla, A. Javaid, and V. Devabhaktuni, "A near real-time automatic speaker recognition architecture for voice-based user interface," *Mach. Learn. Knowl. Extr.*, vol. 1, no. 1, pp. 504–520, Mar. 2019. doi: 10.3390/make1010031
[21] F. Ye and J. Yang, "A deep neural network model for speaker identification," *Appl. Sci.*, vol. 11, no. 8, 3603, Apr. 2021. doi: 10.3390/app11083603
[22] N. Chauhan, T. Isshiki, and D. Li, "Enhancing speaker recognition models with noise-resilient feature optimization strategies," *Acoustics*, 2024.
[23] I. Shahin, A. B. Nassif, and S. Hamsa, "Novel cascaded Gaussian mixture model-deep neural network classifier for speaker identification in emotional talking environments," *Neural Comput. Appl.*, vol. 32, no. 7, pp. 2575–2587, Apr. 2020. doi: 10.1007/s00521-018-3760-2
[24] A. Al-Qaisi, "Arabic word dependent speaker identification system using artificial neural network," *International Journal of Circuits, Systems and Signal Processing*, vol. 14, 2020.
[25] A. H. Meftah, H. Mathkour, S. Kerrache, and Y. A. Alotaibi, "Speaker identification in different emotional states in Arabic and English," *IEEE Access*, vol. 8, pp. 60070–60083, 2020. doi: 10.1109/ACCESS.2020.2983029
[26] F. Roumiassa and F.-Z. Chelali, "Speaker identification and verification system for Arabic and Berber language," in *Proc. 2020 1st International Conference on Communications, Control Systems and Signal Processing (CCSSP)*, EL OUED, Algeria: IEEE, May 2020, pp. 242–247. doi: 10.1109/CCSSP49278.2020.9151633
[27] A. Mahmood, "Arabic speaker recognition system based on phoneme fusion," *Multimedia Tools and Applications*, 2024.
[28] V. Nasteski, "An overview of the supervised machine learning methods," *Horizons.B*, vol. 4, pp. 51–62, Dec. 2017. doi: 10.20544/HORIZONS.B.04.1.17.P05
[29] J. D. Byrum and O. Madison, "Multi-script, multilingual, multi-character issues for the online environment," in *Proc. Workshop Sponsored by the IFLA Section on Cataloguing*, 2013.
[30] Cut audio - Cut MP3, M4A, OGG, WAV online. (Feb. 27, 2023). [Online]. Available: https://www.aconvert.com/audio/split/
[31] G. I. Sapijaszko, "An overview of recent window based feature extraction algorithms for speaker recognition," in *Proc. 2012 IEEE 55th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2012.
[32] V. Tiwari, "MFCC and its applications in speaker recognition," *International Journal on Emerging Technologies*, no. 1, pp. 19–22, Feb. 2010.
[33] M. Musaev, M. Abdullaeva, and M. Ochilov, "Advanced feature extraction method for speaker identification using a classification algorithm," presented at the II International Scientific Forum on Computer and Energy Sciences (WFCES-II 2021), Almaty, Kazakhstan, 2022, 020022. doi: 10.1063/5.0106614
[34] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Expert Syst. Appl.*, vol. 171, 114591, Jun. 2021. doi: 10.1016/j.eswa.2021.114591
[35] S. G. Bagul and R. K. Shastri, "Text independent speaker recognition system using GMM," in *Proc. 2013 International Conference on Human Computer Interactions (ICHCI)*, Chennai, India, 2013, pp. 1–5. doi: 10.1109/ICHCI-IEEE.2013.6887781
[36] S. Mohanty and B. K. Swain, "Speaker identification using SVM during Oriya Speech Recognition," *International Journal of Image, Graphics and Signal Processing*, vol. 7, pp. 28–36, 2015. doi: 10.5815/ijigsp.2015.10.04
[37] S. M. Kamruzzaman, "Speaker identification using MFCC-domain support vector machine," arXiv preprint, arXiv:1009.4972, 2010.
[38] F. R. Chowdhury, S.-A. Selouani, and D. O'Shaughnessy, "Text-independent distributed speaker identification and verification using GMM-UBM speaker models for mobile communications," in *Proc. 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010)*, 2010.
[39] A. Sarkar, J.-F. Bonastre, and D. Matrouf, "Speaker verification using M-Vector extracted from MLLR super-vector," in *Proc. 2012 the 20th European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, 2012, pp. 21–25.
[40] L. Dino. Accuracy in Python. *Medium*. (May 27, 2024). [Online]. Available: https://medium.com/@24littledino/accuracy-in-python-980074154e52
[41] sklearn.metrics. scikit-learn. (Dec. 18, 2024). [Online]. Available: https://scikit-learn/stable/api/sklearn.metrics.html
[42] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995. doi: 10.1109/89.365379
[43] G. Nijhawan and M. K. Soni, "Speaker recognition using support vector machine," *Int. J. Comput. Appl.*, vol. 87, no. 2, pp. 7–10, Feb. 2014. doi: 10.5120/15178-3379
[44] X. Wang, C. Xie, Q. Wu, H. Zhan, and Y. Wu, "A novel phoneme-based modeling for text-independent speaker identification," in *Proc. Interspeech 2022*, ISCA, Sep. 2022, pp. 4775–4779. doi: 10.21437/Interspeech.2022-617
[45] T. A. de Lima and M. C. Da-Costa Abreu, "Phoneme analysis for multiple languages with fuzzy-based speaker identification," *IET Biom.*, vol. 11, no. 6, pp. 614–624, 2022. doi: 10.1049/bme2.12078
[46] G. Costantini, V. Cesarini, and E. Brenna, "High-level CNN and machine learning methods for speaker recognition," *Sens. 2023*, vol. 23, no. 7, Mar. 2023.
[47] A. A. Khan *et al.*, "An efficient text-independent speaker identification using feature fusion and transformer model," *Comput. Mater. Contin.*, vol. 75, no. 2, pp. 4085–4100, 2023. doi: 10.32604/cmc.2023.036797