Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2022, Article ID 4980920, 16 pages
https://doi.org/10.1155/2022/4980920

Research Article

A Robust Approach for Speaker Identification Using Dialect Information

Shahid Munir Shah,[1] **Muhammad Moinuddin**,[2,3] **and Rizwan Ahmed Khan**[1]

[1] *Faculty of IT, Salim Habib University, Karachi, Pakistan*
[2] *Center of Excellence in Intelligent Engineering Systems, King Abdul Aziz University, Jeddah, Saudi Arabia*
[3] *Electrical and Computer Engineering Department, King Abdul Aziz University, Jeddah, Saudi Arabia*

Correspondence should be addressed to Shahid Munir Shah; shahidmunirshah@yahoo.com

Received 21 October 2021; Revised 6 February 2022; Accepted 14 February 2022; Published 7 March 2022

Academic Editor: Upaka Rathnayake

The present research is an effort to enhance the performance of voice processing systems, in our case the speaker identification system (SIS) by addressing the variability caused by the dialectical variations of a language. We present an effective solution to reduce dialect-related variability from voice processing systems. The proposed method minimizes the system's complexity by reducing search space during the testing process of speaker identification. The speaker is searched from the set of speakers of the identified dialect instead of all the speakers present in system training. The study is conducted on the Pashto language, and the voice data samples are collected from native Pashto speakers of specific regions of Pakistan and Afghanistan where Pashto is spoken with different dialectal variations. The task of speaker identification is achieved with the help of a novel hierarchical framework that works in two steps. In the first step, the speaker's dialect is identified. For automated dialect identification, the spectral and prosodic features have been used in conjunction with Gaussian mixture model (GMM). In the second step, the speaker is identified using a multilayer perceptron (MLP)-based speaker identification system, which gets aggregated input from the first step, i.e., dialect identification along with prosodic and spectral features. The robustness of the proposed SIS is compared with traditional state-of-the-art methods in the literature. The results show that the proposed framework is better in terms of average speaker recognition accuracy (84.5% identification accuracy) and consumes 39% less time for the identification of speaker.

1. Introduction

Voice processing systems (VPSs) such as speech, speaker, accent, and dialect and language recognition systems play a vital role in various daily life tasks. Biometric authentication (authenticating people using their voice) [1], conducting forensic tests (searching people voice in large public speech databases) [2], personalizing services based on smart devices (controlling home devices in smart home environments) [3], etc., are some of the popular applications of VPSs.

Recently, with the rapid increase in handheld devices, i.e., smartphones, tablets, and digital assistant devices at smart homes, the usage of VPSs as the user interface has also expeditiously increased [4]. One of the key reasons behind this rapid increase in the usage of VPSs as the user interface is their easy adoption mechanism and capability of operating devices in a hand-free manner [5].

Unfortunately, VPSs are vulnerable to different adversarial attacks [6–8] and contain various other performance degrading factors/variabilities such as channel mismatch (using different channels for enrollment and test data sets) [9], room or space reverberation (decay in sound intensity with time) [10], background noise, and speaker's internal variations such as language, emotions, health, and vocal efforts. [11–14]. The presence of all such variability factors limits the performance of VPSs, and hence, these systems cannot provide robust and accurate recognition [15].

Among the aforementioned variability factors, one of the most important performance degrading factor in VPSs is dialectical variations of a language [16, 17]. Due to the mismatch of the speaker's accent or dialect during the training and testing phases, the performance of VPSs significantly decreases. Accent and dialect recognition systems can decrease this mismatch by detecting the speaker's accent or dialect before the recognition phase of VPSs by adapting an appropriate model [18, 19].

Accent and dialect recognition (classification and identification) systems are not only designed as a solution to address the accent and dialect-related variability but also for other several application areas, for example, development of forensic software, targeted advertisements, and service customization[20]. Another important application of designing accent and dialect recognition systems is the identification of speakers' region of origin [21]. This application may be utilized by law enforcement agencies for the identification of suspects involved in criminal/terrorist activities. Keeping in view the above discussion and the significance of accent and dialect recognition systems, in this research, we propose a Pashto SIS framework based on the information taken from the cascaded Pashto dialect identification (DID) system. Following are the main reasons, which motivated us to select the Pashto language for our research work:

(i) It is a broadly spoken language in two Asian countries, i.e., Pakistan and Afghanistan (refer section 3 for more details).

(ii) It is one of the popular regional languages of Pakistan.

(iii) It is a very rich language in terms of accentual and dialectical variations (refer Table 1 for more details), and therefore, it is a reasonable choice to design Pashto language-based accent and dialect robust VPSs.

(iv) It is an under-resourced language in terms of text, speech, and language tools such as dictionaries, text-to-speech and speech-to-text systems, grammar and spell checker resources, and available data sets to design VPSs [22, 23]. Developing an accent and dialect-based Pashto speaker's data set and designing an efficient SIS for Pashto speakers can be a useful resource for an under-resourced language such as Pashto.

The main purpose of the research was to design a Pashto SIS framework that is robust against dialectical variations. Irrespective of the previous approaches, which are based on feature augmentation, acoustic and pronunciation modeling, or adaptation-based techniques, the proposed framework identifies speakers based on their dialect information combined with spectral and prosodic features. The proposed framework introduces a novel hierarchical framework that identifies a speaker in two unique steps, i.e., speakers' dialect-based training and speaker's dialect-based testing. In speakers' dialect-based training step, initially, the GMM is used for the speakers' dialect classification. The classified dialects' information is then embedded into the spectral and prosodic features, and the combined feature vector is used to train an MLP-based SIS. In this way, the SIS gets trained not only on spectral and prosodic features but also on the speakers' dialect information obtained from the precursory GMM-based speakers' dialect classification system.

During the speaker's dialect-based testing, initially, the speaker's dialect is identified using the trained GMM-based speaker's dialect identification system. In the second step, the speaker's identified dialect information (obtained from the first step) along with spectral and prosodic features is passed to the subsequently trained speaker identification model (trained with speaker's dialect information along with the spectral and prosodic features as stated above). In this way, the SIS identifies the speaker not only with the help of spectral and prosodic features but also with the help of speaker's dialect information. Hence, with the embedding of the speaker's identified dialectical information into the speaker's feature vector, the identification accuracy of the SIS gets improved and a dialect robust identification is achieved. Refer Section 7 for more details on the proposed system design.

The following are the contributions of the proposed novel framework:

(i) The proposed system identifies a speaker using $1:n$ matching, where n is the number of speakers in the identified dialect. Thus, the proposed framework minimizes speaker identification complexity as $n < N$. On the other hand, traditional SISs identify speakers using $1:N$ matching, where N is the total number of speakers in the database.

(ii) As speaker identification complexity is reduced due to dialectical identification, the robustness of system is improved.

The rest of the study is organized as follows: Section 2 presents the comprehensive related work related to the different categories of the accent and dialect recognition systems. Section 3 describes dialectical variations present in the Pashto language. Section 4 describes the development of speech corpora based on Pashto dialectical variations. The speech processing and feature extraction are explained in Section 5. Section 6 illustrates the basic modeling achieved by GMM and ANN. Section 7 describes the basic framework of the proposed system. The results and discussions are presented in Section 8, and finally, Section 9 presents the conclusion of the proposed system along with the future directions.

2. Related Work

The research on designing accent and dialect recognition systems has gained the attention of the researchers in the early 90s, and since then, a variety of accent and dialect recognition systems have been proposed [24]. Mainly, these systems can be grouped into three different categories.

The first category deals with the classification of native and nonnative accents of speakers [25]. The basic purpose of such systems is to overcome the performance loss caused in VPSs due to the accent or dialect mismatches of native and nonnative speakers [26]. Furthermore, the discrimination of nonnative accents from native accents is useful in many

Table 1: Important subdialects of Northern, Southern, and Central variety of dialects and their regions of use.

Main dialects	Subdialects	Regions of use
Southern dialectical variations	Banuchi	Lakki Marwat, Tank, Jandola, Bannu, and Northern Dera Ismail Khan areas of KPK Province of Pakistan
	Kakar	Quetta (Pakistan)
Northern dialectical variation	Afridi	FATA and Khyber Agency (Pakistan)
	Yusufzai	Peshawar, Dir, Swat, Swabi, Mardan, Mohmand, Khyber, and Bajaur agencies, and Hazara Division (Pakistan)
	Wazirwola	South and North Waziristan (Pakistan)
Central dialectical variations	Kabli or Gilji	Kabul and suburb areas (Afghanistan)

intelligence and other applications, such as automated customer service and border control management [27]. One of the other aspects of such types of systems is the recognition of the foreign accent of speakers. Foreign accent recognition has proven to be very helpful in intelligence and security-related applications such as immigrant screening where the passports of the holders can be verified as fake or original just by recognizing their foreign accents. It is also used in other commercial uses such as services based on user agents, voice commands, and targeted advertisement [28, 29].

The second type deals with the automatic accent and dialect identification. In this category, various dialect classification [30, 31], dialect identification [32], and dialect detection [33] systems have been presented. The main purpose of such types of systems is to enhance the performance of voice-based recognizers by detecting, identifying, or classifying their specific accent or dialect well before the recognition. If the accent and dialect of a speaker are identified well before the recognition, the parameters of the recognizer can be adapted for that specific accent and dialect using some adaptation technique [34]. In this way, the variability caused by different accents and dialects of speakers can be minimized, which in turn increases the performance of the recognizers.

The third category of accent and dialect-based systems deals with the identification of regional accents and dialects [35–38]. Identifying regional accents and dialects can help in personalizing synthetic speech of a text-to-speech (TTS) systems according to a speaker of a specific regional accent or dialect. Such systems can also be beneficial for personalizing speech-to-speech translation (S2ST) systems for synthesizing the recognized and translated speech from one language to a specific regional accent in another language. Designing regional accent and dialect recognition systems can also be useful for enhancing regional security of different regions of the world because they can help in identifying suspects' region of origin. In designing accent and dialect-based systems, frequently two approaches have been used, i.e., phonotactic and acoustic [39]. The phonotactic approach is based on the fact that every language consists of phonemes (smallest meaningful words). A phone recognizer tokenizes (breaking the raw text into small chunks) the speech samples into a sequence of its phonemes. Most probably, different dialects of a language differ in phone sequence distributions, which in turn can be used to discriminate among different dialects of a language. In the phonotactic approach, phone sequence distributions and language patterns such as vowel listing, diphthong creation, tense marking, and tones are used to classify among different accents or dialects of a language [40].

On the other hand, the acoustic approach is based on the idea that different accents or dialects of a language differ in terms of spectral characteristics. In this approach, some sort of spectral and prosodic features such as formant frequencies, pitch, pitch slope, duration, and intensity for vowel sounds are used to discriminate among different accents and dialects of a language [30].

Other than the phonotactic and acoustic approaches, some of the popular speaker recognition and language identification techniques such as Gaussian mixture models (GMMs) [41], hidden Markov models (HMMs) [42], support vector machines (SVMs) [43], and i-vectors [44] have also been employed to achieve accent and dialect recognition tasks. I-vector is a front-end factor analysis-based method that usually employs GMMs to reduce the dimensionality of the input into simpler representations, i.e., total variability space [45]. Hence, a probabilistic linear discriminant analysis (pLDA) can then be effectively used to classify dialects or speakers. A newer version of i-vectors, i.e., x-vectors, has been recently proposed by Shon et al. [46]. The x-vector approach replaces i-vectors with language embeddings based on text-based linguistic features that in turn improve accent and dialect identification performance [47, 48].

Since the main purpose of designing accent and dialect recognition systems is to overcome accent and dialect-related variability from the VPSs, therefore, mostly these systems are designed as an accent and dialect interface prior to VPSs to use them as an exact pronunciation dictionary. For each separate accent and dialect class, a separate accent and dialect interface is built. Such accent and dialect interfaces are used by the VPSs to adapt acoustic, morphological, and language models of different accents and dialects for minimizing accentual and dialectical mismatches during the training and testing processes of such systems [49]. Although it is an effective method to achieve accent and dialect robustness, however, a complex feature engineering and sufficient language knowledge are required for designing accent and dialect recognition systems as an exact pronunciation dictionary interface to VPSs. Therefore, it is a difficult task to create such dictionaries. Adding pronunciation variations to the dictionary is another hard task that could lead to substitution errors. Adding a large number of pronunciations of each single word also increases the computational cost because adding alternatives increases the search space. Because of all such reasons, recently, the research community has diverted his attention towards DNN-based techniques and, in the recent past, has proposed various DNN-based approaches to achieve accent and dialect identification tasks more effectively [50–55].

Although acoustic modeling using DNN is an effective technique, it requires a large amount of training data with potential accentual and dialectical variations to design accent and dialect-based systems. Furthermore, for designing accent and dialect robust VPSs, a sufficient amount of data per accent and dialect class is required to train a separate acoustic model for each accent and dialect class. However, even in relatively large available accented speech corpus (such as Accents of the British Isles (ABI) corpus [56]), the amount of data per accent and dialect is limited, and therefore, fewer data are available to train the models with each accent and dialect class. In such cases, identifying accent and dialect classes by selecting model for each accent and dialect is challenging. Instead of constructing model for each separate accent and dialect, the best alternative is to construct a single multiaccent and dialect model by including data from all accents and dialects in the training data set [57]. We have used thesame approach here in our research to classify dialects.

As discussed earlier, Pashto is an under-resourced language, and basic language resources to develop VPSs are limited. In particular, no accented speech data are available to develop accent and dialect-based systems. Therefore, for designing a Pashto accent and dialect-based SIS, we created our own multidialect Pashto data set and used it to train MLP classifier for speaker identification. Prior to speaker identification, GMM has been used to identify speaker's dialect, and then, with the help of the identified dialect, the speaker is identified by the trained speaker identifier (refer Section 7 for more details on the proposed system design). Our own collected Pashto language-based multidialect data set has been used to construct the speaker identification model. It is because the Pashto language is very rich in dialectical variations. The subsequent section outlines the dialectical variations of Pashto in detail.

3. Dialectical Variations in Pashto

Pashto is a national language of Afghanistan and one of the regional languages of Pakistan, which is widely spoken in different regions of Pakistan [58]. The green part of Figure 1 shows regions where the Pashto language is spoken as a first language, and these regions include Lashkar Gah, Umerkot, Kandahar, Kabul, and Jalalabad (Afghanistan side), and Quetta, North and South Waziristan, and Peshawar (Pakistan side).

In all of the abovementioned regions, Pashto is spoken with different dialectical variations. There are three major dialectical variations in Pashto, i.e., Southern, Northern, and Central [60]. Northern and Central dialectical variations are mostly spoken in Pakistan, whereas the Southern dialectical variations are mostly spoken in Afghanistan [61]. Each of these major dialectical variations is further divided into various subdialects [62]. The important subdialects of Northern, Southern, and Central variety of dialects and their respective regions of use are listed in Table 1.

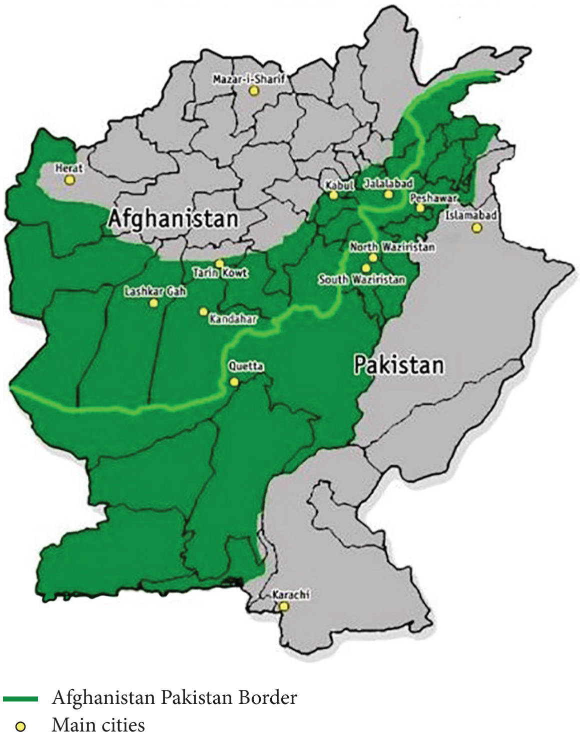


Figure 1: Regions where predominantly Pashto is spoken as the first language [59].

Pashto is also spoken in different regions of the world with different dialectical variations. For example, its Northern dialect is most commonly spoken in India, Canada, UAE, UK, and the USA, while the Southern dialect is spoken in Iran and Tajikistan [63].

4. Corpora Design

To design a SIS for Pashto speakers with accent and dialect identification approach, Pashto speech corpora containing different dialectical variations of Pashto was required. There is no such Pashto speech corpora available, which covers all the major dialectical variations in Pashto; therefore, at the beginning of our research, the required corpora was designed. To include maximum dialectical variations in the designed corpora, the voice data were collected from the native Pashtu speakers of the different regions of Pakistan and Afghanistan. The following regions were considered: Quetta, Pashin, Swat, Hazara division, Mardan, Mohmand Agency, Bannu, South and North Waziristan, Kandahar, Paktika, Kabul, and Federally Administered Tribal Areas (FATA) of Pakistan. All the six dialects listed in Table 2 have been covered by selecting speakers from the above listed regions. Table 2 provides information regarding selected regions and their respective spoken dialects. For

Table 2: Selected dialects with their respective regions of use.

Selected regions	Spoken dialect	Simplified symbol
Quetta and Pishin	Kakar	D1
FATA and Khyber Agency	Afridi	D2
Bannu	Bonucci	D3
South and North Waziristan	Wazirwola	D4
Peshawar, Mardan, Mohmand Agency, Swat, and Hazara Division	Yousufzai	D5
Kabul, Paktika, Qandahar	Kabli or Gilji	D6

convenience, the dialects have been represented as D1, D2, D3, D4, D5, and D6.

A total of 160 native Pashto speakers (including 100 males and 60 females) have been chosen from the selected regions. Table 3 provides the detail of the speakers chosen from different regions.

Table 3 indicates that different numbers of speakers have been chosen from different selected dialects. It is because the spoken areas (geographic) of each dialect are different (small, large, or larger). More numbers of speakers were chosen from the dialects, which are spoken in large area and vice versa. It is evident from Table 3 that there are more male speakers in the recoded database. This is due to the fact that Pashto-speaking regions are culturally conservative, and talking to unknown females is near to impossible. Secondly, the speakers (among available) were selected with their age ranges between 15 and 55 years to remove any age bias. Table 4 illustrates the speakers' age distribution.

To collect voice samples from the speakers, a phonetically rich written script was designed. In the designed script, 25 separate daily conversational-based short-duration (1 to 3 seconds) sentences were included. In particular, those words were included in the sentences, which are pronounced or spoken in a different way in the selected dialects (refer [61] for detail on the written script and the included words in the script). The designed script was then provided to each speaker to read. While reading the script, the voice of the speakers was recorded using high-quality SONY voice recorder and five smartphones manufactured by different manufacturers. All the recordings were performed in quiet rooms located at different locations. Voice samples were recorded in ten different sessions spread between December 2016 and December 2019. All the voice data were collected using 16 kHz sampling frequency.

5. Speech Preprocessing and Feature Extraction

After voice data have been collected, the next step was to process the data through front-end and feature extraction processes. Initially, all the voice data were processed through front-end processing, and then, the features, i.e., MFCC, pitch, and energy, were extracted from the front-end processed data. The details of front-end processing and feature extraction are provided below.

5.1. Front-End Processing

5.1. Front-End Processing. The front-end processing is the basic part of all VPSs. In such systems, the front-end processing is a four-step process, which is mostly used to omit inadequate parts of the collected voice samples for its further processing. The block diagram in Figure 2 illustrates the steps involved in front-end processing. Front-end processing includes preprocessing, preemphasis, framing, and windowing steps, and refer Figure 2. During the preprocessing step, the voice activity detection is performed on raw voice samples through which unvoiced parts of the samples are identified and discarded. After preprocessing, speech samples are preemphasized by passing through a first-order finite impulse response (FIR) filter where the energies of the high-frequency components of the samples are amplified. The preemphasized signal $y(n)$ can be obtained using the following equation:

$$y(n) = x(n) - 0.95x(n-1), \tag{1}$$

where $y(n)$ is preemphasized form of raw speech sample $x(n)$ and 0.95 is the preemphasized parameter. After preprocessing, each speech sample is divided into overlapping frames of duration 20 ms to 30 ms. This process is called framing. After framing, each frame of the sample is analyzed using some sort of window. The Hamming window (one of the most widely used windows in signal processing) is used for this purpose. Equation (2) represents the Hamming window $w(n)$:

$$w(n) = 0.54 - 0.46\cos\frac{2\pi n}{(N-1)}, \quad 0 < n < N-1, \tag{2}$$

where N represents the number of samples in each frame. The window is applied on whole audio sample frame to frame. The resultant sample after the application of window can be achieved using convolution between input signal and filter window as follows:

$$Y(n) = y(n) * w(n), \tag{3}$$

where $Y(n)$ is the resultant sample produced from the input sample $y(n)$ after the application of a window $w(n)$. In this way, the whole speech sample is analyzed.

After the front-end processing, the next step is feature extraction, which is discussed in the ensuing section.

5.2. Feature Extraction Process

5.2.1. Spectral Features. Spectral features approximate the changes produced in the shape and size of vocal tract because of the produced speech signals. Since dialectical variations occur because of different ways of pronouncing words, therefore, the shape and size of vocal tract behave differently with different pronounced sound patterns.

Table 3: Selected dialects with their respective regions of use.

Regions	Dialects	Speakers	Males	Females
Quetta and Pashin	D1	25	16	9
FATA and Khyber Agency	D2	25	16	9
Bannu	D3	20	13	7
South and North Waziristan	D4	25	16	9
Peshawar, Mardan, Mohmand Agency, Swat, and Hazara Division	D5	40	23	17
Kabul, Paktika, Qandahar	D6	25	16	9

Table 4: Selected dialects with their respective regions of use.

Speakers' age interval (years)	No. of speakers
16–20	20
21–25	30
26–30	20
31–35	25
36–40	20
41–45	30
46–50	10
51–55	5

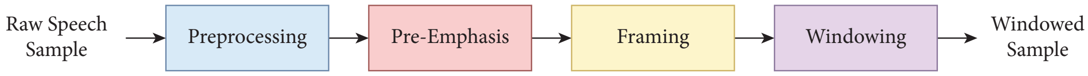


Figure 2: Block diagram of front-end processing.

To extract spectral characteristics from speech signals, various techniques have been used in literature, but the most popular spectral feature extraction technique is MFCC, which has been widely used in all VPSs [64–67].

MFCC is based on the Mel scale, which is linear frequency scaling below 1000 Hz and logarithmic scaling above 1000 Hz. Like the MFCC scaling, the auditory response of human ear is also nonlinear, and therefore, a similar nonlinear approximation causes recognition performance to increase. To extract MFCC features, some steps are followed, which have been illustrated in Figure 3.

After front-end processing, when the windowed frames of a speech sample are achieved, each windowed frame is then processed through fast Fourier transform (FFT), where each frame of the sample is converted from spatial domain to frequency domain. The following equation is used for converting samples from spatial domain to frequency domain.

$$Y(w) = \sum_{0}^{N-1} Y(n)e^{(-j2\pi nk/N)}, \tag{4}$$

where $Y(w)$ is the FFT of $Y(n)$. Once the sample is converted into frequency domain, the power spectrum is obtained by extracting the modulus of the FFT. The power spectrum is then multiplied with a bank of triangular Mel filters, and the logs of the filter bank energies are computed. The following equation is used for this purpose;

$$Y(m) = \log \sum_{0}^{N-1} |Y(w)|^2 T_m(w), \tag{5}$$

where $T_m(w)$ is a triangular filter. Finally, the Mel frequency cepstral coefficients $C(n)$ are calculated by taking discrete cosine transform (DCT) of the log of energies obtained from Equation (5). DCT is obtained using the following equation:

$$C(n) = \sum_{m-1}^{M} Y(m)\cos\left[\frac{\pi n(m-(1/2))}{M}\right], \quad m = 0, 1, \ldots\ldots, M. \tag{6}$$

5.2.2. Prosodic Features. Spectral features analyze shorter frames of the speech signals. However, analyzing shorter frames of speech signals is not enough to capture the maximum information contained in speech signal. It is therefore important to analyze speech signals using longer frames of speech also. Prosodic features of speech such as pitch, energy, duration, and formants use longer frames of speech to analyze speech signals. We propose to use pitch and energy to analyze speech signal and to capture information at longer frame levels. Pitch is commonly known as the fundamental frequency (f_0) of the voiced sound of speech signals. It is directly related to the rate of vibration of vocal fold and is based on the mechanical movement of glottis. It is a perceptual property based on which grave and shrill sounds can be easily identified. Fundamental frequency plays an important role in the identification processes such as speaker identification, dialect identification, and language identification because it contains unique information about speakers' voice quality and dialects [40, 68]. For the calculation of pitch, the autocorrelation method is used. In this method, the correlation of a signal is taken with itself. Time shifting of zero or time shifting equals the fundamental period of the signal gives maximum similarity [69]. In the proposed framework, fundamental frequency or f_0 was computed using the YAAPT pitch tracking algorithm described in [70]. Other than pitch, energy is also an

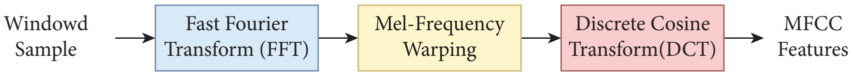


Figure 3: MFCC feature extraction.

important characteristic that is based on varying stress patterns. Since different dialects show variations in stress patterns, therefore, their energies show variations. Energies computed at the frame level are important characteristics for dialect identification tasks [40]. Equation (7) is used to obtain the energy of signal at the frame level.

$$E(n) = \frac{1}{L}\sum_{n=1}^{L} x(n)^2, \tag{7}$$

where $E(n)$ is the normalized energy of the speech signal $x(n), n = 1, \ldots, L$, and L is the frame length.

6. Modeling Techniques

The proposed framework initially classifies/identifies Pashto dialects from a multidialect Pashto data set and then uses classified/identified dialect information to identify a speaker, and refer Section 7 for details on the proposed system design. We propose to use the Gaussian mixture models (GMMs) for dialect classification/identification and artificial neural networks (ANNs) for speaker identification. The working principle of these techniques is briefly explained in this section. It is important to note here that during training, the GMM classifies the speakers' dialects, whereas during testing, initially, the trained GMM identifies a speaker's dialect and passes that information to the cascaded MLP classifier that finally identifies the speaker.

6.1. Modeling with GMM. The Gaussian mixture model (GMM) is a probabilistic modeling technique that takes input data such as a sequence of feature vectors and uses it to create one model per speaker or one model per dialect or accent depending upon the application for which it is used. GMM models each source by a component probability density function (N component densities) and its mixture weights. Each component density is a product of the Gaussian component and a mixture weight. The Gaussian mixture density in the form of a sum of N weighted component densities can be written by the following equation:

$$p(\vec{x}) = \sum_{i=1}^{N} p_i b_i(\vec{x}), \tag{8}$$

where $\vec{x}$ is a D dimensional arbitrary vector, $b_i(\vec{x}), i = 1, 2, \ldots, n,$ is the component densities, and $p_i, i = 1, 2, \ldots, n,$ is the mixture weights. Each component density can be represented by the Gaussian function given in:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{(D/2)}\|\Sigma_i\|^{(1/2)}} e^{\left[(-1/2)\left(\vec{x}-\vec{\mu}_i\right)^T\left(\vec{x}-\vec{\mu}_i\right)\right]}, \tag{9}$$

where $\vec{\mu_i}$ is the mean vector of the extracted feature matrices and Σ_i is the covariance matrix.

The component densities of a mixture collectively form a set of acoustic classes. The speaker's voice can be treated as an acoustic space with a set of acoustic classes. These acoustic classes contain relevant phonetic characteristics of the speaker's vocal such as vowels, nasals, and consonants. In other words, it would be appropriate to say that these acoustic classes provide numerous speaker-dependent vocal tract configurations, which makes them very beneficial for speaker identity. To perform identification by GMM, a set of parameters (model parameters) of mixture density is used, which are given by:

$$y_i = \left[p_i, \sum_i, \vec{\mu_i}\right], \quad i = 1, 2, \ldots, n, \tag{10}$$

where y_i is used as a speaker model or test vector. Identification is performed by obtaining appropriate y_i for each speaker or for each accent or dialect. There are several methods to find y_i, and the most common method is the maximum-likelihood criterion (MLC). MLC is based on the maximization of the likelihood of GMM in finding the model parameters. For a sequence of R training vectors, $X = [\vec{x_1}, \vec{x_2}, \ldots, \vec{x_n}]$, the GMM likelihood criterion may be written as follows:

$$p\left(\frac{x}{y}\right) = \prod_{r=1}^{R} p\left(\frac{\vec{x}_r}{y}\right). \tag{11}$$

The rule of the decision is to select the model that has the largest score as the result.

6.2. Modeling with ANN. Artificial neural network (ANN) is a classification technique inspired by the working mechanism of human biological nervous system. It is a very powerful classification technique among VPSs. We propose to use variant of ANN, called multilayer perception (MLP) with backpropagation (BP) learning algorithm. MLP belongs to the family of neural nets, which consists of interconnected group of artificial neurons called nodes and connections for processing information called edges. A neural network consists of an input, hidden, and output layer and uses optimization algorithm, i.e., stochastic gradient descent to optimize the randomly initialized weights [71, 72]. The input layer transmits inputs in form of feature vector with a weighted value to the hidden layer. The hidden layer is composed of activation units, carries the feature vector from the first layer with weighted value, and performs some calculations as output. The output layer is made up of a single activation unit that carries the weighted output of the hidden layer and predicts the corresponding class [73].

The training mechanism (parameter optimizations) of MLP with BP algorithm is time-consuming as MLP is a fully connected network and number of trainable parameters scales with number of layers, but its importance still persists for applications where the information regarding the input characteristic is limited [74], the same is the case in this research. As already mentioned, MLP uses the stochastic gradient descent method for updating the randomly generated weights. Equation (12) is used for the weight update using the stochastic gradient approach.

$$\delta(w)_{ij}(n+1) = \delta(w)_{ij}(n) + \eta \frac{\partial C}{\partial (w)_{ij}}, \qquad (12)$$

where η represents learning rate, C represents cost function, and w_{ij} represents weights.

7. Proposed System Framework

The proposed framework's block diagram is presented in Figure 4. As shown in Figure 4, the proposed system is designed in a hierarchical format, where, initially, it identifies the speaker's dialect, and then, using the information of the identified dialect, it identifies speaker in cascading manner. The proposed framework has two phases, i.e., training and testing. The detail of each phase is presented below.

7.1. Training Phase. The complete training phase of dialect and speaker identification is shown in Figure 5. Initially, training feature vectors containing spectral and prosodic features of dialects with the correct class labels are provided to dialect recognizer (GMM in our case) to classify speakers' dialects (refer Figure 5(a)). The dialect recognizer generates binary codes for the dialects (for the six dialects, a three-bit binary code was generated). The error is calculated between the desired and generated binary codes for dialects. The calculated error is then fed back to the recognizer, and the process is repeated until the error is reached to a desired minimum level. By this iterative process, the dialect recognizer gets trained on the collected dialects. The generated binary codes are the dialects classification by the recognizer (GMM in our case). It is important to note that GMM is an unsupervised clustering algorithm; however, here we used it as supervised clustering method for speakers' dialect classification.

Figure 5(b) shows the training phase of speaker identification. The training process of speaker recognizer is the same as the training process of dialect recognizer (refer Figure 5(a)). The only difference is that in training feature vectors of speaker identification, binary codes for dialects (obtained from trained dialect model, the first phase of the proposed framework) were also provided to MLP classifier. The purpose of providing binary codes of dialects along with the training feature vectors is to train speaker recognizer not only with the speakers' information but also with the information of their dialects. In this way, the classifier gets trained on the training feature vectors of speakers and on the generated labels (generated by the dialect recognizer) of the speakers' dialects.

After training of speaker and dialect recognizers, the next phase is testing, which is described below.

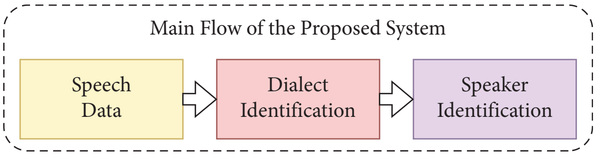


Figure 4: Main flow of the proposed system.

7.2. Test Phase. The test phase of the proposed framework is shown in Figure 6. In the test phase, extracted features from voice recordings are provided to trained dialect and to trained speaker classifiers. A trained dialect classifier identifies dialect information, whereas speaker identification classifier recognizes speaker. The proposed framework is hierarchical in nature as an output of the dialect identification classifier (first step) is fed as an input to the speaker identification classifier (second step). The abovementioned two steps proposed framework recognizes speaker along with identifications of his/her dialect. This approach will be helpful, especially where the voice data set is large and contains a variety of different dialects.

The next section presents results achieved by our proposed framework for dialect identification-based speaker identification.

8. Results and Discussion

As described in Section 4, a total of 160 speakers have been selected from different regions of Pakistan and Afghanistan for voice sample recording. From all the selected speakers, a total of 4000 short-duration sentences were recorded, i.e., 25 samples from each speaker. After recording, the recorded voice samples were processed through front-end and feature extraction processes, and refer Section 5 for further detail about front-end and feature extraction processes. During front-end processing, the recorded voice samples were preprocessed, whereas during the feature extraction process, spectral (MFCC) and prosodic (pitch and energy) features were extracted from them. The extracted features were then grouped by the following arrangement:

(1) Only MFCC features

(2) Pitch and energy features

(3) Combination of MFCC, pitch, and energy features

Each group of features was then separately used during the identification process of the speakers.

As shown in Figure 5, the proposed system works in two phases, i.e., system training phase and testing phase. The system training phases are further divided into dialect identification system training and speaker identification system training phases. The details of achieved results for each phase are provided below.

8.1. System Training Phase and Results

8.1.1. Dialect Identification System Training (Dialect Classification). During DID training, 70% of the data (combination of spectral and prosodic features) was used. A

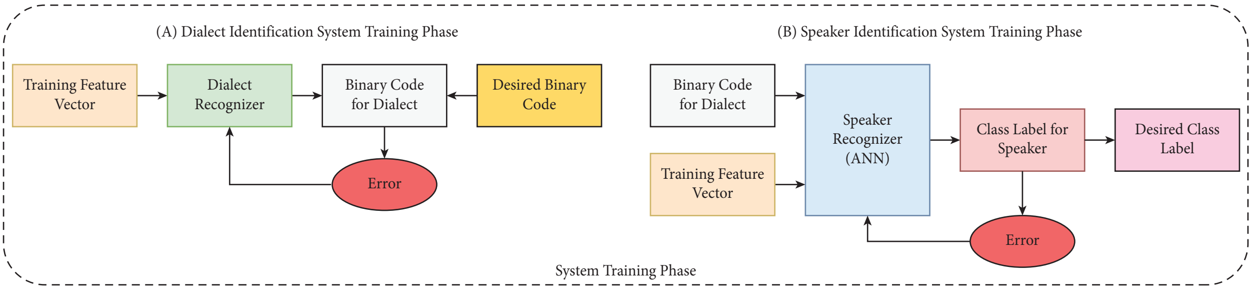


Figure 5: System training. (a) Training of dialect identification system. (b) Training of speaker identification system based on dialect information.

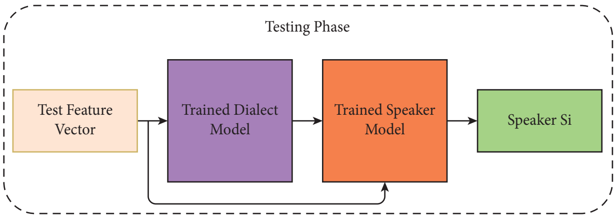


Figure 6: System testing phase.

probabilistic model, i.e., GMM (as explained earlier), was used to classify the dialects. To train this model, the training feature data were provided along with their respective dialect labels in the form of binary codes for each dialect, i.e., 000 for D1 and 001 for D2. The complete training process of GMM is shown in Figure 5(a), which shows that during the training process, GMM generates its own dialect labels (through clustering) and calculates the error between the generated labels and the provided desired labels. The calculated error is then fed back to the model, and the process is repeated until the error is reached to a desired minimum level. Hence, in this way GMM gets trained on the recorded dialects of the Pashto language and it classifies the input multidialect data set into different dialects in the form of binary codes.

Table 5 provides the training accuracies achieved by GMM on each separate dialect, while Figure 7 shows the confusion matrix of the same. The confusion matrix shows the classification of each dialect and correspondence between the target classes along the x-axis and output classes along the y-axis. Average correct classification is shown at the diagonal of the matrix. Referred table and figure show that dialect D5 is the highest recognized dialect, whereas dialect D2 is the lowest recognized dialect. D2 is spoken in limited area, and therefore, it is mostly influenced by the other nearby dialects, and the same is reflected through the achieved recognition accuracy of D2, where D2 is influenced by D5 and D4.

It is also important to note that dialect D5 is the most spoken dialect of Pashto, and in our data set, its number of samples is more as compared to the other dialects (refer section 3, which is one of the reasons of D5 being the most recognized dialect).

8.1.2. Speaker Identification System Training. Similar to the training of DID system, SIS was also trained using labeled data. However, for its training the data were labeled with speaker labels along with the speakers' dialect labels generated by the trained dialect model, i.e., GMM. In this way, the speaker identifier is trained not only by their respective speaker labels but also with their dialect labels. For the effective training of the speaker identifier, the training data were used with different percentage splits and with varying learning rate values. This is done to train system robustly and to identify the effect of splits and learning rates on the framework's performance. Table 6 shows the training performance of the speaker identifier with different percentage splits of the data.

Table 6 shows that the highest training accuracy is achieved when 80% and 85% of the data were used for training. This is because with more data system is able to better model data distribution.

Table 7 shows that the average training accuracy achieved in our experiments is 76.6% and the standard deviation of the achieved training accuracy is 8.84, which shows that the achieved training accuracies are closer to the average training accuracy and there is no outlier in terms of achieved accuracy.

Figure 8 shows the training performance of the speaker identifier with varying learning rate values. Learning rate values of 0.1 and 0.2 provide the highest training accuracy. After system's training, the testing was performed. The detail of the system testing is provided below.

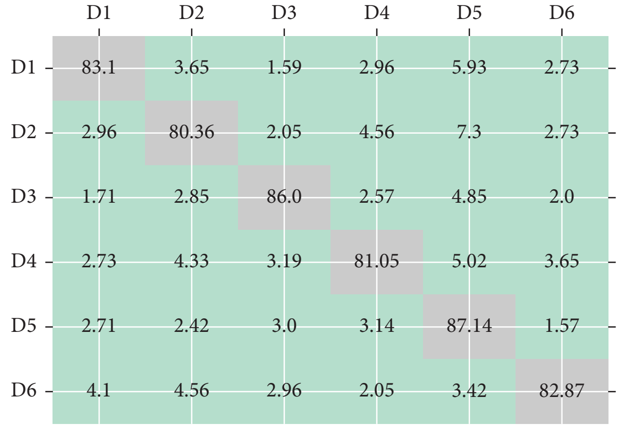
8.2. System Testing Phase and Results. Complete testing procedure is shown in Figure 6. For testing, the test feature vector containing the speakers' label along with their respective dialect labels (generated by the trained dialect model) was provided to the trained dialect model and trained speaker model. With these labels, 200 samples from each dialect were included in the test feature vector.

Table 8 shows the dialect and speaker identification test performance with the combination of spectral and prosodic features, i.e., MFCC + pitch + energy. The referred table shows that with the combination of spectral and prosodic features, the dialect identification system achieved overall average identification accuracy of 80.9%, whereas speaker identification system achieved average identification accuracy of 84.5%.

The performance of SIS and DID systems was further investigated using the aforementioned different groups of features. Figures 9 and 10, respectively, show the

Table 5: Training accuracies of the selected dialects.

Dialect label	Total instances	Training instances	Training (classification) accuracy (%)
D1	625	438	83.10
D2	625	438	80.36
D3	500	350	86.0
D4	625	438	81.05
D5	1000	700	87.14
D6	625	438	82.87



	D1	D2	D3	D4	D5	D6
D1	83.1	3.65	1.59	2.96	5.93	2.73
D2	2.96	80.36	2.05	4.56	7.3	2.73
D3	1.71	2.85	86.0	2.57	4.85	2.0
D4	2.73	4.33	3.19	81.05	5.02	3.65
D5	2.71	2.42	3.0	3.14	87.14	1.57
D6	4.1	4.56	2.96	2.05	3.42	82.87

Figure 7: Confusion matrix of training dialect identification/dialect classification system.

Table 6: Training performance of speaker identification system with different percentages of feature vectors having initial learning rate value = 0.1.

Total instances	Training split (%)	Training instances	Training accuracy (%)
4000	60	2400	66.6
4000	65	2600	68.1
4000	70	2800	72.8
4000	75	3000	79.0
4000	80	3200	86.6
4000	85	3400	86.6

Table 7: Average training accuracy and standard deviation of the achieved training accuracies.

Average training accuracy (%)	Standard deviation
76.6	8.84

identification performance of speaker and dialect identifiers with different groups of features.

Figures 9 and 10 highlight that both the dialect and speaker models achieved the highest identification accuracies when prosodic (pitch and energy) and spectral (MFCC) features were used in combination. Furthermore, the speaker identification system also achieved better performance with spectral (MFCC) features only as compared to the prosodic features, i.e., pitch + energy. It is due to the fact that, in speech processing, MFCC is a better approach to approximate the human auditory response [75].

Figure 11 shows the confusion matrix of DID system testing. It shows that the highest identification accuracy is achieved by dialect D5, whereas the lowest identification accuracy is achieved by dialect D3. It further shows that dialects D2 and D3 were the most confusing dialects for dialect D1. Pashto dialects contain overlapping boundaries, and it is usually difficult even for native Pashto speakers to differentiate the dialects. This is the main reason behind the mixing behavior of the classifier to explicitly identify the dialects. Figure 12 shows the confusion matrix of the SIS, which shows that all the speakers whose voice samples were included in the test sample have been well recognized by the designed SIS. It further shows that as an individual speaker, S8 was the highest identified speaker, whereas S2 was the least identified speaker.

Table 9 provides the detailed identification report achieved by the SIS by illustrating the weighted average values of the popular performance measures, i.e., true-positive rate (TPR), true-negative rate (TNR), precision, and F score.

All the performance indicators illustrated in Table 9 show that the designed SIS using the proposed approach performed well in all respect.

In the next section, the results of SIS achieved with the proposed approach have been compared with the results of SIS designed using the traditional approach.

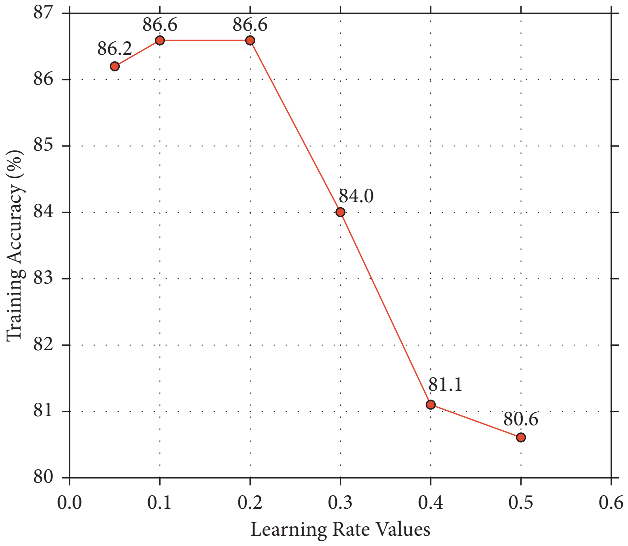


Figure 8: Training performance of speaker identification system with 80% training split of data and varying learning rate values.

Table 8: Test performance of speaker and dialect identification systems with the combination of spectral and prosodic features.

System	Number of test instances	Overall identification accuracy (%)
Dialect identification	200	80.9
Speaker identification	200	84.5

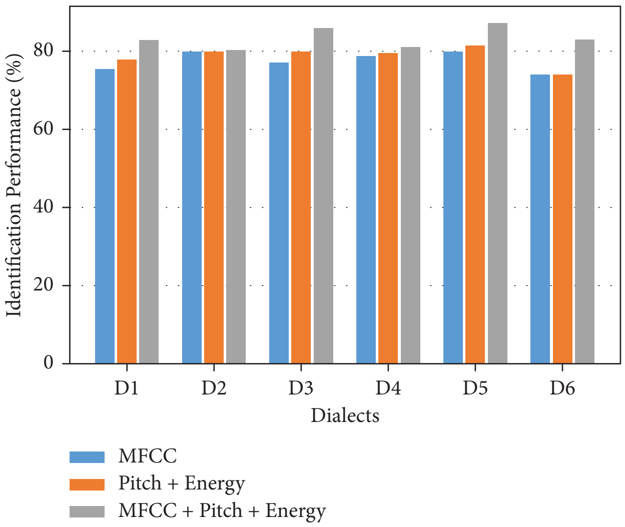


Figure 9: Identification performance of dialect identification system with a different set of features.

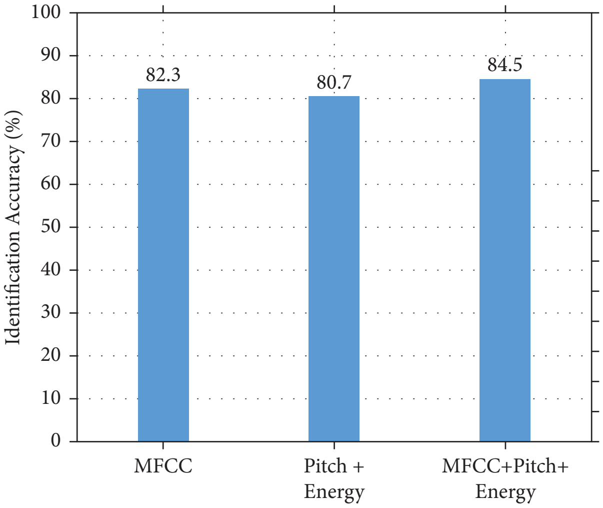
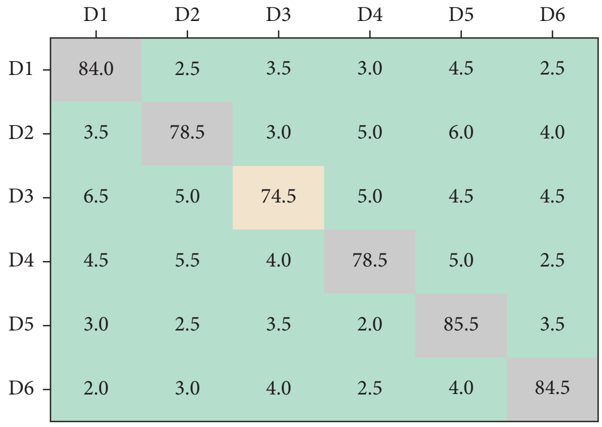


Figure 10: Identification performance of speaker identification system with a different set of features.

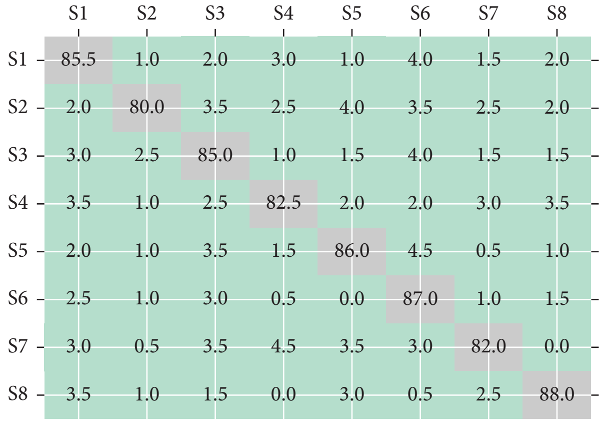
8.3. Comparison of the Proposed Approach with the Traditional Approach. To further validate the performance of designed SIS, another SIS was designed using the traditional approach (trained and tested on the same data set, i.e., our own collected data set) and the result was compared. Here, the meanings of the traditional approach refer to the common approach used in designing SIS as described in [76]. Figure 13 provides the comparison of recognition accuracies achieved by the SIS designed using the proposed approach and SIS using the traditional approach.

It is clear from Figure 13 that the proposed framework outperformed the traditional approach in identifying speakers and achieved 11.0% improvement in speaker identification accuracy. Furthermore, Table 10 shows that the proposed SIS framework reduced the computational time by 39% in identifying the speaker.



	D1	D2	D3	D4	D5	D6
D1	84.0	2.5	3.5	3.0	4.5	2.5
D2	3.5	78.5	3.0	5.0	6.0	4.0
D3	6.5	5.0	74.5	5.0	4.5	4.5
D4	4.5	5.5	4.0	78.5	5.0	2.5
D5	3.0	2.5	3.5	2.0	85.5	3.5
D6	2.0	3.0	4.0	2.5	4.0	84.5

Figure 11: Confusion matrix of testing dialect identification system.



	S1	S2	S3	S4	S5	S6	S7	S8
S1	85.5	1.0	2.0	3.0	1.0	4.0	1.5	2.0
S2	2.0	80.0	3.5	2.5	4.0	3.5	2.5	2.0
S3	3.0	2.5	85.0	1.0	1.5	4.0	1.5	1.5
S4	3.5	1.0	2.5	82.5	2.0	2.0	3.0	3.5
S5	2.0	1.0	3.5	1.5	86.0	4.5	0.5	1.0
S6	2.5	1.0	3.0	0.5	0.0	87.0	1.0	1.5
S7	3.0	0.5	3.5	4.5	3.5	3.0	82.0	0.0
S8	3.5	1.0	1.5	0.0	3.0	0.5	2.5	88.0

Figure 12: Confusion matrix of speaker identification system.

Table 9: Detailed analysis report of the speaker identification system using the proposed approach.

Class	TPR (sensitivity)	TNR (specificity)	Precession	*F* score
Weighted average	0.845	0.810	0.856	0.850

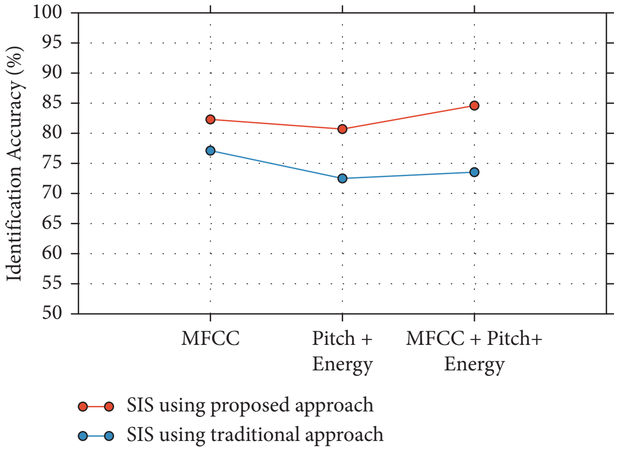


Figure 13: Identification performance of SIS with traditional approach vs SIS with proposed approach.

Table 10: Identification time of the proposed approach vs traditional approach.

Approach	Accuracy with spectral and prosodic features (%)	Average identification time (s)	Difference in average identification time
Proposed	84.5	62	39% more time
Traditional	73.5	158	39% less time

9. Conclusion and Future Work

For minimizing the dialect-related variability and to increase the performance of speaker identification system, we proposed a novel framework. The following are the main achievements of the presented framework:

(i) Presented approach identified speaker by a simple test feature vector not only containing only the speaker's information in form of spectral and prosodic features but also containing the speaker's dialect information (binary codes generated by trained dialect model).

(ii) Identifying speakers using their dialect information reduced the dialect-based variability (variability present because of variety of dialects of a language) from the designed SIS. This way the designed SIS achieved enhanced performance (11.0% increase) as compared to traditionally designed SIS using the same data set.

(iii) As compared to the traditional approaches of identifying speakers, the presented approach identifies a speaker in reduced search space (traditional SIS identifies speakers using *1:N* matching, where *N* is the total number of speakers in database. On the other hand, the proposed system identifies a speaker using *1:n* matching, where *n* is the number of speakers in the identified dialect only).

(iv) Proposed SIS was found to be time-efficient in identifying speakers. It is because the system search space is minimized during the testing/identification process.

For DID, GMM was used, whereas for speaker identification (SI), MLP was used. Spectral (MFCC) and prosodic (pitch and energy) features have been extracted from the voice samples of 160 speakers of different regions of Pakistan and Afghanistan. Different combinations of extracted features, i.e., (i) only MFCC, (ii) pitch + energy, and (iii) MFCC + pitch + energy (combination of spectral and prosodic), were used to train the identifiers. Testing was also performed using the same combinations of features, and the results were compared. Comparative results show that the combination of spectral and prosodic features provided the highest identification accuracy with dialect and speaker identifiers.

Finally, a SIS was designed using the traditional approach trained and tested on the same set of feature vectors and its performance was compared with the SIS designed using the proposed approach. SIS designed using the proposed approach achieved much better identification accuracy as compared to the traditionally designed SIS and showed 11.0% improvement in identification accuracy. The time efficiency of both systems was also compared. It is also found that the SIS designed using the proposed approach took 39% less average time in identifying speakers as compared to traditionally designed SIS.

To further evaluate the proposed approach, in future, the approach/framework will be tested on the large publicly available data sets.

Data Availability

The data can be obtained from the corresponding author on request.

Disclosure

The data set used in this research is the extension of the data set presented in [61]. Complete data set has been collected with self-efforts.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] R. Roberts and M. Page, "Secure voice biometric authentication," *US Patent App*, vol. 16, p. 434, 2019.

[2] A. Nagrani, J. S. Chung, and A. Zisserman, *Voxceleb: A Large-Scale Speaker Identification Dataset*, Interspeech, Incheon, Republic of Korea, 2017.

[3] H. Ren, Y. Song, S. Yang, and F. Situ, "Secure smart home: a voiceprint and internet based authentication system for remote accessing," in *Proceedings of the 2016 11th International Conference on Computer Science & Education (ICCSE)*, pp. 247–251, Nagoya, Japan, August 2016.

[4] H. Abdullah, W. Garcia, C. Peeters, P. Traynor, K. R. Butler, and J. Wilson, "Practical hidden voice attacks against speech and speaker recognition systems," 2019, https://arxiv.org/abs/1904.05734.

[5] H. Huang, C. Zhang, X. Xu, C. Zhang, and D. Wang, "Voice-based user interface with dynamically switchable endpoints," *NoteUS Patent App*, vol. 15, p. 109, 2019.

[6] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2016, https://arxiv.org/abs/1607.02533.

[7] X. Yuan, Y. Chen, Y. Zhao et al., "A systematic approach for practical adversarial voice recognition," in *Proceedings of the 27th USENIX Security Symposium (USENIX Security 18)*, pp. 49–64, Baltimore, MD, USA, August 2018.

[8] N. Carlini and D. Wagner, "Audio adversarial examples: targeted attacks on speech-to-text," in *Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7, San Francisco, CA, USA, May 2018.

[9] X. Fang, L. Zou, J. Li, L. Sun, and Z.-H. Ling, "Channel adversarial training for cross-channel text-independent speaker recognition," in *Proceedings of the ICASSP 2019-2019*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6221–6225, IEEE, Brighton, UK, May 2019.

[10] D. Ribas and E. Vincent, "An improved uncertainty propagation method for robust i-vector based speaker recognition," in *Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE*, pp. 6331–6335, Brighton, UK, May 2019.

[11] M. K. Nandwana, J. van Hout, M. McLaren et al., *Robust Speaker Recognition From Distant Speech Under Real Reverberant Environments Using Speaker Embeddings*, Interspeech, pp. 1106–1110, Incheon, Republic of Korea.

[12] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: a tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[13] M. McLaren, V. Abrash, M. Graciarena, Y. Lei, and J. Pesán, *Improving Robustness to Compressed Speech in Speaker Recognition*, Interspeech, pp. 3698–3702, Incheon, Republic of Korea.

[14] M. M. Ž. Nedeljković, U. Glavitsch, and Ž. DJurović, "Speaker modeling using emotional speech for more robust speaker identification," *Journal of Communications Technology and Electronics*, vol. 64, pp. 1256–1265, 2019.

[15] M. A. Nematollahi, H. Gamboa-Rosales, F. J. Martinez-Ruiz, J. I. De la Rosa-Vargas, S. A. R. Al-Haddad, and M. Esmaeilpour, "Multi-factor authentication model based on multipurpose speech watermarking and online speaker recognition," *Multimedia Tools and Applications*, vol. 76, no. 5, pp. 7251–7281, 2017.

[16] K. Mannepalli, P. N. Sastry, and M. Suman, "Mfcc-gmm based accent recognition system for Telugu speech signals," *International Journal of Speech Technology*, vol. 19, no. 1, pp. 87–93, 2016.

[17] S. Kibria, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "Acoustic analysis of the speakers' variability for regional accent-affected pronunciation in Bangladeshi bangla: a study on Sylheti accent," *IEEE Access*, vol. 8, pp. 35200–35221, 2020.

[18] G. A. Liu and J. H. Hansen, "A systematic strategy for robust automatic dialect identification," in *Proceedings of the 19th European Signal Processing Conference*, pp. 2138–2141, IEEE, Dublin, Ireland, August 2011.

[19] F. Biadsy, "Automatic dialect and accent recognition and its application to speech recognition," Ph.D. Thesis, Columbia University, New York City, NY, USA, 2011.

[20] M. H. Bahari, R. Saeidi, and D. Van Leeuwen, "Accent recognition using i-vector, Gaussian mean supervector and Gaussian posterior probability supervector for spontaneous telephone speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7344–7348, IEEE, Vancouver, Canada, October 2013.

[21] A. Lazaridis, E. el Khoury, J.-P. Goldman, M. Avanzi, S. Marcel, and P. N. Garner, *Swiss French Regional Accent Identification*Odyssey, Hong Kong, China, 2014.

[22] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: a survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[23] S. Nisar and M. Tariq, "Dialect recognition for low resource language using an adaptive filter bank," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 16, no. 4, Article ID 1850031, 2018.

[24] T. Carlos, I. Trancoso, and A. Serralheiro, "Accent identification, proceeding of fourth international conference on spoken language processing," *IEEE*, vol. 3, pp. 1784–1787, 1996.

[25] J. J. Bird, E. Wanner, A. Ekárt, and D. R. Faria, "Accent classification in human speech biometrics for native and non-native English speakers," in *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pp. 554–560, Rhodes, Greece, June 2019.

[26] F. Biadsy, J. Hirschberg, and M. Collins, "Dialect recognition using a phone-gmm-supervector-based svm kernel," in *Proceedings of the Eleventh Annual Conference of the International Speech Communication Association*, Makuhari, Japan, September 2010.

[27] E. Shriberg, L. Ferrer, S. S. Kajarekar, N. Scheffer, A. Stolcke, and M. Akbacak, *Detecting Nonnative Speech Using Speaker Recognition Approaches*p. 26, Odyssey, Incheon, Republic of Korea, 2008.

[28] H. Behravan, V. Hautamäki, and T. Kinnunen, "Factors affecting i-vector based foreign accent recognition: a case study in spoken Finnish," *Speech Communication*, vol. 66, pp. 118–129, 2015a.

[29] H. Behravan, V. Hautamäki, S. M. Siniscalchi, T. Kinnunen, and C.-H. Lee, "I-vector modeling of speech attributes for automatic foreign accent recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 29–41, 2015b.

[30] S. S. Agrawal, A. Jain, and S. Sinha, "Analysis and modeling of acoustic information for automatic dialect classification," *International Journal of Speech Technology*, vol. 19, no. 3, pp. 593–609, 2016.

[31] F. Huang, "Improved Arabic dialect classification with social media data," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2118–2126, Lisbon, Portugal, September 2015.

[32] S. Malmasi, E. Refaee, and M. Dras, "Arabic dialect identification using a parallel multidialectal corpus," in *Proceedings of the Conference of the Pacific Association for Computational Linguistics*, pp. 35–53, Springer, Bali, Indonesia, May 2015.

[33] A. Ali, N. Dehak, P. Cardinal et al., "Automatic dialect detection in arabic broadcast speech," 2015, https://arxiv.org/abs/1509.06928.

[34] B. Pellom, G. Haupt, and K. Ridgeway, "Voice-based liveness verification," *US Patent*, vol. 10, p. 512, 2019.

[35] M. Najafian, S. Safavi, P. Weber, and M. J. Russell, *Identification of British English Regional Accents Using Fusion of I-Vector and Multi-Accent Phonotactic Systems*, pp. 132–139, Odyssey, Incheon, Republic of Korea, 2016.

[36] S. Darjaa, R. Sabo, M. Trnka, M. Rusko, and G. Múcsková, "Automatic recognition of Slovak regional dialects," in *Proceedings of the 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pp. 305–308, Košice, Slovakia, August 2018.

[37] B. Uslu and H. Tora, "Turkish regional dialect recognition using acoustic features of voiced segments," *IJSPS*, vol. 8, 2018.

[38] M. Djellab, A. Amrouche, A. Bouridane, and N. Mehallegue, "Algerian modern colloquial Arabic speech corpus (amcasc): regional accents recognition within complex socio-linguistic environments," *Language Resources and Evaluation*, vol. 51, no. 3, pp. 613–641, 2017.

[39] E. J. Harfash and A. H. Abdul-kareem, "Automatic Arabic dialect classification," *International Journal of Computer Application*, vol. 8887, 2017.

[40] A. Etman and A. Louis, "American dialect identification using phonotactic and prosodic features," in *Proceedings of the SAI Intelligent Systems Conference (IntelliSys)*, pp. 963–970, IEEE, London, UK, November 2015.

[41] P. P. Das, S. M. Allayear, R. Amin, and Z. Rahman, “Bangladeshi dialect recognition using mel frequency cepstral coefficient, delta, delta-delta and Gaussian mixture model,” in *Proceedings of the Eighth International Conference on Advanced Computational Intelligence (ICACI)*, pp. 359–364, IEEE, Chiang Mai, Thailand, February 2016.

[42] I. Kardava, J. Antidze, J. Antidze, and N. Gulua, “Solving the problem of the accents for speech recognition systems,” *International Journal of Signal Processing Systems*, vol. 4, no. 3, pp. 235–238, 2016.

[43] S. Wray, “Classification of closely related sub-dialects of Arabic using support-vector machines,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018.

[44] S. Malmasi and M. Zampieri, “Arabic dialect identification using ivectors and asr transcripts,” in *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pp. 178–183, VarDial), Valencia, Spain, April 2017.

[45] I. Salwani and R. Athiar, “I-vector extraction for speaker recognition based on dimensionality reduction,” *Procedia Computer Science*, vol. 126, pp. 1534–1540, 2018.

[46] S. Shon, A. Ali, and J. Glass, “Convolutional neural networks and language embeddings for end-to-end dialect recognition,” https://arxiv.org/abs/1803.04567.

[47] S. Yoo, I. Song, and Y. Bengio, “A highly adaptive acoustic model for accurate multi-dialect speech recognition,” in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5716–5720, IEEE, Brighton, UK, May 2019.

[48] A. Hanani and R. Naser, “Spoken Arabic dialect recognition using X-vectors,” *Natural Language Engineering*, vol. 26, no. 6, pp. 691–700, 2020.

[49] F. Biadsy and J. Hirschberg, “Using prosody and phonotactics in Arabic dialect identification,” in *Proceedings of the Tenth Annual Conference of the International Speech Communication Association*, Brighton, UK, September 2009.

[50] Z. Ren, G. Yang, and S. Xu, “Two-stage training for chinese dialect recognition,” 2019, https://arxiv.org/abs/1908.02284.

[51] C. Themistocleous, “Dialect classification from a single sonorant sound using deep neural networks,” *Frontiers in Communication*, vol. 4, p. 64, 2019.

[52] C. Suman, P. Kumar, S. Saha, and P. Bhattacharyya, “Gender, age and dialect recognition using tweets in a deep learning framework-notebook for fire,” in *Proceedings of the Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings*, pp. 12–15, CEUR-WS.org, Kolkata, India, December 2019.

[53] X. Chen and J. Cheng, “Deep neural network acoustic modeling for native and non-native Mandarin speech recognition,” in *Proceedings of the 9th International Symposium on Chinese Spoken Language Processing*, pp. 6–9, IEEE, Singapore, September 2014.

[54] Y. Huang, D. Yu, C. Liu, and Y. Gong, “Multi-accent deep neural network acoustic model with accent-specific top layer using the kld-regularized model adaptation,” in *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association*, Singapore, September 2014.

[55] M. Chen, Z. Yang, J. Liang, Y. Li, and W. Liu, “Improving deep neural networks based multi-accent Mandarin speech recognition using i-vectors and accent-specific top layer,” in *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association*, Dresden, Germany, September 2015.

[56] G. Silke, R. Stefan, and K. Ralf, “Generating non-native pronunciation variants for lexicon adaptation,” *Speech Communication*, vol. 42, pp. 109–123, 2004.

[57] N. Maryam and M. Russell, “Automatic accent identification as an analytical tool for accent robust automatic speech recognition,” *Speech Communication*, vol. 122, pp. 44–55, 2020.

[58] I. Ali and B. Shah, “The 19th and early 20th-century us women’s rights struggle: implications for contemporary afghani and pakistani pashtun women,” *Central Asia*, vol. 85, pp. 133–160, 2019.

[59] World Press, “Archive for the “pashtunistan” category,” https://outofcentralasianow.wordpress.com/category/pashtunistan/ Accessed 26 Dec 2020.

[60] S. M. Shah, M. Memon, and M. H. Salam, “Speaker recognition for Pashto speakers based on isolated digits recognition using accent and dialect approach,” *Journal of Engineering Science & Technology*, vol. 15, pp. 2190–2207, 2020.

[61] i. M. Shah, S. A. Memon, K. u. R. Khoumbati, and M. Moinuddin, “A pashtu speakers database using accent and dialect approach,” *International Journal of Applied Pattern Recognition*, vol. 4, no. 4, pp. 358–380, 2017.

[62] S. Khan, H. Ali, and K. Ullah, “Pashto language dialect recognition using mel frequency cepstral coefficient and support vector machines,” in *Proceedings of the 2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT)*, pp. 1–4, IEEE, Karachi, Pakistan, April 2017.

[63] A. W. Abbas, N. Ahmad, and H. Ali, “Pashto spoken digits database for the automatic speech recognition research,” in *Proceedings of the 18th International Conference on Automation and Computing (ICAC)*, pp. 1–5, IEEE, Loughborough, UK, September 2012.

[64] N. B. Chittaragi, A. Prakash, and S. G. Koolagudi, “Dialect identification using spectral and prosodic features on single and ensemble classifiers,” *Arabian Journal for Science and Engineering*, vol. 43, no. 8, pp. 4289–4302, 2018.

[65] A. Ashar, M. S. Bhatti, and U. Mushtaq, “Speaker identification using a hybrid cnn-mfcc approach,” in *Proceedings of the International Conference on Emerging Trends in Smart Technologies (ICETST)*, pp. 1–4, IEEE, Karachi, Pakistan, March 2020.

[66] P. B. Ramteke, S. Supanekar, and S. G. Koolagudi, “Gender identification using spectral features and glottal closure instants (gcis),” in *Proceedings of the Twelfth International Conference on Contemporary Computing (IC3)*, pp. 1–6, IEEE, Noida, India, August 2019.

[67] H. Mukherjee, S. M. Obaidullah, K. C. Santosh, S. Phadikar, and K. Roy, “A lazy learning-based language identification from speech using mfcc-2 features,” *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 1, pp. 1–14, 2020.

[68] H. S. Das and P. Roy, “Optimal prosodic feature extraction and classification in parametric excitation source information for indian language identification using neural network based q-learning algorithm,” *International Journal of Speech Technology*, vol. 22, no. 1, pp. 67–77, 2019.

[69] S. M. Jagdale, A. A. Shinde, and J. S. Chitode, “Robust speaker recognition based on low-level- and prosodic-level-features,” in *Advances in Data Sciences, Security and Applications*-Springer, Berlin, Germany, 2020.

[70] S. A. Zahorian and H. Hu, “A spectral/temporal method for robust fundamental frequency tracking,” *Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4559–4571, 2008.

[71] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[72] G. Hinton, L. Deng, D. Yu et al., "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[73] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8599–8603, IEEE, Vancouver, Canada, May 2013.

[74] A. Sharma, S. P. Singh, and V. Kumar, "Text-independent speaker identification using backpropagation mlp network classifier for a closed set of speakers," in *Proceedings of the Fifth IEEE International Symposium on Signal Processing and Information Technology*, pp. 665–669, IEEE, Ajman, UAE, December 2005.

[75] V. Tiwari, "MFCC and its applications in speaker recognition," *International Journal on Emerging Technologies*, vol. 1, pp. 19–22, 2010.

[76] J. P. Campbell, "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.