Journal of Engineering Science and Technology
Vol. 15, No. 4 (2020) 2190 - 2207
© School of Engineering, Taylor's University

SPEAKER RECOGNITION FOR PASHTO SPEAKERS BASED ON ISOLATED DIGITS RECOGNITION USING ACCENT AND DIALECT APPROACH

SHAHID MUNIR SHAH[1,*], MUHAMMAD MEMON[2],
MUHAMMAD HAMMAD U. SALAM[3]

[1]Department of Computer Science, Faculty of Information Technology, Barrett Hodgson University, Korangi Creek, Karachi, Pakistan
[2] Institute of Business Administration, University of Sindh, Jamshoro, Pakistan
[3]University of Kotli, Azad Jammu and Kashmir, Pakistan
*Corresponding Author: shahid.munir@bhu.edu.pk

Abstract

This paper presents a Pashto speech recognition based on Pashto isolated digits' recognition. In order to develop the system, initially, a Pashto isolated digits' database containing different dialectical variations of Pashto was developed. In the database, Pashto isolated digits one (yao) to ten (las) were recorded from sixty native Pashto speakers of different areas of Pakistan and Afghanistan. To the best of our knowledge, it is the first Pashto digits database that contains all the major dialectical variations of Pashto. After the speech data has been collected, spectral features (Mel Frequency Cepstral Coefficients (MFCC)) and prosodic features (Pitch and Energy) have been extracted from the collected data and Multi-Layer Perceptron (MLP) algorithm of Artificial Neural Network (ANN) was used to classify the digits. The presented system achieved 97.5% recognition accuracy with spectral features, 96.0% recognition accuracy with prosodic features and 98.0% recognition accuracy with the combination of spectral and prosodic features. With these achieved results, the system outperformed some of the recently proposed Pashto based speech and dialect recognizers by showing enhancement in recognition accuracy. Support Vector Machine (SVM) and Hidden Markov Model (HMM) classifiers were also tested on our collected dataset and results were compared with the results achieved by MLP. SVM showed improvement over HMM but overall, the performance of MLP was the best in classifying Pashto isolated digits.

Keywords: Accents and dialects, Dialect recognition, Isolated digits recognition, Pashto language, Under resourced language.

1. Introduction

Speech is a main mode of human communication but at the present era of Human Computer Interactions (HCI) and expert systems, human speech is used much beyond the human communication [1]. Voice dialling (making calls by voice), automatic call distributing (answering and distributing an incoming call), demotic appliance control (control of appliances in a smart home), easy data entry (entering sequence of numbers), and text to speech processing (converting audio to speech) etc. are the best examples of HCI through speech.

HCI through speech is quite demanding because it entails important speech characteristics to be captured from a speech signal. Speech signals are non-stationary in nature; therefore, complex algorithms are used to extract important characteristics from them. Automatic Speech Recognition Systems (ASRS) are one of innovative applications of HCI through speech.

ASRS detect “what is spoken” by a speaker from his produced sound. It is done by transforming the produced sound into digital signal from its original analogue form, splitting it into essential language phonemes (the smallest possible units of sound, which can discriminate one word of a language from the other), create words from produced phonemes and finally, examine the words contextually to check the spelling of the words according to the sound wave [2]. Linking the created word patterns with the stored patterns, the recognition decision is made.

ASRS are generally divided into Isolated Speech Recognition Systems (ISRS) and Fluent Speech Recognition Systems (FSRS). ISRS reorganize the words uttered by a speaker with a single break or pause. In such cases, it is easy to recognize the boundaries of the words pronounced. On the other hand, FSRS reorganize the words without or least break. In such systems, the words’ boundaries are relatively difficult to recognize because of overlying of the words [3].

Like the other pattern recognition systems, training and testing are the essential stages of ASRS. During training stage, the recorded speech signals are transformed into parametric representation (representation of input speech signal into set of symbols) [4] and stored as a reference template. During testing, the parametric representation of an unknown speaker are matched with the stored template and decision is made.

Unfortunately, the matching process of ASRS is not well achieved because of certain unwanted factors (variability) present in them. Background noise, speakers’ age and emotions and use of different devices for recording training & testing data are the most persuasive variability of ASRS. All such variability cause training and testing data mismatch during matching process of recognition and hence cause performance degradation in the designed systems [5].

Other than the aforementioned variability, dialectical variations of a language are also a major variability and a great source of performance degradation of ASRS [6]. Performance of ASRS degrade if one of the dialect of a language is used for system training and some other dialect of the language is used for system testing [7].

In order to enhance performance of ASRS, dialect variations and the other present variability need to be addressed properly and timely. However, ASRS designed for the languages that are rich in dialectical variations need more work on addressing dialectical variations. The case of Pashto is the same because it is a language, which is very rich in dialectical variations and is spoken with different dialects in different regions [8]. It is an under resourced language in form of available resources for developing ASRS. For example, rich databases, lexicons, grammar checkers, spell checkers etc. are not commonly available in this language as compared to English, Japanese, Mandarin, French, etc. [9-12], which are comparatively much developed. Because of less available resources like the other under resourced languages, Pashto language has very less progress of ASRS development [13]. Even though, the Pashto language is one of the widely spoken languages in the world (approximately 45-55 million speakers around the world speak Pashto [14]), it has insufficient progress for ASRS development.

In order to strengthen the research on designing Pashto based ASRS here in this paper, a Pashto ASRS with its major dialectical variations has been proposed. The basic objective of the proposed system is the speech recognition based on isolated uttered words of Pashto and to minimize the dialect related variability from the designed ASRS. For developing the system, initially, a voice database (in form of isolated Pashto digits) including major dialectical variations of Pashto was developed by collecting voice samples from different native Pashto speakers. Spectral and prosodic features (MFCC, pitch, energy) were extracted from the collected data and classification was performed using MLP algorithm of ANN. To the best of our knowledge, spectral and prosodic features combined with MLP for isolated digit recognition has not yet reported for Pashto based ASRS. The purpose of using MLP for the Pashto isolated digits recognition is the overlap of Pashto dialects with each other. Pashto dialects do not contain separate/sharp/linear boundaries rather have mixed boundaries (dialects of Pashto cannot be exactly distinguishable from each other). In such cases, MLP is a better choice because of its capability of recognizing overlapping boundaries. Using MLP for Pashto isolated digits recognition and identifying it as a best choice for Pashto (based on the nature if its dialects) is a major contribution of this study. Furthermore, the developed Pashto digits dataset covering all major dialectical variations of Pashto is also a unique contribution of this study. The authors believe that the present work will serve as a baseline for the forthcoming research on Pashto based ASRS.

After the system has been developed, the results achieved by the MLP classifier were compared with the recent published work in literature and SVM and HMM based classifiers tested on same dataset. Comparative study shows that the MLP classifier achieved the best performance over SVM and HMM classifiers and also achieved the improved recognition accuracies over some of the existing systems in literature.

It is important to mention here that research presented here, is the extension of the work presented in [15] in which the dataset was collected on small scales. In this paper, the speech data in form of isolated digits has been collected on much larger scale as compared to [15] and also new set of features (prosodic and sum of spectral and prosodic have been tested to enhance the system performance (refer section 9).

The rest of the paper is organized as follows: in section 2 related work is discussed. Section 3 provides a note on phonetic inventory and dialectical variations of Pashto. Section 4 describes voice database development phase. Section 5 presents feature extraction techniques used in this paper. Section 6 provides the detail of MLP classifier. Section 7 presents the results achieved by the

MLP classifier on the collected dataset. Section 8 provides the results of SVM and HMM classifiers tested on the same dataset. Section 9 presents the comparative results of the MLP, SVM and HMM classifiers using spectral and prosodic features and their combination. Section 10 provides the comparative study of the proposed system with some of the other recently proposed similar systems (in literature). Section 11 presents the conclusion with future directions and finally the references have been provided.

2. Related Work

In relation with the present study, here in this section only Pashto language based ASRS have been described.

Ahmed et al. [16] developed an automatic Pashto speech recognition system based on isolated words. Initially, a medium-vocabulary speech corpus was developed containing 161 most common daily used isolated words such as the names of the 7 days of week and Pashto digits from 0 to 25. The words were pronounced by 50 Pashto speakers (28 males and 22 females) both natives and non-natives and of different ages and genders. MFCC with their first and second derivatives were used as features and recognition was performed using Linear Discriminative Analysis (LDA). The system achieved different error rates (0 to 60 %) in recognizing the isolated uttered words. According to the authors, the main reason behind the high error rates was the accentual variations of Pashto and inadequate volume of training data.

Tanzeela et al. [17] investigated the impact of MFCC and LDA on speaker independent Pashto isolated spoken digits based ASRS. For designing the system, a speech database was developed in which Pashto digits from sefer (0) to sul (100) were recorded from 50 speakers (25 males and 25 females) of different ages. In order to minimize the dialectical variations of Pashto, the digits were recorded in only a single dialect Yusufzai dialect (the most spoken dialect of Pashto). MFCC features were extracted from the recorded speech data and LDA classifier was used to classify the digits. The system achieved 80% accuracy in recognizing the digits.

Ali et al. [18] developed a Pashto speech database and designed an ASRS from the developed database. Pashto isolated spoken digits from sefer (0) to naha (9) were recorded from 50 Pashto native speakers including 25 males and 25 females with their ages ranging between 18 to 60 years. MFCC features were extracted from the recorded speech data and K Nearest Neighbor (KNN) was used for classification. The system achieved overall 76.8% recognition accuracy.

Nisar and Asadullah [19] proposed home automation solution using Pashto digits recognition. Pashto digits from 1 to 10 were recorded from 75 native Pashto speakers both males and females having their ages ranging between 18 to 60 years. MFCC features were extracted from the recorded Pashto digits and classification was accomplished using KNN. The system achieved overall average 78.8% recognition accuracy.

Nisar et al. [20] presented a Pashto spoken digits' recognition system based on spectral and prosodic features. Initially, a Pashto digits database was developed in which 150 native Pashto speakers (75 males and 75 females) uttered Pashto digits from sefor (0) to naha (9). SVM classifier was used for recognizing the words. Overall, 91.5% recognition accuracy was achieved by the system. Comparing the performance of SVM with KNN on the same dataset, SVM was found the best in recognizing the digits.

Khan et al. [21] presented a content-based Pashto dialect classification and retrieval using SVM. To develop the system, voice samples in form of different Pashto dialects were collected from people of different ages and genders. Cepstral coefficients and statistical parameters were extracted from the collected dataset. SVM provided the best result in accurately distinguishing between different dialects.

Nisar and Tariq [22] proposed a dialect recognition system for low resource languages, the case of Pashto. According to the authors, the traditional feature extraction techniques such as MFCC and Discrete Wavelet Transform (DWT) work better for dialect recognition of high resource languages. The same techniques offer degraded performance when applied on under resourced languages. Therefore, believing on this idea, the authors presented a new approach for Pashto (an under resourced language) dialects recognition. They used adaptive filter bank with MFCC and DWT for speech features extraction. Dialect classification was performed through HMM, SVM and KNN. The proposed method achieved overall 88.0% dialect recognition accuracy.

3. Phonetic Inventory and Dialectical Variations of Pashto

Pashto is an Indo-Iranian Language that is spoken natively in Afghanistan and Pakistan. It is a national language of Afghanistan (In 1936, it was considered and made national language of Afghanistan) and one of the regional languages of Pakistan. In Pakistan, it is widely spoken in two provinces , i.e., Khyber Pakhtunkhwa (KPK) and Baluchistan. Other than Pakistan and Afghanistan, Pashto is also spoken in various other countries across the world such as United Arab Emirates (UAE), United Kingdom (UK), United States (US), Canada, India, Singapore, Iran and Malaysia. Overall, 50 to 60 million people in the world speak Pashto [23].

Pashto language consists of 41 phonemes including 32 consonants and 9 vowels. 32 consonants have further categorized into Fricatives (F), Plosives (P), Nasals (N), Lateral (L), Spirants (S), Affricates (A), Flaps (F) and Glides (G) as shown in Table 1 [24, 25]. Table 1 also provides the detail of pronunciation of the listed Pashto consonants , i.e., Bilabial, Dental, Glottal, etc.

Table 1. Pashto consonants.

	Bilabial	Dental	Palato Alveolar	Retroflex	Uvular	Velar	Glottal
P	p b	t̪ d̪		ʈ ɖ	(q)	k ɡ	(ʔ)
F	(f) h					x ɣ	h
A		c j	cˇ jˇ				
S		s z	sˇ zˇ	xˇ gˇ			
N	m	n				ŋ	
L		l			ṇ		
F		r		ɽ			
G	w		y				

The 9 vowels are further classified into long, short vowels as well as front, back and central vowels as listed in Table 2 [26].

Table 2. Pashto vowels.

	Front vowels	Central vowels	Back vowels
Short vowels	i i:	ə	u u:
	e		
Long vowels		a a:	o

Pashto language is spoken with many dialectical variations. Northern, Southern, and Central are its three main dialectical variations [27], which have been further classified into various sub dialects as shown in Table 3 [28].

Table 3 shows that Gilji dialect is mostly spoken in Afghanistan and all the other five sub dialects are mostly spoken in Pakistan. The purpose of studying the Pashto dialectical variations was to sourt out all those regions of Pakistan and Afghanistan where Pashto is spoken with distinct dialects. After identifying the regions, the next step was to recod the pashto digits data from the speakers of the identified regions.

Table 3. Pashto main and sub dialectical variations.

Main dialects	Sub dialects	Regions of use
Sothern	Kakar	Quetta, Pashin and suburban areas of Baluchistan, Pakistan
Central	Ghilji	Kabul, Kandahar, Sanjavi and Harnai provinces of Afghanistan.
Northern	Afridi	Kohat, Darra Adam Khel, Khyber Agency, Momand Agency areas of Khyber Pukhtunkhwa, Pakistan
	Yousufzai	Peshawar, Dir, Swabi, Mardan, Swat, and Hazara Division areas of Khyber Pukhtunkhwa, Pakistan.
	Waziwola	North & South Waziristan and border areas between Afghanistan & Pakistan
	Banuchi	Bannu, Tank and suburban areas of Khyber Pukhtunkhwa, Pakistan

4. Pashto Isolated Digits Database Development

In order to develop Pashto digits database including dialectical variations of Pashto, the Pashto digits were recorded from the native Pashto speakers (Pashtoon) of the regions listed in Table 3. Ten speakers from each region were selected with their ages ranging between 15 to 55 years. Pashto isolated digits from one (yao) till ten (las) were recorded from each speaker. Zero digit of Pashto was not included in the database because, in most of the daily used communication, the zero digit is negligibly used among Pashtoon. To overcome the channel related variability, different recording devices were used, i.e., one good quality SONY voice recorder and three smart phones manufactured by different manufacturers (Huawei CUN-U29, Motorola Turbo and Samsung Note 3). Out of the 60 total speakers, 30 speakers were recorded using the voice recorder, while, the remaining 30 speakers were recorded using smart phones (each 10 speakers were recorded by each smart phone). In order to minimize the effect of background noise, all the recordings were performed in quite rooms. The baithaks (especial wide and empty rooms in Pashtoon homes designed for guests) were utilized for this purpose. To minimize the effect of seasonal variations on the collected dataset, the data was recorded at different locations and at different days and times. The overall spread of data collection was from December 2017 to December 2018. All the recordings were performed digit by digit (digits were recorded separately one by one). 16 kHz sampling frequency was used to record the digits. After recordings, the recorded digits were transferred to a core-m laptop via its USB port. In the laptop, the recorded digits were converted into .WAV format from the original MP3 format using total audio converter software. Finally, the converted digits data was saved as a binary computer files.

5. Features extraction

After the voice data has been recorded and saved, the data was processed for features extraction where spectral and prosodic features have been extracted from the data.

5.1. Spectral features

Spectral features consider the behaviour of shape and size of vocal tract, which behave differently with different sound patterns. The most important spectral features (according to the cited work in Section 2) , i.e., MFCC have been used in this study.

5.1.1. Mel frequency cepstral coefficients (MFCC)

MFCC is the most widely used technique in speech, speaker and accent & dialect recognition applications [29-31]. It uses filter banks to extract spectral information from the input speech signals. The prime advantage of using MFCC is the approximation of nonlinear frequency response of human hearing using Mel scaling, which is linearly spaced below 1000 Hz and logarithmically spaced above 1000 Hz. Due to Mel scaling, MFCC is considered, the most suitable for capturing the speech characteristics possessed by the nonlinear speech spectrum [32].

In order to extract MFCC features from input speech signals, certain steps are followed. Block diagram in Fig. 1 show the steps followed during MFCC computation.

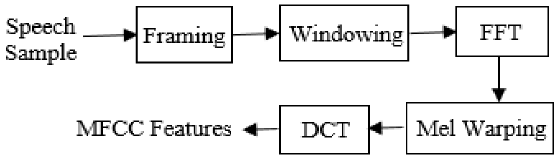

Fig. 1. MFCC features extraction process.

Each step shown in Fig. 1 is explained below.

Pre-processing: In this step, the input raw speech signal is pre-processed by separating the voiced parts of speech (where meaningful speech is present) from the unvoiced parts (where no meaningful speech is present). The purpose of separating voiced and unvoiced parts of speech is to minimize the use of resources and to save the system from complexity.

Pre-emphasising: During the pre-processing step, the signals energy is usually minimized and needed to be amplified before further processing. It is achieved by supressing low frequency components and amplifying the high frequency components contained in speech signal. For this purpose, the pre-processed speech signal is passed through a first order high pass filter. The resultant speech signal after pre-emphasizing becomes:

$$X(n) = x(n) - \alpha\,\{n-1\} \tag{1}$$

where *X (n)* is the pre emphasized form of the original discrete time speech signal $x\ (n)$. α is pre-emphasized parameter whose typical value is 0.95, in some cases 0.97 is also used.

Framing: After the pre-processing and pre-emphasizing is done, the signal is sent for analysis. It is difficult to analyse the whole speech signal at once because of its nonstationary nature; therefore, the whole speech signal is broken into small overlapping parts of 20 ms to 30 ms duration. The duration is kept smaller for the purpose that the speech signal appears to be stationary in that duration, but it should not be taken less than 20 ms because the speech signal will reduce its distinguishability among the consecutive frames.

Windowing: During this step, each frame of the speech signal is analysed using some type of window (Hamming, Rectangular or Triangular) for the purpose of removing discontinues at the beginning and the end of each frame. Hamming window is the most widely used window in speech processing, which can be described through the following mathematical relation Eq. (2).

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

where ***N*** represents the total number of overlapping blocks whose typical value is 256.

Fast Fourier Transform: After windowing, the pre-emphasized signal *X (n)* is transformed into frequency domain from time domain by simply applying Fast Fourier Transform (FFT) on it using Eq. (3).

$$y(n) = \sum_{\mathrm{n}=-\infty}^{\infty} X(n)\, e^{-jwn} \tag{3}$$

where *y(n)* is the achieved signal in frequency domain after applying FFT on *X(n)* in time domain.

Mel frequency warping: In order to calculate the average energy confined in each frame in frequency domain, each frame of the signal is passed through a series of triangular filters (Mel filters) using Eq. (4).

$$mel(F) = 2595log10(1 + f/700) \tag{4}$$

While passing the signal through Mel filters, the natural logarithm of the filter energies are obtained using Eq. (5).

$$x(m) = \log\{\textstyle\sum_{k}^{N-1} |y(n)|^2 T_m(w)\} \tag{5}$$

In Eq. (5), $T_m(w)$ represents the triangular Mel filters and *x (m)* represents the resultant signal with Mel-energies.

Discrete Cosine Transform: Finally**,** Discrete Cosine Transform (DCT) is used to transform signal frames back into time domain from the frequency domain. Equation (6) is used for the transformation and to compute Mel-coefficients that are the resultant MFCC.

$$c(n) = \sum_{m-1}^{M} x(m) \cos\{\frac{\pi n(m-\frac{1}{2})}{M}\}, m = 0, M \tag{6}$$

where *c(n)* are the MFCC obtained after DCT of the log energies of *x(m)*.

5.2. Prosodic features

Spectral features are extracted using shorter frames of the speech signals. However, the features at shorter frame level are limited to capture whole information contained in speech signal. For this purpose, some information should be captured using longer frames of speech also. Prosodic features of speech such as pitch and energy use longer frames of speech.

5.2.1. Pitch

Pitch is also known as fundamental frequency of the sound. It is directly related to the rate of vibration of vocal fold. It is a perceptual property based on which grave and shrill sounds can be easily identified. Fundamental frequency plays an important role in any identification process such as speaker identification, dialect identification and language identification because, it contains much information about speakers and their different dialects [33, 34]. For pitch estimation, subharmonic-to-harmonic ratio algorithm is used [35].

5.2.2. Energy

Different dialects of same language yield varying stress patterns due to which energy profiles of each dialect shows variations. Therefore, energy of dialects obtained at frame levels is important in dialect identification [33]. Energy of dialects can be estimated by

$$E_d(i) = \frac{1}{L}\sum_{n=1}^{L}|x_i(n)|^2 \tag{7}$$

where $E_d(i)$ is the normalized energy of the speech signal $x_i(n)$, $n = 1, \ldots.., L$, L is the frame length.

6. Multi-layer perceptron

MLP is a network of neurons, in which multiple layers of neurons operate in parallel. Basic structure of MLP is given in Fig. 2.

MLP is a popular binary classification algorithm that usually performs classification by the help of Back Propagation Feed Forward Neural Network (FFBPNN) algorithm. There are certain steps to be followed by FFBPNN whose detail is provided below.

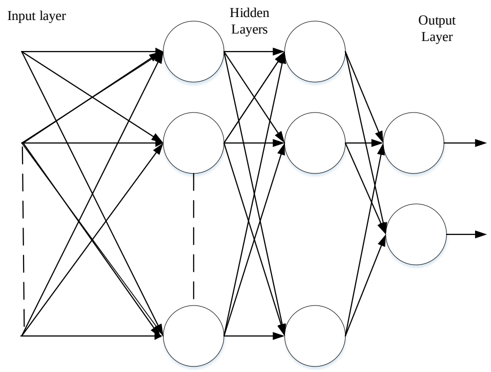

Fig. 2. Simple MLP network.

Presenting inputs: Initially, the input patterns such as the voice features' set is presented at the input layer of MLP. The MLP processes the presented inputs combined with their weights from input layer to the first hidden layer and then to the second hidden layer of neurons until the output layer is reached. The output layer produces the output patterns.

Error calculation: At the output layer, the produced output patterns are compared with the desired (target) output patterns of the neurons and an error is calculated using Eq. (8).

$$E = \frac{1}{2}\left|y_d(k) - y_p(k)\right|^2 \tag{8}$$

where *E* is the error calculated at the output of the k^{th} neuron of the output layer by simply taking the difference of its desired output y_d *(k)* and the produced output y_p *(k)*. The error of all the output neurons at the output layer is combined together to find the total output error E_i that is commonly known as the error function or cost function and can be written as

$$E_i = \frac{1}{2}\sum_{i=0}^{n}\left|y_{d(i)} - y_{p(i)}\right|^2 \tag{9}$$

The purpose of the BPFFNN algorithm is to minimize the cost function by using an iterative process of gradient descent and by updating the weights using Eq. (10).

$$\nabla E = \frac{\sigma E}{\sigma w_1}, \frac{\sigma E}{\sigma w_2}, \ldots\ldots, \frac{\sigma E}{\sigma w_n} \tag{10}$$

Equation (10) shows that the error calculated at the output layer is reduced by taking gradient and updating weights. This process can be understood by a simple three layer network of neurons , i.e., the input layer (*i*), hidden layer (*j*) and the output layer *(k*). Furthermore, it is considered that each layer contains only a single neuron with single input only. Thus, the error at the output of the output neuron (k^{th} neuron) will simply be computed by Eq. (8).

Back propagation of error: In order to minimize the error calculated at the output of the kth neuron Eq. (8), the backpropagation simply computes the gradient of the error *E*., i.e.,

$$\frac{\sigma E}{\sigma w_{jk}} = y_p(k)\,\frac{\sigma E}{\sigma y_p(k) w_{jk}} \tag{11}$$

where w_{jk} is the weight of the input of the k^{th} neuron and the output of the j^{th} neuron. The error calculated at the output layer (k^{th} neuron) is then sent back towards the hidden layer (j^{th} neuron) and then to the input layer (i^{th} neuron). During propagation of the error in backward direction, weight linked with each neuron is modified using Eq. (12).

$$w_{+i} = w_i - \Delta\, w_i \tag{12}$$

where, for the i^{th} neuron, w_{+i} is the updated weight, w_{i} is the previous weight and Δw_i is the weight modification factor, which can be computed using Eq. (13).

$$\Delta w_i = \alpha\, O_j \delta_i \tag{13}$$

where α is learning rate and the δ_i is the error gradient of the i^{th} neuron.

Producing desired outputs: The modified input patterns are again sent to the output layer at which again error is calculated and sent back to the input. At input, again the weights are modified. This process is repeated until a predefined desired level of error is achieved.

7. Experimental Results

The current research presents an isolated digits recognition system to recognize Pashto isolated digits from yao (1) to las (10). As discussed earlier, the Pashto speech data in form of isolated digits was collected from 60 Pashto speakers in different native dialects. Once the speech data has been collected, the MFCC features have been extracted from it. Table 4 lists the MFCC parameters, which have been used to extract the MFCC features. Table 4 also provides the comparison of the MFCC parameters used by the recently proposed research on speech recognition based on Pashto isolated digits.

Table 4. Comparison of MFCC parameters.

Parameters used in this research		Parameters used by Ali et al. [18]	Parameters used by Abbas et al. [36]
Linear filters	17	11	11
Log filters	19	12	13
Log spacing	100	100	1.148
Window size	128	128	128
FFT size	512	512	512
Sampling rate (Hz)	16000	8000	8000
Cepstral coefficients	19	12	13

Table 4 shows, for calculating MFCC, we increased the linear filters, cepstral coefficients and sampling rate as compared to [18] and [36]. The purpose was to reduce the amount of low frequency components, determining more magnitudes and to increase the overlapping widow during framing of speech signal respectively.

After extracting MFCC features, the pitch and energy features were also extracted. Once the features have been extracted from all the collected voice samples, an MLP classifier was designed for classification of digits. Table 5 elucidates the initial design parameters of the MLP classifier.

Table 5. MLP initial design parameters.

Parameters	Assigned values
Learning Rate	0.3
Momentum	0.2
Hidden layers	01
Epochs	500
Number of neurons in input layer	600
Number of neurons in hidden layer	500
Number of neurons in output layer	01

The purpose of settling initial design parameters was to vary them until the highest recognition accuracy is achieved. Designed MLP classifier was then used to classify the Pashto isolated digits using the obtained feature vectors.

During classification, for training and testing splits of data, ten-fold cross validation technique was used. Using this technique, the system splits the input feature vectors into ten equal parts out of which, nine parts are used for training and the remaining one is used for testing. The process is repeated until all the parts have been tested.

Features extraction algorithms were implemented in MATLAB, while, MLP was run in WEKA. Table 6 provides the detail of the results achieved during the performed experiment.

Table 6. Classification result of MLP classifier.

Correctly classified instances	Incorrectly classified instances	Total number of instances	Recognition accuracy (%)
585	15	600	97.5

Table 7 is the confusion matrix of MLP classifier during testing. It was hard to draw the original 60 X 60 confusion matrix, therefore, only the confusion matrix of the misclassified instances has been drawn.

Table 7. Confusion Matrix of the misclassified speakers.

Classified As Speakers	**1**	**10**	**23**	**40**	**55**	**59**
16	1	0	2	0	0	0
5	0	1	0	0	0	1
11	0	0	2	0	0	0
7	0	0	0	3	0	0
60	1	0	0	0	3	0
43	0	0	0	0	0	1
Overall Recognition						**585/600**

Table 7 shows that speaker 16 once misclassified as speaker 1 and twice misclassified as speaker 23, speaker 5 once misclassified as speaker 10 and once misclassified as speaker 59, speaker 11 twice misclassified as speaker 23, speaker 7 three times misclassified as speaker 40, speaker 60 once misclassified as speaker 1 and three times misclassified as speaker 55 and speaker 43 once misclassified as speaker 59 by the MLP classifier. Rest all the speakers were 100% accurately classified by the classifier, which have not shown in Table 7. Out of total 600 instances, 585 were correctly classified hence, overall, 97.5% recognition was achieved.

In order to evaluate the performance of the proposed MLP classifier, several popular Machine Learning (ML) performance measuring functions have been used such as Precession, Recall, F-measure, True Positive Rate (TPR), False Positive Rate (FPR), Mathews Correlation coefficients (MCC) and area under ROC curve (AUC). Table 8 reports all these evaluation parameters achieved in the conducted experiment. Table 8 shows that the values achieved of the evaluation parameters are quite satisfactory.

Table 8. Evaluation measures achieved in the conducted experiment.

Class	TPR	FPR	Precession	Recall	F-measure	MCC	AUC
Weighted Average	0.968	0.008	0.971	0.968	0.968	0.961	0.999

8. SVM and HMM Based Classifiers

SVM and HMM were also used to classify the Pashto digits using the collected dataset. The purpose was to test those classifiers on the collected dataset, which have been mostly used for Pashto based speech recognition systems (in literature) and to compare the results. Table 9 provides the recognition accuracy achieved by SVM and HMM over the collected dataset.

Table 9. Recognition accuracies of SVM and HMM classifiers in recognizing Pashto isolated digits.

Classifier	Recognition Accuracy (%)
SVM	98.3
HMM	95.0

Table 9 shows that SVM classifier achieved better accuracy, whereas, HMM achieved reduced accuracy as compared to MLP (refer Table 6). It is because SVM uses Kernel functions for classification. Because of using Kernel functions, it has the capability to classify dialects/speakers/words/digits etc. even if these are not clearly distinguishable from each other. Even though the Kernel function has the capability of classification in case of mixed boundaries still its use is not always trusted to draw a valid conclusion because it results in overfitting the data. Some special techniques are required to prevent data overfitting during classification [37].

As compared to SVM, MLP does not result in overfitting because it does not use Kernel function rather use some type of nonlinear activation functions such as sigmoid, which make MLP capable to classify the candidates (speakers, digits or dialects etc.) in case if these are overlapping and are not separated from each other by clear boundaries [38]. Even though, MLP has an advantage over SVM in form of not overfitting the data, still SVM produces better recognition accuracy as compared to MLP because of using Kernel functions. The same case has observed here in our study where SVM achieved better recognition accuracy as compared to MLP.

9. Results with spectral and prosodic features and with their combination

Keeping in view the fact that the dialect related information from the speech cannot be fully covered through using spectral features only, the MLP, HMM and SVM classifiers were also trained by the prosodic features (detail of prosodic features has been provided in section 5) and then by the combination of prosodic and spectral features. Fig. 3 provides the comparison of the recognition accuracies using prosodic and sum of prosodic and spectral features.

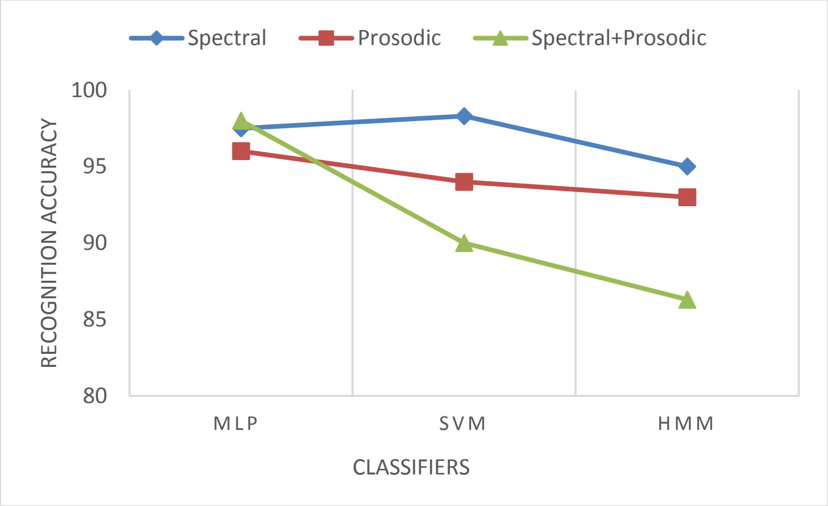

Fig. 3. Comparison of recognition accuracies achieved by MLP, HMM and SVM classifiers using spectral features, prosodic features and their sum.

It is shown in Fig. 3 that the accuracy of the MLP classifier improved by 0.5% when the combination of prosodic and spectral features is used. Furthermore, the accuracies of HMM and SVM classifiers could not be improved more.

For better elaboration of the achieved results, the performance of the proposed system is compared with some of the recently proposed speech and dialect recognition systems designed for Pashto language in literature. Comparative study shows that the results achieved by the proposed system outperformed some of the recently proposed Pashto based systems. Detail of the comparison is provided below.

10. Comparison with the recently proposed systems in literature

Following is the summary of the comparative results of the proposed system with some of the recently proposed related studies from literature.

- The proposed system outperformed MFCC-KNN based isolated Pashto spoken digits recognition system proposed in literature by Ali et al. [18] by showing

20.7% enhancement in recognition accuracy. In proposed system, different MFCC parameters were used as compared to [18] (refer Table 4).

- The proposed system showed 3.7% improvement in recognition accuracy over MFCC-SVM based dialect recognition system designed for Pashto language proposed by Khan et al. [21].
- Another MFCC-SVM, MFCC-HMM based dialect recognition system for Pashto language designed by Nisar and Tariq [22] was outperformed by the proposed system through showing 9.5% enhancement in recognition accuracy.

11. Conclusion and Future Work

In this research, a Pashto speech recognition system based on Pashto isolated spoken digits has been proposed. The proposed system is different from the previous systems in literature because, the speech data in form of isolated spoken digits of Pashto was collected in form of different dialects of Pashto. There were two reasons for including dialectical variations

- To reduce the accent and dialect related variability from the designed system (because of dialectical variations, the performance of most of the speech recognizers greatly degrade).
- Dialectical variations are important to include because, based on such variations, region of origin of speakers can be recognized by simply knowing the area where the particular dialect is spoken. This application can be useful in many security applications.

In the proposed system, Pashto spoken digits (from 1 to 10) were collected from sixty native speakers of six different regions of Afghanistan and Pakistan where Pashto is spoken natively with different dialectical variations. MFCC, pitch and energy features were extracted from the collected speech samples and MLP of ANN was used to classify the Pashto digits. The system achieved overall 97.5% recognition accuracy in recognizing Pashto isolated digits.

The system's performance was compared with SVM and HMM classifiers tested on same collected speech dataset. SVM classifier showed improved performance, while, HMM showed degraded performance over the proposed MLP classifier. When the classification performance of MLP, SVM and HMM was tested on spectral (MFCC), prosodic (Pitch and Energy) and their combination, The result achieved by MLP was 0.5% improved, whereas, no further improvement in the results of SVM and HMM was observed.

The proposed system's performance was also compared with recently proposed Pashto speech and dialect recognition systems in literature. The comparative study shows that the proposed system achieved much better recognition performance as compared to [18, 21, 22].

The Speech dataset for this research was collected with self-efforts. Collecting data from Afghani Pashtoon (Pashto native speakers from Afghanistan) was one of the challenging phase of the research because of the tense situation of Afghan and Pakistan Authorities. It was main reason of collecting the speech from less number of speakers.

In future, the speech data will be collected on large scales in which gender balance will also be taken into account. The other state of the art classifiers such as

Deep Neural Network (DNN) and identity vectors will also be tested on the collected data set.

References

1. Rehmam, B.; Halim, Z.; Abbas, G.; and Muhammad, T. (2015). Artificial neural network-based speech recognition using dwt analysis applied on isolated words from oriental languages. *Malaysian Journal of Computer Science*, 28(3), 242-262.
2. Heigold, G.; Nguyen, P.A.P.; Weintraub, M.; and Vanhoucke, V.O. (2014). *Speech recognition process*. Google Inc, U.S. Patent 8,775,177.
3. Sharma, K.; and Singh, P. (2015). Speech recognition of Punjabi numerals using synergic HMM and DTW approach. *Indian Journal of Science and Technology*, 8(27), 1-6.
4. Gamit, M.R.; and Dhameliya, K. (2015). English digits recognition using MFCC LPC and Pearson's correlation. *International Journal of Emerging Technology and Advanced Engineering*, 5(5), 364-367.
5. Vimala, C.; and Radha, V. (2015). Isolated speech recognition system for Tamil language using statistical pattern matching and machine learning techniques. *Journal of Engineering Science and Technology (JESTEC)*, 10(5), 617-632.
6. Yusnita, M.A.; Paulraj, M.P.; Yaacob, S.; and Shahriman, A.B. (2012). Classification of speaker accent using hybrid DWT-LPC features and K-nearest neighbors in ethnically diverse Malaysian English. *Procedings of IEEE International Conference on Computer Applications and Industrial Electronics (ISCAIE)*, 179-184.
7. Najafian, M.; Safavi, S.; Hansen, J.H.; and Russell, M. (2016). Improving speech recognition using limited accent diverse British English training data with deep neural networks. *Proceedings of 26th IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 1-6.
8. Mostefa, D.; Choukri, K.; Brunessaux, S.; and Boudahmane, K. (2012). New language resources for the Pashto language. *Proceedings of LREC*, 2917-2922.
9. Ahn, T.Y.; and Lee, S.M. (2016). User experience of a mobile speaking application with automatic speech recognition for EFL learning. *British Journal of Educational Technology*, 47(4), 778-786.
10. Mufungulwa, G.; Tsutsui, H.; Miyanaga, Y.; Abe, S.I.; and Ochi, M. (2017). Robust speech recognition for similar Japanese pronunciation phrases under noisy conditions. *Proceedings of IEEE International Symposium on Signals, Circuits and Systems (ISSCS)*, 1-4.
11. Zhou, S.; Dong, L.; Xu, S.; and Xu, B. (2018). Syllable-based sequence-to-sequence speech recognition with the transformer in Mandarin Chinese. *ArXiv Preprint ArXiv*, 1804.10752.
12. Gupta, V.; Kenny, P.; Ouellet, P.; and Stafylakis, T. (2014). I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6334-6338.

13. Ali, H.; Jianwei, A.; and Iqbal, K. (2015). Automatic speech recognition of Urdu digits with optimal classification approach. *International Journal of Computer Applications*, 118(9), 1-5.
14. Pashto language, alphabet and pronunciation - Omniglot. Retrieved January 18, 2019,from https://omniglot.com/writing/pashto.htm.
15. Shah, S.M.; Memon, S.A.; Khoumbati, K.U.; and Moinuddin, M. (2017). A Pashtu speakers database using accent and dialect approach. *International Journal of Applied Pattern Recognition*. 4(4), 358-380.
16. Ahmed, I.; Ahmad, N.; Ali, H.; and Ahmad, G. (2012). The development of isolated words Pashto automatic speech recognition system. *Proceedings of IEEE 18th International Conference on Automation and Computing (ICAC),*1-4.
17. Tanzeela; Abbas, A.W.; Ali, Z.; and Uddin, B. (2014). Analyzing the impact of MFCC and LDA for the development of isolated Pashto spoken numbers ASR. *Proceedings of IEEE 12th International Conference on Frontiers of Information Technology (FIT),* 350-354.
18. Ali, Z.; Abbas, A.W.; Thasleema, T.M.; Uddin, B.; Raaz, T.; and Abid, S.A.R. (2015). Database development and automatic speech recognition of isolated Pashto spoken digits using MFCC and K-NN. *International Journal of Speech Technology*, 18(2), 271-275.
19. Nisar, S.; and Asadullah, M. (2017). Home automation using spoken Pashto digits recognition. *Proceedings of IEEE International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT),* 1-4.
20. Nisar, S.; Shahzad, I.; Khan, M.A.; and Tariq, M. (2017). Pashto spoken digits recognition using spectral and prosodic based feature extraction. *Proceedings of 9th IEEE International Conference on Advanced Computational Intelligence (ICACI)*, 74-78.
21. Khan, S.; Ali, H.; Ullah, K. (2017). Pashto language dialect recognition using mel frequency cepstral coefficient and support vector machines. *Proceedings of IEEE International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT),* 1-4.
22. Nisar, S.; and Tariq, M. (2018). Dialect recognition for low resource language using an adaptive filter bank. *International Journal of Wavelets, Multiresolution and Information Processing*. 16(4), 1850031.
23. Omniglot-Pashto. Retrieved January 18, 2019, from https://www.omniglot.com/writing/pashto.htm.
24. Iqbal, M.; and Rahman, G. (2017). *A comparative study of Pashto and English consonants*. Hazara University, Pakistan.
25. Farooq, M. (2015). *An acoustic phonetic study of six accents of Urdu in* Pakistan*: M. Phil Thesis*. Department of Language and Literature, University of Management and Technology, Johar Town, Lahore, Pakistan.
26. Ijaz, M. (2003*). Phonemic inventory of Pashto, CRULP, annual student* report. National University of Computer and Emerging Sciences, Center for Research in Urdu Language Processing (CRULP), B Block, Faisal Town, Lahore, Pakistan.
27. Tegey, H.; and Robson, B. (1996). *A reference grammar of Pashto*. Centre for Applied Linguistics, Washington DC.

28. Pashto dialects-Wikipedia. Retrieved January 18, 2019, from https://en.wikipedia.org/wiki/Pashto_dialects.
29. Ittichaichareon, C.; Suksri, S.; and Yingthawornsuk, T. (2012). Speech recognition using MFCC. *Proceedings of International Conference on Computer Graphics, Simulation and Modeling (ICGSM)*, 28-29.
30. Tiwari, V. (2010). MFCC and its applications in speaker recognition. *International journal on emerging technologies*, 1(1), 19-22.
31. Rao, K.S.; and Shashidhar, G.K. (2011). Identification of Hindi dialects and emotions using spectral and prosodic features of speech. *International Journal of Systemics, Cybernetics and Informatics,* 9(4), 24-33.
32. Muda, L.; Begam, M.; and Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *International Journal of Computing*, 2(3), 138-143.
33. Etman, A.; and Louis, A.A. (2015). American dialect identification using phonotactic and prosodic features. *In IEEE SAI Intelligent Systems Conference*, 963-970.
34. Das, H.S.; and Roy, P. (2019). 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*, 22(1), 67-77.
35. Sun, X. (2000). A pitch determination algorithm based on subharmonic-to-harmonic ratio. *In Sixth International Conference on Spoken Language Processing,* 676–679.
36. Abbas, A.W.; Ahmad, N.; and Ali, H. (2012). Pashto Spoken Digits database for the automatic speech recognition research. *Proceedings of 8th IEEE International Conference on Automation and Computing (ICAC*), 1-5.
37. Lameski, P.; Zdravevski, E.; Mingov, R.; and Kulakov, A. (2015). SVM parameter tuning with grid search and its impact on reduction of model over-fitting. *In Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, 464-474.
38. Tsapanos, N.; Tefas, A.; Nikolaidis, N.; and Pitas, I. (2019). Neurons with paraboloid decision boundaries for improved neural network classification performance. *In IEEE Transactions on Neural Networks and Learning Systems*, 30(1), 284-294.