Journal of Big Data

RESEARCH

Open Access

Development of a regional voice dataset and speaker classification based on machine learning

Muhammad Ismail[1,2*], Shahzad Memon[1], Lachhman Das Dhomeja[1], Shahid Munir Shah[3], Dostdar Hussain[2], Sabit Rahim[2] and Imran Ali[2]

*Correspondence:
muhammad.ismail@kiu.edu.pk
[2] Department of Computer Science, Karakoram International University, Gilgit, Pakistan
Full list of author information is available at the end of the article

Abstract

At present, voice biometrics are commonly used for identification and authentication of users through their voice. Voice based services such as mobile banking, access to personal devices, and logging into social networks are the common examples of authenticating users through voice biometrics. In Pakistan, voice-based services are very common in banking and mobile/cellular sector, however, these services do not use voice features to recognize customers. Therefore, the chance to use these services with false identity is always high. It is essential to design a voice-based recognition system to minimize the risk of false identity. In this paper, we developed regional voice datasets for voice biometrics, by collecting voice data in different local accents of Pakistan. Although, there is a global need for voice biometrics especially when voice-based services are common, however, this paper uses Pakistan as a use case to show how to build regional voice dataset for voice biometrics. To build voice dataset, voice samples were recorded from 180 male and female speakers with two languages English and Urdu in form of five regional accents. Mel Frequency Cepstral Coefficient (MFCC) features were extracted from the collected voice samples to train Support Vector Machine (SVM), Artificial Neural Network (ANN), Random Forest (RF) and K-nearest neighbor (KNN) classifiers. The results indicate that ANN outperformed SVM, RF and KNN by achieving 88.53% and 86.58% recognition accuracy on both datasets respectively.

Keywords: Speaker recognition systems, Speakers classification, Voice database, Accents and dialects

Introduction

For identification and verification, human body characteristics like voice, face, fingerprint, and gait etc. have been used since long ago [1]. Such characteristics are preferably used for biometric identification because these satisfy the desirable properties of biometrics such as universality, distinctiveness, permanence and collectability [2]. Biometric identification is based on biometric traits, which broadly fall into two categories i.e. physiological biometric traits (fingerprint, face, iris, vein, ear, DNA, etc.) and behavioral biometric traits (voice, key strokes dynamics, signature, and gait etc.). These individual traits are unique and are often difficult for scammers' to replicate.

Hence biometric approaches may provides superior security and convenience than recognition techniques based on PIN, passwords and identity cards [3, 4]. Biometric traits can be used in various applications such as ATM, credit cards, physical access control, cell phone, national ID cards, passport control, driver licenses, dead body identification, and criminal investigation, etc. [5]. Each biometric has its own features and limits. The selection of a biometric particularly depends on the applications for which it is being used. No single biometric is optimal and nor it efficiently fulfills all the requirements of various applications. For example, in some situations, the fingerprint biometric trait is more desirable than the voice biometric trait. In another situation, the voice biometric is preferable than finger print, such as access control for bank transactions via cell phones or landline telephones, voice mails and verification of credit cards, distant access to computers through a modem on the dial-up telephone line in call-centers and forensic applications where speaker recognition is required [4, 6]. The human voice carries different characteristics such as the meaning/words a speaker wants to pass to a listener, spoken language information, emotions, gender, identity, health and speaker's age-related information etc. [6]. The objective of speaker recognition is to extract information about the speaker's identity and based on that information it recognizes the speaker [7]. Speaker recognition is usually subdivided into speaker verification and speaker identification tasks. The speaker verification is the task of verifying a claimed person from his/her voice and verification system must perform a 1:1 comparison hence the cost of computation is independent of the records in the voice database. On the other hand, the speaker identification task is to determine the specific speaker speaking from a speaker's database. In this task the unknown person does not claim identity and there must be 1:*N* comparisons. In this way, the cost of computation depends on the number of records in the voice database [8].

An important step in designing Voice Based Systems (VBS) i.e. speech and speaker recognition systems is the voice database design. A comprehensive voice database plays significant role for design and development of VBS specifically designed for particular applications, same is the case here (refer to "Database" section for further detail about the designed database). In this research, we developed an Urdu and English languages based voice dataset particularly designed for voice based customer verification services in banking sector in Pakistan. To the best of our knowledge, there is no such voice dataset available, which has been particularly designed for the banking sector in Published literature.

During the database design we particularly focused on minimizing different performance degrading factors/variability mostly encountered in voice databases such as background noise, channel mismatch, speaker's age, health & emotions etc.

Like the other performance degrading factors, different accents and dialects of a language can also cause performance degradation in VBS [9]. Typically, VBS do not perform well if the accent of a speaker, who is going to be recognized, is different from the accent of the speakers by whom the system was trained. Incorporation of accents can minimize the variability caused by different accents of a language, which in turn enhances the performance of the recognition system [10].

Based on the above mentioned detail about VBS and their possible performance degradation factors, here in this research, a voice database (containing Urdu and English datasets) has been designed and tested on several popular ML methods (KNN, RF, SVM and ANN in our case). The specific contributions of the presented research are:

- Designing a voice database for the five regional accents of Urdu and English spoken in Pakistan.
- Including accents variations in the designed database so that it could be used to design robust Speaker Recognition Systems (SRS) based on this dataset.
- Database design to particularly help banking sector applications in Pakistan.
- Database design based on a questionnaire/script containing question answers particularly asked by bank representatives for authentication during voice calls in case of lost/theft of credit/debit cards or any other query with the banks in Pakistan.
- Testing KNN, RF, SVM and ANN algorithms on the designed dataset.

The other aforementioned performance degrading factors (other than different accents and dialects) have also been taken into consideration during database design, however, to fully address all the performance degrading factors from a single designed system is still a challenging task, which is one of the limitation of the present research. Further detail of the development of the voice database is presented in "Database" section. Related research is presented in "Related research" section. Methodology is presented in "Methodology" section. Results and discussions are provided in "Results and discussion" section. Finally in 'Conclusion" section, conclusion is provided.

Related research

This section provides a review of some of the recent SRS particularly robust against different types of variability (performance degrading factors like accents and dialects of a language) present in them.

Development of voice datasets is an important aspect of designing SRS. Particularly, SRS designed for specific applications need their own voice dataset to develop because such datasets are usually not available, same is the case here in this research.

Some of the variability present in SRS (room reverberation, channel mismatch, seasonal variations) are tried to be addressed during voice database design. However, most of them (accents and dialects, emotions, background noise) need to be addressed during the design and development phase of the SRS [11, 12].

National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) [13] is an ongoing series of developing speaker recognition and exploring new promising ideas in this field. The task specifically includes voice database design and development of state of the art SRS robust against different variability/mismatch conditions.

RedDot [14] is a project to collect speech data over mobile devices for speaker recognition. The designed database contains speech data of 45 English speakers (both native and non-native) from 16 countries. The content of the database consists of a short duration test utterances with variable phonetic content. The main focus for the RedDots database was to include a high degree of inter-speaker variations and intra-speaker variations. To achieve the main focus, the speakers were selected worldwide and data was collected from speakers in 91 different sessions.

Although RedDot and NIST speaker recognition evaluations provide great opportunity to the research community to use their datasets and to evaluate their systems but their focus was mostly towards the controlled data collection over mobiles or landline telephones, which restricts the dimensions of different variability present in the data [15].

To fill some of these gaps present in the previous datasets (particularly NIST-SRE), a new voice dataset "Speakers in the Wild (SITW)" was created [16].

SITW [15] is a speaker recognition database specifically collected for the text-independent speaker recognition applications. This database contains audio recordings of 299 speakers that were collected from open source media, with an average of 8 sessions per speaker. The audio recordings in database include unconstrained/wild acoustic conditions like background noise, reverberation as well as large intra-speaker variability. The database also contains audios of speakers in different scenarios like interview, dialog or uncontrolled conditions where multiple speakers are involved. SITW filled much of the gaps present in the previous datasets like NIST but because of manually annotations, its size was quite small.

VoxCeleb2 [17] is a very large-scale audio-visual speaker recognition publicly available dataset collected from open-source media, which contains over a million utterances from over 6000 speakers. Along with its large size, the dataset specially focused to improve the limitations exhibited by the most of the previous datasets like recording under controled conditions [18–22] limited in size because of manually annotations, and not freely available to the research community [18, 23].

Accent and dialect classification systems have also provided a great help in solving accent and dialect related variability and designing robust SRS [24].

A weighted accent classification system has been designed by using multiple words [25]. In this research, extreme learning machines (ELMs) and SVMs were used to the problem of accent/dialect classification on TIMIT dataset. The TIMIT comprises utterances from 630 speakers indicating eight different dialect regions of the United States.

Design of regional accented databases as well as accent and dialect classification is another important area of research that helps in improving performance of speaker and speech recognition systems designed for specific regions.

An Algerian Speech corpus [26] was designed to support research in speech recognition. The corpus represents 300 Algerian native speakers who could speak Modern Standard Arabic (MSA) language. The speakers were selected from 11 different regions of Algerian, with both genders (148 males and 152 females), with different age groups and with different educational levels (primary school to postgraduate level). Finally, using a subset of the collected dataset, the author has designed a text-independent ASR system using Hidden Markove Model (HMM) that achieved a 91.65% recognition rate.

Shah et al. [10] designed a voice database considering the same strategy i.e. the voice samples were recorded using different recording devices to minimize the channel variability, quite rooms were used to record the voice samples to minimize the effect of the room reverberation, data was recorded with different sessions spread over ample amount of time to cope with the seasonal variations' effect on the collected dataset as well as a particular script containing the words spoken differently in different accents and dialects to make the designed dataset robust against accentual and dialectical variations present in Pashtu language. On the other hand, Gaussian noise was added in the extracted feature vectors to make the designed SRS robust against background noise.

An accent classification system for classifying the regional accent of Philippine was presented by Dano et al. [27]. Voice data was collected from 150 native residents of Philippine. MFCC features were extracted to train MLP and k-NN classifier. In this research, MFCC-MLP based accent classification system outperforms k-NN by achieving recognition accuracy of 56.19%.

A small scale database of 11 speakers including 7 males and 4 females, with their age ranging from 19–36 years was collected by Thullier et al. [28]. All of the speakers were native French. Some speakers from all had unique Canadian French accents and Hexagonal French accents. The voice samples were collected in a silent room (university meeting room) as well as in a noisy environment (i.e. University cafeteria).

Based on the above discussion and the importance of designing regional voice databases, here in this research we designed a voice database for five regional accents (spoken in Gilgit Biltistan (GB), Pakistan) of Urdu and English, which are the important regional accents of Pakistan. The next section describes the database development phase of our research.

Database

This paper presents development of voice datasets used to evaluate the performance of various ML algorithm. The datasets contain 7200 voice samples of 180 different speakers of GB region of Pakistan. The proposed datasets add diversity to existing datasets in terms of different local accents. Datasets development consists of various steps, each step is explained below.

A regional voice dataset collection with accent variation

To design Urdu and English voice datasets with various accentual and dialectical variations, the Gilgit-Baltistan (GB) Provence of Pakistan has been selected. GB is an important region of the China-Pakistan Economic Corridor (CPEC) and has great diversity in voice accents. It borders with Azad Kashmir, Jammu Kashmir, Khyber Pakhtunkhwa, Afghanistan, and China. GB is an area of high mountains and has an area of over 72,496 km^2. The capital city of GB is Gilgit and the population of GB is about 2.0 millions. There are ten districts in GB i.e. Gilgit, Nagar, Hunza, Ghizer, Astore, Skardu, Diamer, Ghanche, Shigar, and Kharmang. The people of this region have different native languages and have different cultures and backgrounds. There are five different native languages spoken in these districts, which are Shina, Balti, Burushishki, Khuwar and Wakhi [29, 30]. Speakers of each of these districts speak their standard native Languages (which are used as a mean of official communication in offices as well as are communicated on radio broadcasts) with different accents. Based on these five regional accents, several speakers were chosen from each district as indicated in Table 1.

Table 1 District wise speaker selection

Districts	Accent	Language	Participants		
			Male	Female	Total
Gilgit, Hunza, Nagar, Ghizer, Astore, Diamer, Skardu, Ghanche, Shigar, Kharmang	Shina	English and Urdu	33	33	66
Skardu, Shigar, Kharmang. Ghanche	Balti	English and Urdu	21	21	42
Gilgit, Hunza, Nagar, Ghizer,	Burushiski	English and Urdu	18	18	36
Hunza, Ghizer	Wakhi	English and Urdu	9	9	18
Hunza, Ghizer	Khuwar	English and Urdu	9	9	18
Total participants 180					

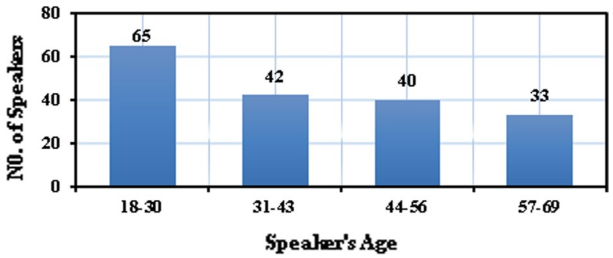

Fig. 1 Age-wise speakers' distribution

Age distribution of the speakers

For the design of voice database, the voice data was collected from male and female speakers with different ages ranging from 18–69 years. The purpose of selecting speakers with different age groups is to include acoustic variation, which arises in the voice of speakers at different stages of age to cover maximum telephone banking customers and mobile users. The Age-wise distribution of the speakers is shown in the Fig. 1

Design of script for speakers

The voice data was collected from each selected speaker based on two specifically designed scripts in Urdu and English languages. The scripts contain all possible conversational talk between a phone banking officer/mobile call center agent and their customers. These scripts contain sentences in the form of words, 10–16 digit strings, and the speaker's personal information mostly related to bank and mobile network services. All together 20 sentences with average time duration ranging between 10 and 100 ms were included in each of the scripts. The scripts were provided to each speaker to read for recording data. The designed written script for the English language is shown in Table 2, whereas, the same script was translated into Urdu for the recording of voice samples in Urdu language.

Table 2 Designed written script for English

S. No	Authentication questions by a phone banking officers	Recorded customer's response
1	Asalamualaikum	Asalamualaikum
2	What is your Name?	My name is: ———
3	Where are you calling from ?	I am calling from: ———
4	What is your father's name?	My father name is: ———
5	What is your mother's name?	My mother name is: ———
6	What is your National Identity Card (NIC)?	My NIC number is: ———
7	What is your postal address?	My postal address is: ———
8	What is your mobile number?	My mobile number is: ———
9	Is your mobile registered with this bank?	Yes/No: ———
10	What is your current location?	My current location is: ———
11	What is your account number?	My account number is: ———
12	What is your debt card number?	My debt card number is: ———
13	What is your credit card number?	My credit card number is: ———
14	What is the expiry date of your debit card?	The expiry date of my debit card is: ———
15	What is the expiry date of your credit card?	The expiry date of my credit card is: ———
16	What is the secret code of your debit card?	The security code of my debit card is: ———
17	What is the secret code of your credit card?	The security code of my credit card is: ———
18	What is the expiry date of your NIC?	The expiry date of my NIC card is: ———
19	What is your occupation?	My occupation is: ———
20	What is your Date of Birth?	My Date of Birth is: ————

Voice data recording environment and device allocation

The data was recorded from speakers in the university office, seminar room, and rest room using different smartphones and landline. The specification of the smart phones which have been used for data recording is as follows.

1. Huawei P8, CPU Octa-core 1.2 GHZ, 2.0 GB RAM, 16.00 GB internal memory and Android version 6.0
2. Oppo A371W, Processor Qualcomm snapdragon 410 quad-core processor msm8916, 16 GB internal memory, 2 GB RAM and OS version lollipop 5.1.1
3. Samsung S6 (Samsung-sm-g920v), 32 GB internal memory, and Android version 7.0.
4. iPhone X, CPU Hexa core, 256 GB internal memory, 3 GB RAM
5. Micromax Q349, 16 GB internal memory, 2 GB RAM, and Android version 6.0.
6. Landline

The voice data was recorded from a total of 180 speakers. The speakers were sub divided into 6 groups where each group contained 30 speakers. All the groups were assigned a different specific device for recording.

Recording sessions

The voice data was collected in seven recording sessions with a gap of at least one month. Figure 2 depicts recording sessions and their corresponding speakers. The

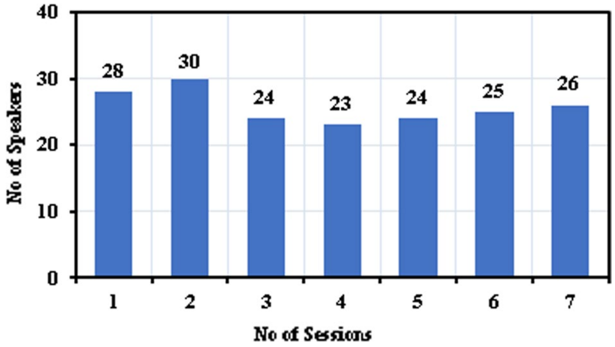

Fig. 2 Recording sessions with corresponding speakers

purpose of recording data in different sessions was to track the effects of intersession variability on ASR performance [31].

Recording of voice samples

The designed scripts (Urdu and English) were provided to each speaker who was selected in a particular session for recording. Before the start of each session, each speaker was communicated on how to record their voices, and afterword, they were supposed to rehearsal for a short period. Finally, the data was collected sentence by sentence according to the script. After the recording of each sentence, the recorded sample was verified by just replaying the recorded sentence to ensure the acquisition of appropriate sample. Since the scripts contained 20 different sentences each, therefore, each speaker recorded 20 separate sentences for the Urdu language as well as for English. A total of 3600 (20 * 180) voice samples have been recorded for the English Language. Similarly, a total of 3600 (20 * 180) voice samples have been recorded for the Urdu language. So overall a total of 7200 (3600 + 3600) voice samples have been recorded. All the recorded samples are then transferred to a laptop and converted to .wav from the default format of the allocated devices using audio converter (4dots software) for further processing. The voice samples were recorded in a systematic way as shown in Fig. 3. As per Fig. 3, the designed scripts were distributed to the speakers selected for a particular session for recording. During a practice session, the speakers were given instructions on how to read the script and making them familiar with the acquisition process. It was a kind of practice session before the actual recording. Afterwards, the speaker's voice samples were recorded sample by sample. Each sample was cross-checked with the script to ensure the consistency of the acquired voice sample with the script. All consistent samples were kept as voice database and inconsistent samples were discarded and the process was continuing until the collection of all voice samples.

Methodology

Voice samples preprocessing

After data collection, the collected voice samples were pre-processed and features were extracted using MFCC. Pre-processing of speech plays an important role in the development of an efficient automatic speech/speaker recognition systems. Pre-processing is an important step for ML algorithms to produce better results. In speech

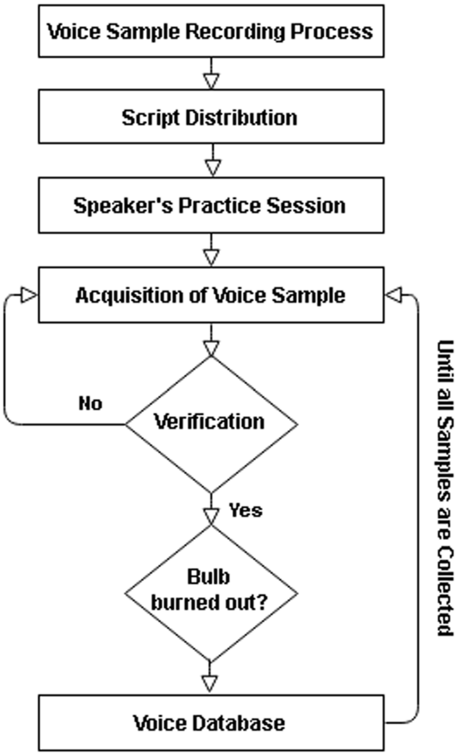
Voice Sample Recording Process
Script Distribution
Speaker's Practice Session
Acquisition of Voice Sample
No
Verification
Yes
Bulb burned out?
Voice Database
Until all Samples are Collected

Fig. 3 Voice samples recording process

processing, the pre-processing includes noise cancellation, pre-emphasis and silence removal. Pre-processing facilitates the voice-based recognition systems to be computationally efficient [32]. Due to the characteristics of the human vocal system, glottal airflow and lip radiations depress higher frequency components of the voiced part of the sound signal. For the voiced sound signal, the glottal pulse has a slope of approximately − 12dB/octave, and the lip radiation has a slop of approximately + 6dB/octave. Resultantly sound signal introduces a slope of − 6 dB/octave down-ward if compared with the spectrum of vocal tract. The process to remove the slope of − 6 dB /octave is known as pre-emphasis and it removes the effects of the glottal pulse from actual vocal tract and balance power spectrum dynamics. Thus, all voice samples have been passed through a high pass filter. It amplifies high-frequency components with respect to low-frequency components. The pre-emphasis method ensures that all formants of the voice signal have identical amplitude so that have equal importance in subsequent processing steps [33]. After pre-emphasis, the voice samples were further processed to remove silence using amplitude based silence removal technique [34] This technique divides the whole audio sample into components of short fixed length called frames and calculates maximum amplitude of each frame. It then finds those frames with the maximum amplitude is greater than 0.03 and considers those frames as voice portion of the speech and discards the frames with lesser amplitude then 0.03. This technique assumes that the silent part of the voice signal has amplitude < 0.03 and the voice part of voice signal contains amplitude > 0.03. The pre-processing is shown in Fig. 4.

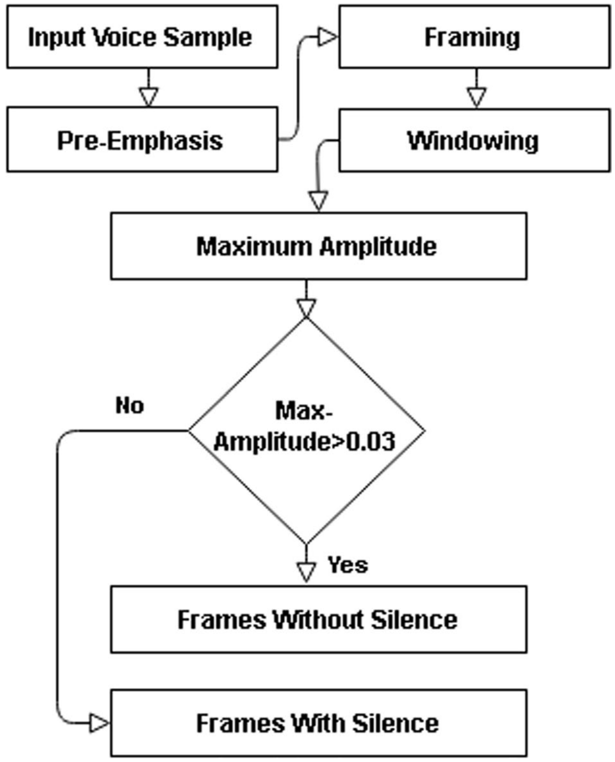

Fig. 4 Voice samples' pre-processing

As per Fig. 4, for pre-processing of voice samples, initially, each voice sample was pre-emphasized by passing through a first order high pass filter. Each pre-emphasized voice sample is then divided into 20 ms to 30 ms duration overlapping frames for analysis. Each frame is then analyzed using Hamming window. Based on the amplitude of each windowed frame, frames with silence and without silence were separated and frames with silence were removed.

Features extraction

After pre-processing, MFCC features were extracted from all voice samples. Feature extraction is the next important step after pre-processing for developing voice based recognition systems. The output from the feature extraction process is the main input for speaker model development and matching processes.

The MFCC technique is a most popular, has a huge achievement and extensively used in the speaker and speech recognition systems [35, 36]. It is based on a logarithmic scale and is able to estimates human auditory response in a better way than the other cepstral feature extraction techniques [37, 38]. MFCC features are derived from short-term Fast Fourier Transform (FFT) power spectrum of the pre-emphasized input speech samples. To obtain these features, initially, the pre-emphasized input voice sample is divided into fixed length segments known as frames. The purpose of framing is to analyze the input voice sample in nearly non-varying/static form (by nature, speech signals are non-stationary). After framing each frame is passed through a window (hamming window in our case) to analyze. Each frame is analysed by applying window to remove discontinues at the beginning and the end of each frame. The resulting speech sample is then transformed into frequency domain from time domain by simply applying FFT. Transformed signal values are then plotted against the Mel scale (Mel-scale has linear frequency spacing lower than 1000 Hz and a logarithmic frequency interval higher than 1000 Hz) [39]. Finally, MFCC coefficients are obtained by using a Discrete Cosine Transform (DCT) of the logarithm of the power on each Mel frequency.

The process of feature extraction is shown in Fig. 5. In this research, 19 MFCC were obtained and used as feature vectors for ML algorithms.

Classification models

After extraction of voice features (MFCCs), some of the popular ML algorithms relevant and suitable to the present research such as SVM, ANN, k-NN, and RF [40–42] were trained using training set of feature vectors and tested on the set of test feature vectors. To build learning models and for train-test splits of feature vectors, ten-fold cross-validation (TFCV) technique was used. In TFCV method, the original feature vector is randomly divided into ten nearly equal sub samples. One of the ten samples is randomly chosen as a test feature vector, whereas, all the remaining sub-samples are used as training feature vectors. Similar, process is repeated until all the ten sub-samples of feature vectors have been tested [43]. In this procedure, since, the classification accuracy is based on ten estimates rather than just a single estimate, therefore, TFCV produces a more precise estimate of classification accuracy than cross validation [44].

SVM is a supervised machine learning model [45] that analyzes data and recognizes patterns, used for regression and classification. It is a well-known discriminative classifier that models the boundary between the speaker and a group of impostors. We implemented a Sequential Minimal Optimization (SMO) algorithm to train SVM classifier [46].

K-NN algorithm is a family of lazy and instance-based learning algorithms. Whenever it is necessary to classify a sample of unknown data from a test data set, the KNN's task is to examine training data set for the most related k samples. Instance based classification algorithm, provide an efficient implementation of the KNNs [47]. We used euclidean distance based approach for KNN implementation [48].

RF belongs to a family of supervised learning model, used for the task of classification and regression. RF works by constructing a huge amount of decision trees during

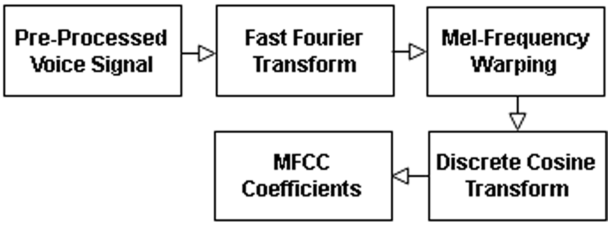

Fig. 5 Features extraction process

the training phase and producing a class that is the mode of the classes. RF is a very precise and robust classifier and is not subject to overfitting problem [49]. We implemented RF with bootstrapping of samples technique [40] in this research.

ANN is used in a wide range of applications. It consists of a collection of various neurons often called nodes of network connection and is a simplified version of the human brain [50]. It consists of an input, hidden, and output layers. Its objective is to get inputs and transform them into meaningful outputs. Here in this research, Multilayer Perceptron (MLP) algorithm of ANN with back propagation feed-forward algorithm is implimentd to classify instances and log-sigmoid function as a neuron activation function [51].

Performance evaluation measurements

The behavior of each classification model is assessed on the basis of certain parameters for measuring its efficiency. The performance of model is influenced by the training data size, the quality of voice records, and most significantly the type of ML model used. We have used the following measurement matrices to assess the efficiency of the ML models [52, 53]:

Accuracy: It shows how frequently the classifier predicts the correct values and can be calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision: The segment of the relevant examples among the retrieved examples. The precision can be calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall: The segment of relevant examples, which are retrieved from the total relevant examples and can be defined mathematically as bellow:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F-measure: It is the harmonic mean of the precision and recall. It can be expressed mathematically as bellow:

$$F\text{-}measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

where TP is the number of samples predicted as positive that are actually positive; FP is the number of samples predicted as positive that are actually negative; TN is the number of samples predicted as negative that are actually negative; FN is the number of samples predicted as negative that are actuall positive

Root Mean Squared Error: It is used to measure the mean magnitude of the errors in an experiment using a quadratic scoring rule and it can be calculated as bellow::

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{5}$$

where $\hat{y}_i$ is the estimate of y_i.

Results and discussion

To test the effectiveness of ML models on the collected voice datasets, we performed two experiments. In the first experiment, the performance of ANN, SVM, KNN and RF models was evaluated on the feature vectors obtained from the English voice dataset, whereas, in the second experiment, the performance of the same models was evaluated on the feature vectors obtained from the Urdu voice dataset. In each experiment, for the training and testing splits of features data, TFCV method was used. WEKA was used as implementation tool to implement ML classifiers. Table 3 outlines the design parameters of the employed ML algorithms.

Tables 4 and 5 outlines the results obtained during first and second experiments respectively.

Table 3 Design parameters of ML classifiers

ML Models	Parameters
SVM	Kernal= polynomial, Seeds=1
RF	Trees= 100, Bag size = 100
KNN	K= 3, Distance = Euclidian
ANN	Learning Rate = 0.3, Momentum = 0.2, Epochs = 500, Input nodes = 19, Hidden layers = 01, Hidden units = 10

Table 4 Performance of ML models on English dataset

Performance measures	Classifier models			
	ANN	SVM	KNN	RF
Accuracy	88.53%	85.54%	86.11%	85.28%
RMSE	0.032	0.074	0.0384	0.0499
Precision	0.889	0.845	0.869	0.841
Recall	0.885	0.855	0.861	0.853
F-Measure	0.886	0.835	0.862	0.861

Table 5 Performance of ML models on Urdu dataset

Performance measures	Classifier models			
	ANN	SVM	KNN	RF
Accuracy	86.58%	81.75%	83.03%	81.12%
RMSE	0.0346	0.074	0.0424	0.0517
Precision	0.869	0.829	0.840	0.800
Recall	0.866	0.818	0.830	0.811
F-Measure	0.865	0.817	0.821	0.813

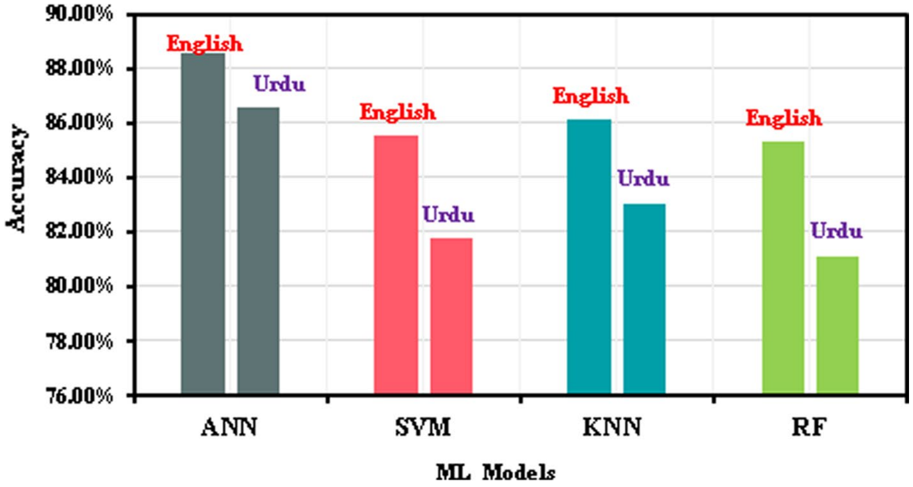

Fig. 6 The accuracy of the classification models with TFCV for English and Urdu Dataset

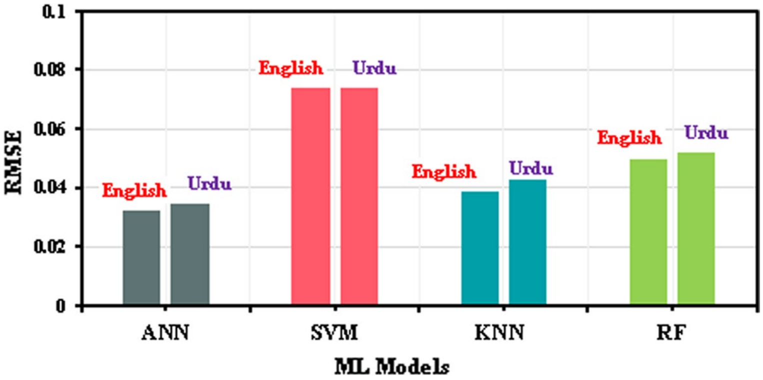

Fig. 7 The RMSE of the classification models with TFCV for English and Urdu Dataset

Table 4 shows that ANN classifier outperformed SVM, RF and KNN by showing 2.99%, 3.25% and 2.42% improvement in classification accuracy respectively in classifying speakers of English dataset. Root Mean Square Error (RMSE) is the standard deviation of the recognition errors. SVM has the highest RMSE at 0.074 followed by RF at 0.049. ANN and KNN have the lowest RMSE of 0.032 and 0.038.

Table 5 shows that again ANN outperformed RF, KNN and SVM by 5.4%, 3.5% and 4.8% respectively in classifying speakers on Urdu dataset. Furthermore, SVM has the uppermost RMSE of 0.074 followed by RF of 0.051, whereas ANN and KNN have the lower RMSE of 0.034 and 0.042 respectively. This also indicates that SVM, RF and KNN misclassified most of the Urdu voice dataset compared to ANN.

Figures 6 and 7 provide the comparison of the performance of ML models on English and Urdu datasets.

Figure 6 provides the comparision of recognition accuracies of ML models (i.e. ANN, SVM, KNN and RF) on English and Urdu datasets. All the models achieved higher accuracy on English dataset as compared to Urdu dataset. However, ANN achieved high accuracy for both the datasets as compared to other classifiers.

Figure 7 provides the comparision of RMSE of the models on English and Urdu datasets. All the models provided Lower RMSE on English dataset as compared to Urdu dataset. However, ANN model showed lower RMSE for both the datasets.

It can be seen through the results provided that ANN outperformed all the other tested ML models on the developed datasets. As already mentioned, Multilayer Perceptron (MLP) algorithm of ANN was used to classify different speakers based on different accents. MLP better approximates the classes overlapping in nature [54]. The same is case here in this research, where the different regional accents of GB, Pakistan are overlapping in nature, that is one of the reason why ANN outperformed the other used ML algorithms.

Comparison with the other systems in literature

Finally, the achieved experimental results were compared with some of the state of the art recently proposed classifier models. During the literature review, it was found that:

- Rizwan et al. [25] applied SVM on TIMIT dataset and achieved 77.8% recognition accuracy.
- Danao et al. [27] applied MLP on their own designed Philippine dataset and achieved 56.19% recognition accuracy.
- Shah et al. [10] applied MLP on their own developed Poshtu speakers' dataset and achieved 87.5% recognition accuracy.
- Liu et al. [55] developed an MFCC-based text-independent speaker identification system for access control. In their system, along with MFCC features, they used Gaussian Mixture Models (GMM) as a classifier. Their system achieved overall 86.87% identification accuracy.

Comparative studies indicates that the proposed ANN model using collected English dataset outperformed all the above mentioned models reported in literature.

Conclusion

In this paper, the authors have designed voice datasets in Urdu and English languages with five different regional accents spoken in GB, located at the north of Pakistan. These datasets are specifically designed to support and extend research in the domain of speaker recognition systems. The designed voice datasets represents 7200 voice samples of 180 speakers and the content is in the form of single words, 10–16 digit strings and speaker's personal information. Designed datasets were pre-processed to extract voice features and used in training four ML algorithms including SVM, ANN, RF and KNN. The recognition accuracy indicates that ANN classifier model outperforms SVM, RF and KNN by achieving 88.53% and 86.58% on English and Urdu dataset respectively. Similarly, RMSE of ANN is best among other ML algorithms as the ANN model is simple and does not overfit. Moreover, it was found that ANN outperformed some of the recently proposed classifiers in literature.

Abbreviations

MFCC: Mel Frequency Cepstral Coefficient; ANN: Artificial Neural Network; RF: Random Forest; KNN: K-nearest neighbors; SVM: Support Vector Machine; VBS: Voice Based Systems; MLP: Multi layer perceptron.

Acknowledgements

Authors would like to acknowledge and thanks to the participants in the collection of voice samples.

Authors' contributions

MI: I am submitting this paper as part of my PhD work. SM: Supervised my work. LMD is acted as my co-supervisor. SMS: helped me in data collection and revising my whole paper. DH: supported me in compiling the paper. SR: Support me in data collection. SR and IA also helped me in revising my paper and also in financial contribution. All authors read and approved the final manuscript.

Funding

This research is not funded by any private/public organization.

Availability of data and materials

The dataset used in this study can be obtained from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

All authors are agreed to publish their work in journal of Big Data

Competing interests

The authors declare that they have no competing interests.

Author details

[1] AHS Bukhari Institute of Information and Communication Technology, Faculty of Engineering and Technology, University of Sindh, Jamshoro, Pakistan. [2] Department of Computer Science, Karakoram International University, Gilgit, Pakistan. [3] Department of Computer Science, Barrett Hodgson University, Karachi, Pakistan.

Received: 20 October 2020 Accepted: 22 February 2021
Published online: 02 March 2021

References

1. de Luis-Garcia R, Alberola-López C, Aghzout O, Ruiz-Alzola J. Biometric identification systems. Signal Process. 2003;83(12):2539–57.
2. Jain AK, Ross A, Prabhakar S. An introduction to biometric recognition. IEEE Trans Circ Syst Video Technol. 2004;14(1):4–20.
3. Memon S, Shah SGS. Securing sensitive edatabases using multibiometric technology. Bahria Univ J Inf Commun Technol BUJICT 2010;3(1):46–9.
4. Prabhakar S, Pankanti S, Jain AK. Biometric recognition: security and privacy concerns. IEEE Secur Privacy. 2003;1(2):33–42.
5. Ghayoumi M. A review of multimodal biometric systems: Fusion methods and their applications. In: 2015 IEEE/ACIS 14th international conference on computer and information science (ICIS). New York: IEEE; 2015. p. 131–6.
6. Reynolds DA. An overview of automatic speaker recognition technology. In: 2002 IEEE international conference on acoustics, speech, and signal processing, vol. 4. New York: IEEE; 2002. p. 4072.
7. Chauhan T, Soni H, Zafar S. A review of automatic speaker recognition system. Int J Soft Comput Eng. 2013;3(4):132–5.
8. Yadav S, Rai A. Learning discriminative features for speaker identification and verification. In: Interspeech; 2018. p. 2237–41.
9. Shah SM, Memon M, Salam MHU. Speaker recognition for pashto speakers based on isolated digits recognition using accent and dialect approach. J Eng Sci Technol. 2020;15(4):2190–207.
10. Shah SM, Memon SA, Khoumbati KUR, Moinuddin M. A pashtu speakers database using accent and dialect approach. Int J Appl Pattern Recogn. 2017;4(4):358–80.
11. Xu M, Zhang L, Wang L. Database collection for study on speech variation robust speaker recognition. In: O-COCOSDA: proceedings; 2008.
12. Saquib Z, Salam N, Nair RP, Pandey N, Joshi A. A survey on automatic speaker recognition systems. In: Signal processing and multimedia. Berlin: Springer; 2010. p. 134–45.
13. Kozlov A, Kudashev O, Matveev Y, Pekhovsky T, Simonchik K, Shulipa A. Svid speaker recognition system for nist sre 2012. In: International conference on speech and computer. Berlin: Springer; 2013. p. 278–85.
14. Lee KA, Larcher A, Wang G, Kenny P, Brümmer N, Leeuwen Dv, Aronowitz H, Kockmann M, Vaquero C, Ma B, et al. The reddots data collection for speaker recognition. In: sixteenth annual conference of the international speech communication association; 2015.

15. McLaren M, Ferrer L, Castan D, Lawson A. The 2016 speakers in the wild speaker recognition evaluation. In: Interspeech; 2016. p. 823–7.
16. McLaren M, Ferrer L, Castan D, Lawson A. The speakers in the wild (sitw) speaker recognition database. In: Interspeech; 2016. p. 818–22.
17. Chung JS, Nagrani A, Zisserman A. Voxceleb2: Deep speaker recognition; 2018. arXiv preprint arXiv:1806.05622.
18. Fisher WM. Ther darpa speech recognition research database: specifications and status. In: Proceedings of the DARPA workshop on speech recognition, Feb. 1986; 1986. p. 93–9.
19. Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. STIN 1993;93:27403.
20. Hennebert J, Melin H, Petrovska D, Genoud D. Polycost: a telephone-speech database for speaker recognition. Speech Commun. 2000;31(2–3):265–70.
21. Vloed Dv, Bouten J, van Leeuwen DA. Nfi-frits: A forensic speaker recognition database and some first experiments; 2014.
22. Millar, J.B., Vonwiller, J.P., Harrington, J.M., Dermody, P.J.: The australian national database of spoken language. In: Proceedings of ICASSP'94. IEEE international conference on acoustics, speech and signal processing. New York: IEEE; 1994. p. 97.
23. Greenberg CS. The nist year 2012 speaker recognition evaluation plan. Technical report: NIST; 2012.
24. Farrús M. Voice disguise in automatic speaker recognition. ACM Comput Surveys CSUR. 2018;51(4):1–22.
25. Rizwan M, Anderson DV. A weighted accent classification using multiple words. Neurocomputing. 2018;277:120–8.
26. Selouani SA, Boudraa M. Algerian arabic speech database (algasd): corpus design and automatic speech recognition application. Arab J Sci Eng. 2010;35(2):157–66.
27. Danao G, Torres J, Tubio JV, Vea L. Tagalog regional accent classification in the philippines. In: 2017IEEE 9th international conference on humanoid, nanotechnology, information technology, communication and control, environment and management (HNICEM). New York: IEEE; 2017. p. 1–6.
28. Thullier F, Bouchard B, Menelas B-AJ. A text-independent speaker authentication system for mobile devices. Cryptography. 2017;1(3):16.
29. UNPO: impact of climate change on biodiversity; 2017. https://unpo.org/members/8727. Accessed 20 Dec 2020.
30. Shams SA, Anwar ZUH. Linguistic identity construction of shina speakers: an ethnographic study. Glob Soc Sci Rev. 2019;4(3):278–83.
31. Patil HA, Basu T. Development of speech corpora for speaker recognition research and evaluation in indian languages. Int J Speech Technol. 2008;11(1):17–32.
32. Berdibaeva GK, Bodin ON, Kozlov VV, Nefed'ev DI, Ozhikenov KA, Pizhonkov YA. Pre-processing voice signals for voice recognition systems. In: 2017 18th international conference of young specialists on micro/nanotechnologies and electron devices (EDM). New York: IEEE; 2017. p. 242–5.
33. Ibrahim YA, Odiketa JC, Ibiyemi TS. Preprocessing technique in automatic speech recognition for human computer interaction: an overview. Ann Comput Sci Ser. 2017;15(1):186–91.
34. Kim C, Stern RM. Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction. In: tenth annual conference of the international speech communication association; 2009.
35. Zhen B, Wu X, Liu Z, Chi H. On the importance of components of the mfcc in speech and speaker recognition. In: Sixth international conference on spoken language processing; 2000.
36. Tiwari V. Mfcc and its applications in speaker recognition. Int J Emerg Technol. 2010;1(1):19–22.
37. Lozano-Diez A, Silnova A, Matejka P, Glembek O, Plchot O, Pesan J, Burget L, Gonzalez-Rodriguez J. Analysis and optimization of bottleneck features for speaker recognition. Odyssey. 2016;2016:352–7.
38. Tirumala SS, Shahamiri SR, Garhwal AS, Wang R. Speaker identification features extraction methods: a systematic review. Expert Syst Appl. 2017;90:250–71.
39. Bezoui M, Elmoutaouakkil A, Beni-hssane A. Feature extraction of some quranic recitation using mel-frequency cepstral coeficients (mfcc). In: 2016 5th International Conference on Multimedia Computing and Systems (ICMCS). New York: IEEE; 2016. p. 127–31.
40. Shah SM, Ahsan SN. Arabic speaker identification system using combination of dwt and lpc features. In: 2014 international conference on open source systems & technologies. New York: IEEE; 2014. p. 176–81.
41. Hansen JH, Hasan T. Speaker recognition by machines and humans: a tutorial review. IEEE Signal Process Mag. 2015;32(6):74–99.
42. Khan S, Ali H, Ullah Z, Minallah N, Maqsood S, Hafeez A. Knn and ann-based recognition of handwritten pashto letters using zoning features; 2019. arXiv preprint arXiv:1904.03391.
43. Jiang P, Chen J. Displacement prediction of landslide based on generalized regression neural networks with k-fold cross-validation. Neurocomputing. 2016;198:40–7.
44. Verbyla DL, Litvaitis JA. Resampling methods for evaluating classification accuracy of wildlife habitat models. Environ Manag. 1989;13(6):783–7.
45. Singh S. Support vector machine based approaches for real time automatic speaker recognition system. Int J Appl Eng Res. 2018;13(10):8561–7.
46. Platt J. Sequential minimal optimization: a fast algorithm for training support vector machines. Microsoft Research Technical Report MSR-TR-98-14. 1998.
47. Aha DW, Kibler D, Albert MK. Instance-based learning algorithms. Mach Learn. 1991;6(1):37–66.
48. Guo G, Wang H, Bell D, Bi Y, Greer K. Knn model-based approach in classification. In: OTM confederated international conferences "on the move to meaningful internet systems". Berlin: Springer; 2003. p. 986–96.
49. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
50. Nagy H, Watanabe K, Hirano M. Prediction of sediment load concentration in rivers using artificial neural network model. J Hydraulic Eng. 2002;128(6):588–95.
51. Yee CS, Ahmad AM. Malay language text-independent speaker verification using nn-mlp classifier with mfcc. In: 2008 international conference on electronic design. New York: IEEE; 2008. p. 1–5.

52. Mokgonyane TB, Sefara TJ, Modipa TI, Mogale MM, Manamela MJ, Manamela PJ. Automatic speaker recognition system based on machine learning algorithms. In: 2019 Southern African Universities power engineering conference/robotics and mechatronics/pattern recognition association of South Africa (SAUPEC/RobMech/PRASA). New York: IEEE; 2019. p. 141–6.
53. Hussain D, Hussain T, Khan AA, Naqvi SAA, Jamil A. A deep learning approach for hydrological time-series prediction: a case study of gilgit river basin. Earth Sci Inf. 2020;13(3):915–27.
54. Tsapanos N, Tefas A, Nikolaidis N, Pitas I. Neurons with paraboloid decision boundaries for improved neural network classification performance. IEEE Trans Neural Netw Learn Syst. 2018;30(1):284–94.
55. Liu J-C, Leu F-Y, Lin G-L, Susanto H. An mfcc-based text-independent speaker identification system for access control. Concurr Comput Pract Exp. 2018;30(2):4255.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.