Foreign Accents Classification of English and Urdu Languages, Design of Related Voice Data Base and A Proposed MLP based Speaker Verification System

Muhammad Ismail [1†], Shahzad Ahmed Memon [2 †], Lachhman Das Dhomeja [3], Shahid Munir Shah[4]

[†1] Department of Computer Science, Karakoram International University(KIU), Gilgit, Pakistan

[†2] IICT, University of Sindh, Jamshoro, Pakistan

[†3] IICT, University of Sindh, Jamshoro, Pakistan

[†4] Faculty of Information Technology, Department of Computer Science, Barrett Hodgson University, Korangi Creek, Karachi, Pakistan

Abstract

A medium scale Urdu speakers' and English speakers' database with multiple accents and dialects has been developed to use in Urdu Speaker Verification Systems, English Speaker Verification Systems, accents and dialect verification systems. Urdu is the national language of Pakistan and English is the official language. Majority of the people are non-native Urdu speakers and non-native English in all regions of Pakistan in general and Gilgit-Baltistan region in particular. In order to design Urdu and English speaker verification systems for security applications in general and telephone banking in particular, two databases has been designed one for foreign accent of Urdu and another for foreign accent of English language. For the design of databases, voice data is collected from 180 speakers from GB region of Pakistan who could speak Urdu as well as English. The speakers include both genders (males and females) with different age groups ranging from 18 to 69 years. Finally, using a subset of the data, Multilayer Perceptron based speaker verification system has been designed. The designed system achieved overall accuracy rate of 83.4091% for English dataset and 80.0454% for Urdu dataset. It shows slight differences (4.0% with English and 7.4% with Urdu) in recognition accuracy if compared with the recently proposed multilayer perceptron (MLP) based SIS achieved 87.5% recognition accuracy

Keywords:

Speaker recognition systems; Voice recognition system; Urdu speakers; English speakers; Speaker identification systems; Accent identification systems; Dialect identification systems; Urdu speakers' database; English Speakers' database; Pattern recognition; Multi-Layer Perceptron; MFCC; Noise robustness; Intra speaker variability; GB languages;

1. Introduction

Biometric is a general term used to describe a characteristic or a process. The biometric as a characteristic is a measureable physiological or behavioral characteristic that can be used for automatic personal recognition. The biometric as a process is an automated method of recognizing an individual based on properties like behavioral or physiological [1]. For identification and verification of human beings, their characteristics like voice, face, fingerprint and gait have been used for long ago. These characteristics of human beings can be used as a biometric trait for recognition as long as it satisfied the desirable properties of biometrics such as universality, distinctiveness, permanence and collectability [2]. The fingerprint, faced, iris, vein, ear, DNA etc. are the physiological biometric traits whereas the behavioral biometric traits are voice, key strokes dynamics, signature and gait. These individual traits are unique, can't be lost, stolen and forgotten. For attackers it is quite difficult to replicate and for users it is quite difficult to deny. Hence It 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.

Each biometric has its own advantages and disadvantages. 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 fulfill all the requirements of various applications. For example, in some situations the finger print biometric trait is more desirable than voice biometric trait. In another situation, the voice biometric is preferable than figure 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 modem on dial-up telephone line, in call centers, forensic application where speaker recognition is required [4][5]. The human voice carries different characteristics such as the meaning/words a speaker wants to pass to a listener, speaker emotions, spoken language information, gender and identity of speakers, speaker's health and speaker's age related information etc. The objective of speaker recognition is to extract information about speaker's identity and based on that information it recognizes the

Manuscript received October 5, 2024
Manuscript revised October 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.10.5

speakers [6]. 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. The verification system must perform a 1:1 comparisons hence the cost of computation is independent of the records in the voice database. On the other hand, 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 [5, 6, 7]. The design of voice database is an essential requirement for speaker recognition systems. In this perspective two voice databases one for Urdu language and one for English language with five different accents spoken in Gilgit-Baltistan located at the North of Pakistan have been designed. Our work is basically, a voice database design in foreign accents of Urdu and English spoken in GB region. The database is specifically designed for telephone banking services and mobile network services. To the best of our knowledge, these are the first databases designed for speaker recognition particularly for telephone banking and mobile network services. Incorporation of different accents in the designed voice databases addresses the performance degradation of Automatic Speaker Recognition System (ASRS) due to different accents. The accents of a language can cause performance degradation in ASRS like the other performance degrading factors such as noise, speaker's age, channel mismatch, speaker health and emotions etc.

Typically, ASRS do not perform well if the accent of a speaker, who is going to be recognized, is different from the accent of the speaker by whom the system was trained. Incorporation of accents can minimize the variability caused by different accents of a language and which in turns enhance the performance of recognition system [8]. The other aforementioned performance degrading factors (other than different accents and dialects) have also been addressed during database design. Further details of the development of voice database are presented in Section 4. The rest of the paper has been managed in the following way. The review of the existing speech/speaker recognition databases are presented in Section 2. Introduction about study area for the present research is given in the section 3. The design of Multilayer perceptron (MLP) based speaker verification is described in section 5. Conclusion is provided in section 6 and references are provided at the end of the paper

2. Related Works.

In the present research, voice database designed for speaker recognition systems in general and for telephone banking in particular. The designed database contains the voices of speakers from Gilgit-Baltistan region who could speak Urdu and English. In this way, the database represents foreign accent of Urdu and English. The research in the field of speaker recognition has been started in early 1960. Since then, numerous speaker recognition and speech recognition databases have been designed. The basic design goals of voice databases are to support research in the field of speaker recognition and related areas.

Experiments in [9] are conducted based on a self-recording voice library of 50 people. The voices of 50 speakers were recorded at 16 kHz sampling frequency. The designed database was for Chines language. Each speaker reads 10 times, 20 short Chinese phrases. The duration of the utterances was 1-3 seconds long. The duration of utterances per speaker was 1-2 minute long for training data for specific Gaussian Mixture Model (GMM). The purpose of the designed voice library was to support research in text-related short utterance speaker task.

Experiments in [10] were carried out using a database of 30 different speakers. The data collected on line through mobiles with 8 kHz sampling frequency. The designed database represents 32 voice samples of each speaker. Each speaker uttered 32 times district and mandi name of Jharkhand state.

Experiments in [11] were carried out using a small scale database of 11 speakers including 7 male and 4 female, with their age ranging from 19-36 years. All of the speakers were native French. Some speakers from all had unique Canadian French accents and Hexagonal French accents. The experiments were conducted in a silent room (university meeting room) as well as in noisy environment (i.e. University cafeteria).

A small scale database of Pashtu speakers was developed to support research in Pashtu speaker identification system, accents and dialect identification systems. The designed voice database represents 32 native Pashtu speakers (male), with different age groups (15 to 55 year), from different areas of Pakistan and Afghanistan. The data was recorded in 8 different sessions using smart phones and a sonny recorder. The data was collected with their age ranging from 15 to 55 years. The author has designed a Multilayer Perceptron (MLP)-based Speaker Identification System to test the collected database. The system achieved 87.5% identification accuracy. The author believes that the designed database can also be used in designing systems for recognizing region of Pashto speakers [8].

TCD-TIMIT is an Audio-Visual Corpus designed for support research in continuous audio visual speech recognition. The developed corpus represents 62 speakers. Three speakers among 62 were professional trained lip speakers. The audio/video data was recorded from speakers reading 6913 phonetically rich sentences. The

video clips were acquired from two angles by using two cameras. One camera recorded the speaker from the front side where as the other camera recoded the speaker at 30 degree angle right side of the speakers. Some experiments have been conducted on the lip-speakers and on non-lip speakers (regular speakers) and there results were analyzed. Results on the lip-speakers were higher as compared to the results on the non-lip-speakers [12].

RedDots is a project with the aims 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 consist of short duration test utterances with variable phonetic content. The main focus for the RedDots database was to include 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 [13].

A Pashtu spoken digits database was developed to support research in speech recognition. The developed database represents speech data of 60 speakers including both males and females with different age groups (18 to 60 years). The content of the database consist of digits (0 to 100) spoken by Pashto speakers. The speech data was recorded using a recorder (Sony PCM-M 10) in a noise free environment [14].

To carry out some experiments in [15], A corpus was developed. It represents 15 speakers including male and female with different age group. The database contains a total of 110 speech samples. The voice samples were recorded from each speaker directly using Android mobile device.

An Algerian Speech corpus was designed to support research in speech recognition. The corpus represents 300 Algerian native speakers who could speak Modern Standard Arabic (MSA). 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 post-graduate level). Finally, using a subset of the collected dataset, author has designed a txt-independent ASR system that achieved 91.65% recognition rate [16].

An Urdu Speech Corpus was designed for speaker independent spontaneous Urdu speech recognition system. The designed database represents 45 hours of speech data collected from 82 speakers including 42 males and 40 females with different age groups (20-55years). The content of the database was in the form of spontaneous and read speech. The recording sessions were conducted in office and home environment. The data was collected using microphone connected to a laptop and telephone line [17].

SAAVB database was developed for Arabic language that represents 1103 native Arabic speakers who could speak MSA with Saudi accent. The specified speakers belong to Saudi Arabia. The content of the database was verified internally as well as externally by IBM Cairo. The database can be practiced for automatic speech recognition and verification systems [18].

An Indian Language speech databases in Tamil, Telugu and Marathi was designed to support research in large vocabulary speech recognition systems. The designed database represents 560 speakers (both males and females) of the said languages with different age groups. The data was recorded by using cell phones and landline phones [19].

POLYCOST is a telephone-speech database supports research particularly in speaker recognition application over telephone network. The designed database represents 134 speakers including 74 males and 60 females' foreigner English speakers from 13 European countries which are member of the COST 250. Around ten speakers from each specified countries were selected for database. In the database majority of the speakers were non-native English. Most of the data was recorded in the form of digits and some data with free speech. The speech from speakers was acquired telephonically in more than 8 sessions per speaker up to two month period of time. Moreover, 4 baseline speaker recognition experiments were defined to enable cross-site comparisons of the algorithms [20].

AHUMADA is a Spanish database for speaker characterization and identification. The speech database was designed concerning various sources of intra speaker variability and letting researchers study the underling effects of these variability's in speaker recognition system. Examples of some variations included in this database were read text at different speech rates, different microphones and telephone sets, dialectical variations and read speech versus spontaneous speech. To obtain the said intra-speaker variation factors, the speech data was collected in form of 24 isolated digits, 10 digit strings, phonologically and syllabically balanced phrase and more than one minute spontaneous speech was recorded from 104 speakers. The data was recorded in different recording sessions [21].

The CSLU Speaker Recognition CORPUS was motivated by a need for speech data from 500 speakers over different sessions. The speech data was collected from each speaker in 12 different recording sessions over two year period. To normalize the seasonal effects (hay fever in summer, cold in winter) the speech data was collected from speakers in different sessions. The main designed goal of CSLU was to support research in text independent and text dependent speaker recognition and verification systems. The data was collected from the specified speakers in the form of single words, digit strings,

speaker's personal information, phonetically rich phrases and free speech [22].

GANDALF is a Swedish telephone speaker verification database designed to support research in the field of automatic speaker verification. For its development, voice data from 86 speakers was recorded. The data was collected from speakers speaking during telephone calls. There were 24 telephone calls per speaker during a period of up to 12 months [23].

SIVA is an Italian language based speech database consists of four categories such as female users, male users, female imposters and male imposters. Almost 500 speakers were trained via email to record their voices through Public Switched Telephone Network (PSTN). They were supposed to read the information provided via email before making call. The speech data for SIVA database was collected over PSTN. About two thousands calls were recorded from specified speakers in different sessions. Moreover a text independent speaker verification system was presented and discussed using a subset of the developed database [24].

YOHO CD-ROM is a voice database, designed to compare the performance between various testing voice verification systems. The designed database represents voices from 138 speakers including 32 females and 106 males. There were four sessions per speaker and twenty four notes for each session. Similarly there were also ten sessions for verification each speaker and four notes in each session [25].

The Otago is a Speech Database represents speakers who could speak New Zealand English accent. The database was in the form of words and digits. The data was collected from 43 speakers including males and females. For the digit collection speech of 11 males and 10 females was collected where as for the word acquisition speech of 10 females and 12 males was collected. Additionally, the authors developed database management system to house the designed database. It allows researchers to query the database and they can extract the required content of the database. The developed DBMS is a huge progress over file based methods can be used to household corpora [26].

The OGI 22 language telephone speech corpus is an extension of OGI multi-language telephone speech database designed for support research in language identification as well as spoken language systems. The developed database represents at least 200 speakers per language. The speech data was collected through telephone calls (approx. 2 to 3 minutes) from speakers of 22 languages. The content of the database was in the form of specific information such as what is your native language, continuous speech is in the form of selected topic such as describe your most recent meal and extemporaneous speech such as speech for one minute on any topic [27].

King database was designed in the year 1987 by Alan Higgins. The designed database represents voice of 51 male speakers. The speakers were subdivided into two groups. One group contains 25 and another contains 26 speakers. The data was recorded from each group of speakers from different locations with the time duration 30 to 60 seconds in 10 different sessions. The data was recorded over telephone lines using telephone handset and high quality microphones to have channel variability. The primarily designed goal of KING database was to support research in closed set text independent speaker identification or verification over telephone lines [28].

SWITCHBOARD is a large multi speaker database. Voice data from 500 speakers both males and females from around U.S was recorded automatically over telephone lines. This database includes 2500 American English conversations with the time duration of three to ten minutes. Its designed purpose was to support research in the field of speaker authentication and large vocabulary speaker recognition [29].

3. Gilgit-Baltistan and various mother tongues

The selected region for speaker database is Gilgit-Baltistan (GB) located at the North of the Pakistan. It borders with Azad Kashmir, Jammu Kashmir, Khyber Pakhtunkhwa, Afghanistan and China. Gilgit-Baltistan is an area of highly mountains and has an area of over 72,971. The capital city of GB is Gilgit and the population of GB was 1,800,000 in 2015. There are ten districts in GB such as 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 almost five different native languages spoken in these districts such as Shina, Balti, Burushishki, Khuwar and Wakhi. Majority people of this region speak Shina language. Majority of the people of this region can also speak Urdu and English which are national and official languages of Pakistan respectively. Because of having different mother tongues, cultures and backgrounds these people speaks Urdu and English with different accentual and dialectical variations in different districts. Further details about the language spoken in the corresponding districts are mention in the following sections [30, 31].

4. Database Design

In order to design Urdu and English Databases with various accentual and dialectical variations according to GB region, the voice data was acquired from the speakers of ten districts of the GB. It is an important part of the China-Pakistan Economic Corridor (CPEC). There are

almost five different native languages spoken in these districts such as Shina, Balti, Burushishki, Khuwar and Wakhi. From each district several speakers were chosen. Details about selection of speakers are given in the table 1to 6.

Table 1: District Wise Speaker Selection

Districts	Accent	Language	Speakers		
			Male	Female	Total
Gilgit, Hunza, Nagar, Ghizer, Astore, Diamer, Skardu, Ghanche, Shigar, Kharmang	Shina	English and Urdu	33	33	66
Gilgit, Hunza, Nagar, Ghizer, Astore, Diamer, Skardu, Ghanche, Shigar, Kharmang	Balti	English and Urdu	21	21	42
Gilgit, Hunza, Nagar, Ghizer, Astore, Diamer, Skardu, Ghanche, Shigar, Kharmang	Burushiski	English and Urdu	13	13	36
Gilgit, Hunza, Nagar, Ghizer, Astore, Diamer, Skardu, Ghanche, Shigar, Kharmang	wakhi	English and Urdu	9	9	18
Gilgit, Hunza, Nagar, Ghizer, Astore, Diamer, Skardu, Ghanche, Shigar, Kharmang	Khuwar	English and Urdu	9	9	18
Total Number of Speakers : 180 and the total number of Samples: 7200(3600+3600)					

Table 2: Selection of District-wise Shina Speakers

Accent	District	Speakers			Description
		Male	Female	Total	
Shina	Gilgit	9	9	18	Approx. 90% population is Shina speaker (Very very large pop)
	Hunza	1	1	2	Approx. 15% population is Shina speakers(Small pop)
	Nagar	2	2	4	Approx. 35% population is Shina speakers(Very small pop)
	Ghizer	3	3	6	Approx. 35% population is Shina speakers(Large pop)
	Astore	6	6	12	Approx. 99% population is Shina speaker (Medium pop)
	Diamer	9	9	18	Approx. 99% population is Shina speaker (Very large pop)
	Skardu	2	2	4	Approx. 14% population is Shina speakers(Large pop)
	Kharmang	1	1	2	Approx. 14% population is Shina speakers(Very small pop)
Total Shina Speakers	Sixty Six speakers				

Table 3: Selection of District-wise Balti Speakers

Accent	District	Speakers			Description
		Male	Female	Total	
Balti	Skardu	8	8	16	Approx. 85% population is Balti speakers(Large pop)
	Ghanchi	6	6	12	Approx. 99% population is Balti speakers(Medium pop)
	Shigar	4	4	8	Approx. 99% population is Balti speakers(Very small pop)
	Kharmang	3	3	6	Approx. 85% population is Balti speakers(Very small pop)
Total Balti Speakers	Forty Two				

Table 4: Selection of District-wise Burushiski Speakers

Accent	District	Speakers			Description
		Male	Female	Total	
Burushiski	Gilgit	1	1	2	Approx. 9% population is Burushishiki speakers(Very very large pop)
	Hunza	5	5	10	Approx. 45% population is Burushishki speakers(Small pop)
	Nagar	7	7	14	Approx. 64% population is Burushishki speakers(Very small pop)
	Ghizer	5	5	10	Approx. 35% population is Burushishki speakers(Large pop)
Total Burushishki Speakers	Thirty six Speakers				

Table 5: Selection of District-wise Khuwar Speakers

Accent	District	Speakers			Description
		Male	Female	Total	
Khuwar	Hunza	3	3	6	Approx. 4% population is Khuwar speakers(Small pop)
	Ghizer	6	6	12	Approx. 25% population is Khuwar speakers (Large pop)
Total Khuwar Speakers	Eighteen Speakers				

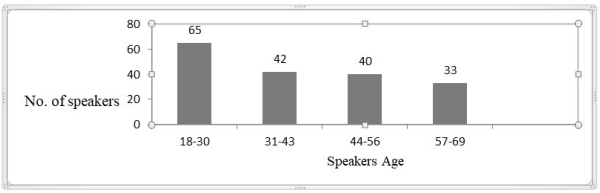
Table 6: Selection of District-wise Wakhi Speakers

Accent	District	Speakers			Description
		Male	Female	Total	
Wakhi	Hunza	6	6	12	Approx. 35% population is Wakhi speakers(Small pop)
	Ghizer	3	3	6	Approx. 4% population is Wakhi speakers (Large pop)
Total Khuwar Speakers	Eighteen Speakers				

4.1 Age distribution of the speakers

For the designed database the voice data was collected from males and females 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. Further to cover maximum telephone banking customers and mobile users. The Age-wise distribution of the speakers is shown in figure 1.

Figure 1: Age-wise speakers' distribution



4.2 Script design

The voice data was collected from each selected speaker based on two specifically designed scripts. One script is for Urdu and another is for English language. The scripts contain all possible conversational talk between a phone banking officer/mobile call center agent and their customers. These scripts contains sentences in the form of words, 10-16 digits strings and speaker's personal information mostly related to bank and mobile network services. All together twenty sentences with average time duration ranging between 10msec to 100ms were included in each of the script. The scripts were provided to each speaker to read for recording data. The designed written script for English language is as shown in the table 7 whereas the script for Urdu language is shown in the table 8.

Table 7: Designed written script for English

S.No	Authentication questions may be asked by phone banking officers/mobile call center agent(Not Recorded)	Answered by the customers (Recorded)
1	Asalamualaikum	Asalamualaikum
2	What is your Name?	My Name is : ……
3	Where are you calling from?	I am calling from (City Name): ……
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 CNIC?	My CNIC 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 CNIC?	The expiry date of my CNIC card is: …
19	What is your occupation?	My occupation is: ……
20	What is your DOB?	My DOB is: ……

Table 8: Designed written script for Urdu

S.No	Authentication Questions may be asked by phone banking officers/mobile call center agent(Not Recorded)	Answered by the customers (Recorded)
1	السلام علیکم	السلام علیکم
2	آپ کا نام کیا ہے ؟	میرا نام.....ہے
3	آپ کہاں سے بات کر رہے ہیں؟	میں.....سے بات کر رہا ہوں
4	آپ کے والد کا نام کیا ہے؟	میرے والد کا نام .....ہے
5	آپ کی والدہ کا نام کیا ہے ؟	میری والدہ کا نام .....ہے
6	آپ کا شناختی کارڈ نمبر کیا ہے؟	میرا شناختی کارڈ نمبر ..... ہے
7	آپ کا مستقل پتہ کیا ہے؟	میرا مستقل پتہ.....ہے
8	آپ کا فون نمبر کیا ہے؟	میرا فون نمبر..... ہے
9	کیاآپ کا فون نمبر اس بینک سے تصدیق شدہ ہے ؟	ہاں یا نا .....
10	آپ کا موجودہ پتا کیا ہے؟	میرا موجودہ پتہ.....ہے
11	آپ کا اکونٹ نمبر کیا ہے؟	میرا اکونٹ نمبر .....ہے
12	آپ کا ڈیبٹ کارڈ نمبر کیا ہے؟	میرا ڈیبٹ کارڈنمبر .....ہے
13	آپ کا کریڈٹ کارڈ نمبر کیا ہے؟	میرا کریڈٹ کارڈنمبر .....ہے
14	آپ کے ڈیبٹ کارڈ کے ختم ہونے کی تاریخ کیا ہے؟	میرے ڈیبٹ کے ختم ہونے کی تاریخ .....ہے
15	آپ کے کریڈٹ کارڈ کے ختم ہونے کی تاریخ کیا ہے؟	میرے کریڈٹ کارڈ کے ختم ہونے کی تاریخ .....ہے
16	آپ کے ڈیبٹ کارڈکا حفاظتی کوڈ کیا ہے ؟	میرے ڈیبٹ کارڈکا حفاظتی کوڈ .....ہے
17	آپ کے کریڈٹ کارڈکا حفاظتی کوڈ کیا ہے ؟	میرے کریڈٹ کارڈکا حفاظتی کوڈ .....ہے
18	آپ کے شناختی کارڈ کے ختم ہونے کی تاریخ کیا ہے؟	میرے شناختی کارڈ کے ختم ہونے کی تاریخ.....ہے
19	آپ کا پیشہ کیا ہے ؟	میرا پیشہ .....ہے
20	آپ کی تاریخ پیدائش کیا ہے؟	میری تاریخ پیدائش.....ہے

4.3 Data Recording Environment and device allocation

The data was recorded from speakers in university office, seminar room and rest room using different smart phones 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 data was recorded from a total of 180 speakers. The speakers were sub divided into 6 groups and each group contains 30 speakers. These groups were assigned to a specific device for recording.

4.4 Recording sessions

The data is collected in different sessions. A total of seven recording sessions from July 2018 to February 2019 were conducted to record voice data. The gap between sessions is at least one month. The details of recording sessions and their corresponding speakers are shown in the figure 2.

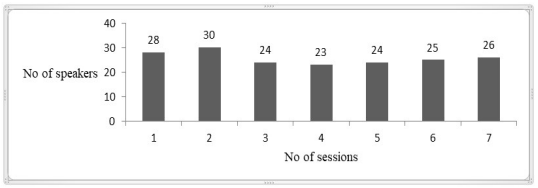


Figure 2: Recording Sessions with corresponding speakers

4.5 Voice samples recording

The designed scripts were provided to each speaker who was selected in a particular session for recording. Before start of each session each speakers were communicated how to record their voices and after words they were supposed to rehearsal for a short period of time. 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 Urdu language as well as for English. A total of 3600(20*180) voice samples have been recorded for English Language. Similarly, a total of 3600(20*180) voice samples have been recorded for Urdu language. So overall a total of 7200(3600+3600) voice sample has 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 shown in the figure 3.

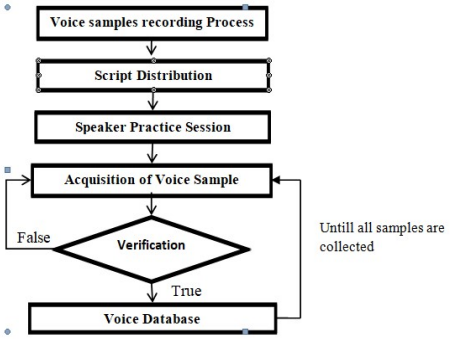


Figure 3: Voice sample recording process

The designed scripts were distributed to the speakers selected for a particular session for recording. During a session the speakers were given instruction how to read the script and making them familiar with the acquisition process. It was a kind of practice session before actual recording. After words the speaker voices samples were recorded sample by sample. Each sample was cross checked with the script to ensure the consistency of 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 all collection of all voice samples.

4.6 Preprocessing and feature extraction

After data collection, the collected voice samples were preprocessed and the features were extracted. Preprocessing of speech plays an important role in the development of an efficient automatic speech recognition system. Preprocessing is the first phase of speech recognition system and it is considered an important step for better results. The process of pre-processing includes noise cancellation, pre-emphasis and silence removal. Preprocessing facilitate the voice based recognition system to be computationally more efficient. Voice is an analog signal. The sampling process actually affects high frequency components of voice signal. To compensate the affect we need to amplify the high frequency components. For this purpose all voice samples have been passed through a high pass filter. It amplifies high frequency components with respect to low frequency components [32]. After pre-emphasis the voice samples were further processed to remove silence. There are various techniques can be used to remove silence from voice sample. Some of these techniques are based on amplitude, Zero crossing rate (ZCR), Short term energy (STE) and so on. In this research work amplitude based techniques has been used to remove silence. This techniques work as it first gets voice sample and it divide the whole audio sample into components of short fixed length called frames. It then calculates maximum amplitude of each frame. It then finds those frames with maximum amplitude is greater than 0.03 and consider those frames as voice portion of the speech and it discard the frames with lesser amplitude then 0.03. This technique assumes that silent part of voice signal have amplitude <0.03 and voice part of voice signal contains amplitude> 0.03 [33]. The process of preprocessing is shown in the figure 4.

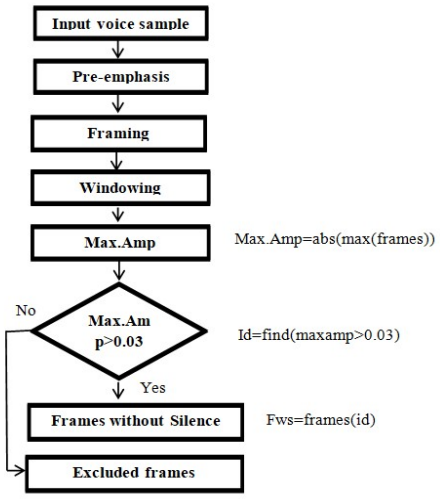


Figure 4: Preprocessing process

After preprocessing, features were extracted from all voice samples. Feature extraction is next important step after preprocessing for developing voice base recognition systems. The output from feature extraction process is the main input for speaker model development and matching processes. There are various techniques such as LPC, PLP, RASTA and MFCC that can be used to extract Cepstral features from voice samples. In this research work Mel Frequency Cepstral Coefficients (MFCCs) has been used because it is the most popular, has a huge achievement and extensively used in speaker recognition system [34]. It is based on logarithmic scale and it estimates human auditory response in a better way than the other Cepstral feature extraction techniques. In order to obtain features the voice sample is taken as input and divide the voice sample into fixed length segments known as frames. The purpose of framing is to make the voice signal static. After framing each frame is multiplied with a hamming window of frame length to minimize the disturbance occurred in the voice signal during the process of framing. The framing process has been done in preprocessing step. Then Fast Fourier Transform (FFT) has been applied on the frames acquired in preprocessing stage to convert the signal into frequency representation and then its values are plotted against Mel scale using the equation: Mel (f) = 2595 log10 (1+f/700). Finally 19 MFCC coefficients were obtained by using discrete cosine transform (DCT) [34][8]. The process of feature extraction is shown in the following figure 5.

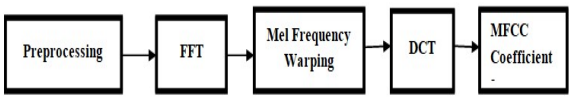


Figure 5: Feature Extraction process

5. Multilayer perceptron (MLP) based speaker verification system

An Artificial neural network (ANN) consists of a collection of various neurons often called nodes of network connected to each other. It is a simplified version of human brain. The typical neural network consists of input layer, hidden layer and output layer. Its objective is to get inputs and transform it into meaningful outputs. Multilayer perceptron (MLP) is one of the popular ANN model used in wide range of applications. It uses back propagation feed forward algorithm to classify instances. In this model neurons/nodes are arranged in different layers. The neurons/nodes in each layer take inputs and weights from the nodes in the preceding layer and transfer their outputs to the nodes of the next layer. An example of feed forward multilayer perceptron is shown in the

following figure 6 [35][36].

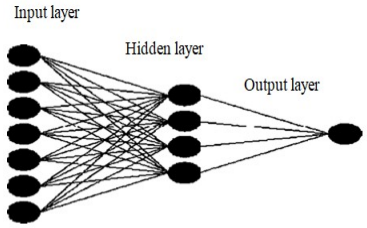


Figure 6: A multilayer perceptron with single hidden layer with 4 neurons [35]

The main idea of this work is to use a multilayer perceptron neural network, which is similar to speakers based on conversational speech used between customers and phone banking officers. A subset of designed database was used as an input to the MLP. The designed database for English language consists of 3600 voice samples from 180 speakers with 20 samples each. Similarly the designed database for Urdu language also consists of 3600 voice samples from 180 speakers with 20 samples each. The subset of the English database used as input to MLP consists of 900 samples from 180 speakers with 5 samples each. Similarly, the subset of Urdu database used as input to MLP consists of 900 samples from 180 speakers with 5 samples each. In the present application 19 dimensional MFCC features obtained during feature extraction process and used as input variables. All the input variables are normalized to minimize the effect of different value ranges and to minimize the effect of outlier and extreme values. The input data for MLP were divided into training set and testing set. The MLP was trained using training set of data and MLP model was tested using testing set of data. In order to build the model different experiments were performed based on different splits of training data such as 50%, 60%, 70%, 80%, 90% and 10 fold cross validation respectively. The experiments were performed with different learning rates, seed values, hidden layers, momentum and epochs. For all experiments the best results were achieved with learning rate 0.3, seed 2, a single hidden layer, momentum 0.2 and epochs 500. The summary of experiments and corresponding results for English dataset as well as for Urdu dataset are shown in the following table 9-14.

Table 9: Recognition accuracy for English dataset

Exp.no	Training data (%)	Testing data (%)	Correctly classified instances	Incorrectly classified instances	No of test instances	Total no of instances	Recognition accuracy (%)
1	50	50	296	144	440	880	67.2727
2	60	40	244	108	352	880	69.3182
3	70	30	187	77	264	880	70.8333
4	80	20	140	36	176	880	79.5455
5	90	10	73	15	88	880	82.9545
6	10 fold CV	10 fold CV	734	146	880	880	83.4091

Table 10: Recognition accuracy for Urdu dataset

Exp.no	Training data (%)	Testing data (%)	Correctly classified instances	Incorrectly classified instances	No of test instances	Total no of instances	Recognition accuracy (%)
1	50	50	285	156	441	882	64.6259
2	60	40	253	100	353	882	71.6714
3	70	30	199	66	265	882	75.0943
4	80	20	137	39	176	882	77.8409
5	90	10	69	19	88	882	78.4091
6	10 fold CV	10 fold CV	706	176	882	882	80.0454

The table 9 and 10 provides recognition accuracy summary of all six experiments for English and Urdu sub datasets respectively. Results listed in table 9 and 10 indicate that the recognition accuracy increases with the increase of training data split. The best result achieved with 10 fold cross validation test scheme. Furthermore the designed classifier provides better results for English dataset if compare with Urdu dataset.

Table 11: weighted precession, recall, F-measure and AUC for English dataset

Exp.no	Training and testing data split (%)	Precession	Recall	F-measure	AUC
1	50/50	0.743	0.673	0.662	0.975
2	60/40	0.739	0.693	0.678	0.983
3	70/30	0.716	0.708	0.688	0.994
4	80/20	0.819	0.795	0.789	0.991
5	90/10	0.856	0.83	0.822	0.989
6	10 fold CV	0.845	0.834	0.829	0.995

Table 12: weighted precession, recall, F-measure and AUC for Urdu dataset

Exp.no	Training and testing data split (%)	Precession	Recall	F-measure	AUC
1	50/50	0.66	0.646	0.612	0.963
2	60/40	0.765	0.717	0.717	0.984
3	70/30	0.803	0.751	0.743	0.989
4	80/20	0.814	0.778	0.765	0.996
5	90/10	0.797	0.784	0.778	0.996
6	10 fold CV	0.82	0.8	0.796	0.994

The table 11 and 12 provides the summary of weighted precession, recall, F-measure and area under curve (AUC) which are most widely used measure to analyzed system accuracy. The AUC results in the table 11 and 12 shows that all experiments gets more than 96% AUC and these are increasing with the increase of training data split and it reaches to 99.5% with 10 fold cross validation test scheme. Similarly all other parameters are increasing with the increase of training data split and reach at its maximum with 10 fold crass validation test scheme.

Table 13: Error analysis of experiments for English dataset

Exp.no	Training and testing data split (%)	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	No of test instances	Total no of instances
1	50/50	0.0061	0.0517	54.8446	69.4323	440	880
2	60/40	0.0057	0.0501	51.3857	67.2963	352	880
3	70/30	0.0053	0.0489	47.9437	65.736	264	880
4	80/20	0.0045	0.0429	40.7583	57.665	176	880
5	90/10	0.004	0.0398	35.8708	53.5535	88	880
6	10 fold CV	0.0038	0.0392	34.78	52.6465	880	880

Table 14: Error analysis of experiments for Urdu dataset

Exp.no	Training and testing data split (%)	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	No of test instances	Total no of instances
1	50/50	0.0062	0.0532	56.5165	71.4521	441	882
2	60/40	0.0056	0.0494	50.6356	66.4139	353	882
3	70/30	0.0051	0.0466	46.2504	62.6768	265	882
4	80/20	0.0047	0.0448	42.6636	60.2203	176	882
5	90/10	0.0043	0.0417	38.4515	38.4515	88	882
6	10 fold CV	0.0042	0.0422	38.1693	56.6841	882	882

The table 13 and 14 provides the summary of error analysis of experiments for English and Urdu sub dataset respectively. It contain mean absolute error, root means squared error, relative absolute error and root relative squared error of all six experiments. Results listed in table 13 and 14 indicate that all mentioned errors are decreasing when the training data split increases. We can get the lowest error rate at 10 fold cross validation.

6. Conclusion

In this paper, the author have designed two databases one for Urdu and another for English language with five different accents spoken in Gilgit-Baltistan located at north of Pakistan. These databases are specifically designed to support biometric research in the area of telephone\mobile banking and in mobile network services. The security situation in Pakistan and particularly in this region compels to incorporate biometric solutions with the existing traditional verification system to provide a strong security mechanism. The author intends to develop a speaker verification system particularly for telephone/mobile bank services and mobile network services and the design of voice database is an essential requirement for speaker verification system. The designed database is in the form of single words, 10-16 digit strings and speaker's personal information related to bank and mobile network services. The designed database is preprocessed and then features have been extracted and binary file have been produced which was further used in speaker verification process. The subsets of the designed database for English as well as Urdu was trained and then tests using our own designed Multilayer Perceptron based speaker verification system with different accents and dialect. Some experiments have been performed and achieved overall accuracy rate of 83.4091% for English dataset and 80.0454% for Urdu dataset with 10 fold cross validation test scheme. It shows slight differences (4.0% with English and 7.4% with Urdu) in recognition accuracy if compared with the recently proposed multilayer perceptron (MLP) based SIS achieved 87.5% recognition accuracy.

References

[1] NSTC, "Biometrics ' Foundation Documents ,'" *Subcomm. Biometrics, Natl. Sci. Technol. Counc.*, pp. 1–166, 2006.

[2] A. K. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," vol. 14, no. 1, pp. 4–20, 2004.

[3] S. Memon, S. Ghulam, S. Shah, K. Khoumbati, and I. A. Ismaili, "Securing Sensitive eDatabases using Multi- Biometric Technology," no. December 2018, 2010.

[4] T. J. Ibm, "Biometric Recognition: Security and Privacy Concerns," pp. 33–42, 2003.

[5] F. Selection, "Speaker Verification:," no. January, pp. 42–48, 1990.

[6] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology • 1," pp. 4072–4075, 2002.

[7] E. Karpov, "Real-Time Speaker Identification Real-Time Speaker Identification Evgeny Karpov University of Joensuu Department of Computer Science Master ' s Thesis," no. January 2004, 2014.

[8] K. Khoumbati, "A Pashtu speakers database using accent and dialect approach Shahid Munir Shah *, Shahzad Ahmed Memon and Muhammad Moinuddin," vol. 4, no. 4, pp. 358–380, 2017.

[9] Z. Liu, Z. Wu, T. Li, J. Li, and C. Shen, "GMM and CNN Hybrid Method for Short," *IEEE Trans. Ind. Informatics*, vol. 14, no. 7, pp. 3244–3252, 2018.

[10] N. Chauhan and M. Chandra, "Speaker Recognition and Verification Using Artificial Neural Network," pp. 1147–1149, 2017.

[11] F. Thullier, B. Bouchard, and B. J. Menelas, "A Text-Independent Speaker Authentication System for Mobile Devices," pp. 1–22, 2017.

[12] N. Harte and E. Gillen, "TCD-TIMIT : An Audio-Visual Corpus of Continuous Speech," vol. 17, no. 5, pp. 603–615, 2015.

[13] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Br, D. Van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, J. Alam, A. Swart, J. Perez, A. Star, M. Lium, S. West, and S. Africa, "The RedDots Data Collection for Speaker Recognition," pp. 2996–3000, 2016.

[14] A. W. Abbas, N. Ahmad, and H. Ali, "Pashto Spoken Digits Database for the Automatic Speech Recognition Research," no. September, pp. 8–11, 2012.

[15] A. Alarifi and I. Alkurtass, "Arabic Text-Dependent Speaker Verification for Mobile Devices Using Artificial Neural Networks," pp. 350–353, 2011.

[16] G. Droua-hamdani, S. A. Selouani, M. Boudraa, and T. H. Boumediene, "Algerian Arabic Speech Database (ALGASD): Corpus design and automatic speech recognition application CORPUS DESIGN AND AUTOMATIC SPEECH," no. December, 2010.

[17] H. Sarfraz, S. Hussain, R. Bokhari, A. A. Raza, I. Ullah, S. Pervez, A. Mustafa, I. Javed, and R. Parveen, "Speech Corpus Development for a Speaker Independent Spontaneous Urdu Speech Recognition System," pp. 1–6, 2010.

[18] M. Alghamdi, F. Alhargan, M. Alkanhal, A. Alkhairy, M. Eldesouki, and A. Alenazi, "Saudi Accented Arabic Voice Bank," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 20, pp. 45–64, 2008.

[19] R. Kumar, S. P. Kishore, A. Gopalakrishna, R. Chitturi, S. Joshi, S. Singh, R. N. V Sitaram, G. Anumanchipalli, R. Chitturi, S. Joshi, R. Kumar, S. P. Singh, R. N. V Sitaram, and S. P. Kishore, "Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems by Vocabulary Speech Recognition Systems," no. July, 2007.

[20] J. Hennebert, H. Melin, D. Petrovska, and D. Genoud, "POLYCOST : A telephone-speech database for speaker," vol. 31, 2000.

[21] J. Ortega-garcia, J. Gonzalez-rodriguez, and V. Marrero-aguiar, "AHUMADA : A large speech corpus in Spanish for speaker characterization and identi ® cation q," vol. 31, pp. 255–264, 2000.

[22] R. Cole, M. Noel, and V. Noel, "THE CSLU SPEAKER RECOGNITION CORPUS," no. Icslp 98, pp. 98–101, 1998.

[23] "GANDALF - A SWEDISH TELEPHONE SPEAKER VERIFICATION DATABASE," pp. 3–6.

[24] M. Falcone, A. Gdlo, and V. B. Castiglione, "THE ' SIVA ' SPEECH DATABASE FOR SPEAKER VERIFICATION : DESCRIPTION AND EVALUATION," pp. 1902–1905, 1902.

[25] "TESTING WITH THE YOHO CD-ROM VOICE VERIFICATION CORPUS Joseph," pp. 341–344, 1995.

[26] J. Sinclair and C. Watson, "The Development of the Otago Speech Database," pp. 298–301, 1995.

[27] "ISCA Archive 4," no. September, pp. 817–820, 1995.

[28] A. S. Recognition, "ISCA Archive," no. April, pp. 39–42, 1994.

[29] J. J. Godfrey and E. C. Holliman, "SWITCHBOARD : Telephone Speech Corpus for Research and Development," pp. 517–520, 1992.

[30] https://en.wikipedia.org/wiki/Districts_of_Gilgit%E2%80%93Baltistan

[31] https://en.wikipedia.org/wiki/Gilgit-Baltistan

[32] Y. A. Ibrahim, J. C. Odiketa, and T. S. Ibiyemi, "Preprocessing technique in automatic speech recogntion for human computer interaction: an overview," *Ann. Comput. Sci. Ser.*, vol. XV, no. 1, pp. 186 – 191, 2017.

[33] https://www.jcbrolabs.org/speech-processing

[34] Nisha, "Voice Recognition Technique : A Review," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 5, no. V, pp. 262–268, 2017.

[35] B. Tomassetti, M. Verdecchia, and F. Giorgi, "NN5: A neural network based approach for the downscaling of precipitation fields - Model description and preliminary results," *J. Hydrol.*, vol. 367, no. 1–2, pp. 14–26, 2009.

[36] B. Choubin, S. Khalighi-Sigaroodi, A. Malekian, and Ö. Kişi, "Multiple linear regression, multi-layer perceptron network and adaptive neuro-fuzzy inference system for forecasting precipitation based on large-scale climate signals," *Hydrol. Sci. J.*, vol. 61, no. 6, pp. 1001–1009, 2016.