Accent Classification among Punjabi, Urdu, Pashto, Saraiki and Sindhi Accents of Urdu Language

Afsheen
CLE, KICS, UET, Lahore
afsheen@kics.edu.pk

Saad Irtza
CLE, KICS, UET, Lahore
saad.irtaza@kics.edu.pk

Mahwish Farooq
CLE, KICS, UET, Lahore
mahwish.farooq@kics.edu.pk

Sarmad Hussain
CLE, KICS, UET, Lahore
sarmad.hussain@kics.edu.pk

Abstract

Automatic Speech Recognition (ASR) is a key component in Human Computer Interaction (HCI) applications. Stability of ASR systems largely depends on accent, gender, age of speakers, background noise and channel variations. In this paper, a study has been conducted to classify five different accents of Urdu language spoken in Pakistan i.e. Punjabi, Urdu, Pashto, Saraiki and Sindhi. Speech data has been collected from native speakers of these accents. The five accents have been classified using mel frequency cepstral coefficient (MFCCs) and feature formants.

1. Introduction

The objective of an automatic speech recognition system is to convert speech into text. The performance of ASR systems depends on the training data on which acoustic models have been trained. Speaker dependent (SD) ASR systems perform better than speaker independent (SI) ASR systems, as the trained speech samples in SD ASR systems cover most of the possible acoustic variations of the targeted speakers. The acoustic variations in speakers are mainly due to age, gender and regional accents. Among these factors, accent is the most leading factor that contributes to a higher error rate in ASR systems.

Accent refers to the articulation pattern that a speaker follows to produce a particular sound and it is also related to speaker's first language, which affects speaker's perception and production of speech. Dialect and accent do not cause much trouble when communicating among humans but gives poor recognition results with ASR systems. ASR trained on a particular accent gives poor accuracy on cross accent speech data. It is a worth challenging task to build an ASR system which gives best accuracy for all accents of a language.

Present speech recognition systems use pronunciation dictionaries, which are based on pronunciation of words followed by a single accent group. Different accents give rise to multiple pronunciations of a word due to the use of different phonemes. But the addition of multiple pronunciations in ASR dictionary against a single word generates additional confusions, resulting in higher recognition error rate [1].

The most widely used approach to address accent variation problem in speech recognition systems is to build multiple accent dependent ASR systems. In order to send accented speech to the accent-specific speech recognizer, the above approach requires a pre-processing step of accent classification.

In different geographical regions of Pakistan, 59 languages are spoken. Based on the accent and native language, there are six prominent accents used in Pakistan, namely Urdu, Punjabi, Pashto, Saraiki, Sindhi and Balochi [2]. In the past, major research has been done on Urdu literature or grammar but no major work is done on the acoustic analysis of the different accents of Urdu. Although less than 8% of Pakistanis speak Urdu as their first language, but it is spoken and understood as a second language by almost all Pakistanis. In learning a second language, vowel articulation is the key factor to take care of [4]. In Phonetics, vowel is a sound least obstructed by any of the articulators during its articulation and are generally called the directors of the sounds [5]. Therefore, in our experiments we use vowels to classify five different accents of Urdu language.

The rest of the paper is structured as follows: Section 2 describes the past techniques used to classify accents of different languages, Section 3 details the two methodologies followed in this paper, to classify five different accents of Urdu language, Section 4

reports the results, Section 5 discusses the results and Section 6 concludes the paper.

2. Literature Review

The performance of speech recognition systems improve significantly over the last few years as this field has received much attention by researchers. In the past, many experiments have been performed in order to classify Chinese, Indian, Korean and American English accents.

A recent research shows that the native accent pronunciation dictionary, used in ASR systems, can be transformed to the accented speech dictionary by simply using knowledge of native language of the foreign speakers. The use of accent adapted dictionary to classify different Chinese and English accents reduces speech recognition error rate by 13.5% [7, 11]. Acoustic level adaptation techniques such as MLLR and an integration of both PDA and MLLR are also performed to trace the detailed acoustic changes in the speech of speakers due to change in speaking speed and style [11]. Results in the paper showed that both techniques are complementary.

Another approach used to identify Beijing, Shanghai, Guangdong and Taiwan accents from multi-accent Mandarin corpus, consisting of male and female speakers of each accent, is by using a set of Gaussian mixture models (GMMs) [8, 11]. The GMMs are used to estimate the probability that the input speech comes from a particular gender and accent. This approach shows that performance of accent dependent systems is generally better than that of accent independent systems.

Multivariate statistical analysis is also used to classify different accents [9]. The corpus used for the experiments consists of 4925 telephone utterances of American English being spoken by native speakers of 23 different languages. The study shows that classification results of support vector machine are better than other classifiers like ZeroR, Naive Bayes, Logistic Regression, SMO and K- Neatest Neighbors using MFCC, FBank and LPC features.

In order to classify Chinese, Indian, Korean and Croatian accents of English speakers, *Q* factor is introduced [10]. The paper reports that clustering of different accents based on formants does not assist to classify accents with high accuracy.

Different geographic regions of Sweden that share common accents are grouped using MFCC feature vectors [6]. The Swedish Speech data FD5000 telephone speech database was used for these experiments and different regions are clustered using Bhattacharya distance.

The above discussion shows that a lot of work has been done to classify Chinese and English accents but no research has been carried out before, to classify different accents spoken in Pakistan. In this paper we proposed a method to classify five widely spoken accents of Urdu language in Pakistan i.e. Punjabi, Urdu, Pashto, Saraiki and Sindhi.

3. Methodology

In this paper, we have classified Punjabi, Urdu, Pashto, Saraiki and Sindhi accents of Urdu language by using vowels. Based on two different acoustic features, two experiments have been conducted to classify above mentioned five accents. The vocabulary of the experiments consists of 139 district names of Pakistan. The proper nouns are chosen, as these are language independent words and are familiar to the speakers of all five languages. Urdu is the national and official language of Pakistan, therefore, district names can also be considered to be words/phrases of Urdu language. The training and testing corpus consists of 760 utterances from native speakers of above mentioned five languages. The speaker distribution is listed in Table 1.

Table 1: Speaker distribution

Accents	Male Speaker	Female Speaker
Punjabi	130	77
Urdu	45	56
Pashto	14	6
Saraiki	13	5
Sindhi	5	14

The speech is recorded over telephone channel at the sampling rate of 8 kHz with random background noise.

In Urdu, there exist 13 vowels: three short, three medial and seven long [3]. In our experiments, we used eight vowels, three short and five long, having sufficient data for analysis. The list of eight vowels along with Urdu letters, CISAMPA and IPA symbols is given in Table 2.

Table 2: Urdu vowel list with IPA and CISAMPA symbols

Sr. #	Urdu Letter	IPA	CISAMPA
1	ءَ،ـَ	ə	A
2	آ،ا	ɑ:	AA
3	ے	e:	AE
4	ِ	ɪ	I
5	ی	i:	II
6	و	o:	OO
7	ُ	ʊ	U
8	ؤ	u:	UU

The data of each of eight vowels listed above is given in table 3.

Table 3: Vowel data statistics

Vowel CISAMPA	Punjabi	Urdu	Pashto	Saraiki	Sindhi
A	276	109	295	80	142
AA	255	82	164	51	90
AE	69	21	37	20	20
I	125	64	52	12	29
II	65	25	57	19	30
OO	39	16	15	7	15
U	62	28	39	14	20
UU	36	23	22	8	13

In Experiment 1, we used feature formants, which are proved to be sensitive features to classify accents of Chinese language [10], to classify accents of Urdu language. Preliminary analysis of first and second formant frequencies (F1 and F2) shows differences in characteristics of vowels. Formants are calculated from the midpoint of each vowel to get a standard value using Praat, Praat is a package for analysis of speech in phonetics. The mean and standard deviation of F1 and F2 for five accents of each vowel have been calculated. Each accent of a vowel is compared with the remaining four accents of the same vowel and all vowels of Urdu accent, to find out how much these accents and vowels are different from each other.

In Experiment 2, we have used mel frequency cepstral coefficients (MFCCs) and formants (F1 and F2) as acoustic features to classify five different accents of Urdu language by calculating distance between them. MFCCs are commonly used as features in speech recognition. They are also common in speaker recognition, which is the task of recognizing people from their voices. The middle frame of 100 samples of each vowel is used to calculate MFCCs. These cepstral features are calculated using HTK toolkit. The Hidden Markov Model Toolkit (HTK) is a portable toolkit for building and manipulating hidden Markov models. Only the first twelve MFCCs have been used here to classify accents of Urdu language. The MFCC data is first normalized using Weka (Weka is a collection of machine learning algorithms for data mining tasks.) and then mean MFCC vector for each accent of a vowel has been computed.

Using 12 dimensional mean MFCC vector, the distance between two accents has been computed by Euclidian distance formula. For example, distance between Punjabi and Urdu accents (of a vowel of Urdu language) is computed using equation 1.

$$d_{(Pun,Urd)} = \sqrt{\begin{matrix}(mfcc_1^{PUN} - mfcc_1^{URD})^2 + ... + \\ (mfcc_7^{PUN} - mfcc_7^{URD})^2 + ... + \\ (mfcc_{12}^{PUN} - mfcc_{12}^{URD})^2\end{matrix}} \quad (1)$$

The distance between Punjabi and Urdu accents of a vowel of Urdu language using formants has been calculated using two dimensional Euclidian distance formula given in equation 2.

$$d_{(Pun,Urd)} = \sqrt{(F_1^{PUN} - F_1^{URD})^2 + (F_2^{PUN} - F_2^{URD})^2} \quad (2)$$

The distances calculated using these two equations are then compared and analyzed in following sections.

4. Classification Results

The classification results of experiment 1 are shown below in graphs. Each vowel with the five accents is plotted in a separate graph along with remaining vowels of Urdu accent and each plot shows standard deviation of formants (F1 and F2) of a particular vowel from their mean.

The figure 1 shows the standard deviation of F1 and F2 of short vowel "A" (of above mentioned five accents) from their mean along with other vowels of Urdu language.

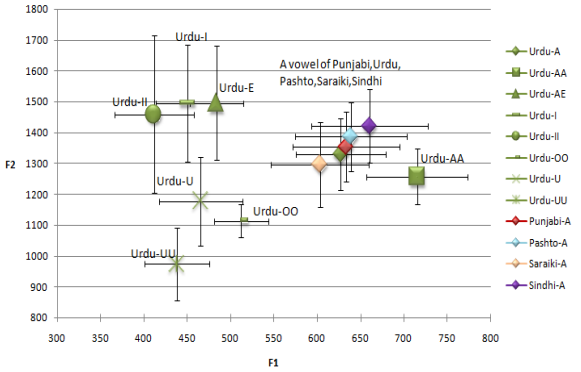

Figure 1: Comparison of A vowel of 5 accents and all vowels of Urdu

The figure 2 shows the standard deviation of formant frequencies of long vowel "AA" (of above mentioned five accents) from their mean along with the remaining vowels of Urdu language.

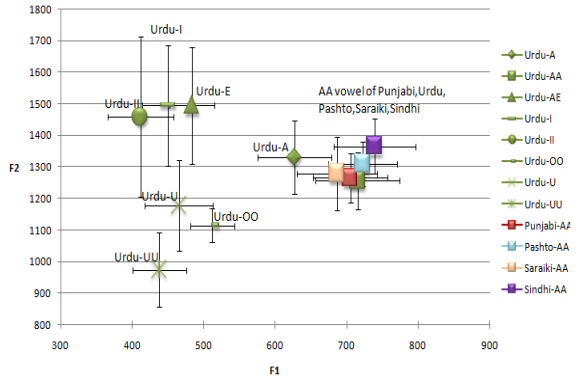

Figure 2: Comparison of AA vowel of 5 accents and all vowels of Urdu

The figure 3 shows standard deviation of formant frequencies of long vowel "AE" of five accents from their mean along with other vowels of Urdu language.

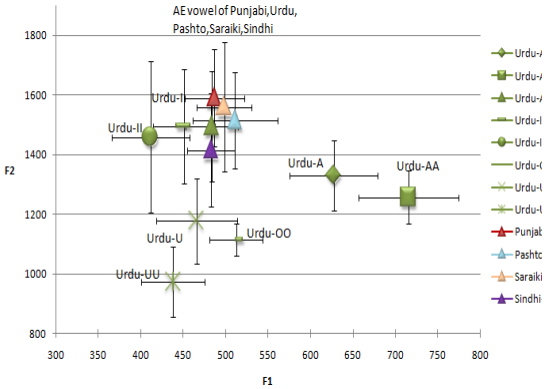

Figure 3: Comparison of AE vowel of 5 accents and all vowels of Urdu

The figure 4 below shows the standard deviation of formant frequencies of short vowel "I" (of above mentioned five accents) from their mean along with remaining vowels of Urdu language.

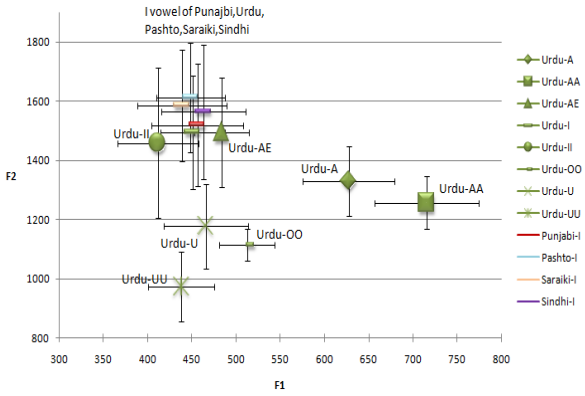

Figure 4: Comparison of I vowel of 5 accents and all vowels of Urdu

The figure 5 below shows standard deviation of F1 and F2 of long vowel "II" (of Punjabi, Urdu, Pashto, Saraiki and Sindhi accents) from their mean along with remaining vowels of Urdu language.

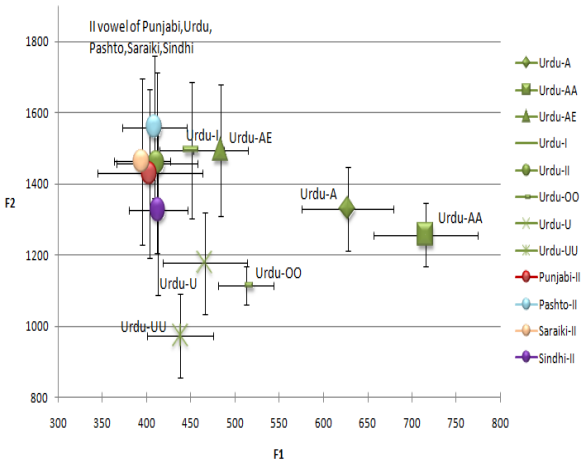

Figure 5: Comparison of II vowel of five accents and all vowels of Urdu

Figure 6 below shows standard deviation of formant frequencies of "OO" vowel of five accents from their mean.

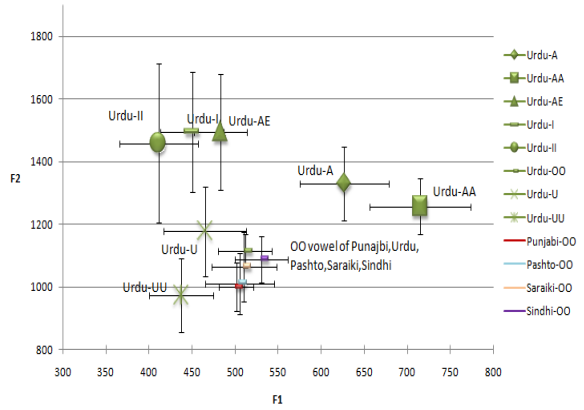

Figure 6: Comparison of OO vowel of 5 accents and all vowels of Urdu

Figure 7 shows the standard deviation of formants of "U" vowel (of five accents) from their mean along with the remaining vowels of Urdu language.

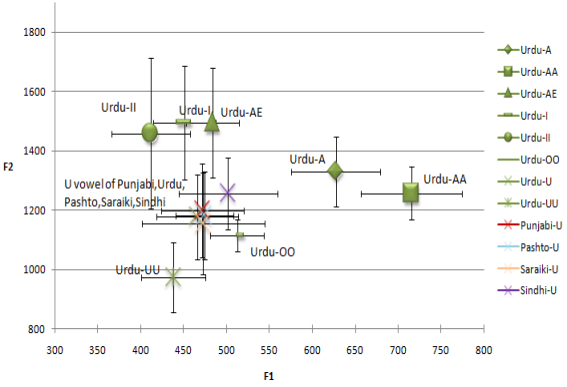

Figure 7: Comparison of U vowel of 5 accents and all vowels of Urdu

Figure 8 shows the standard deviation of formant frequencies of "UU" vowel of five accents from their mean.

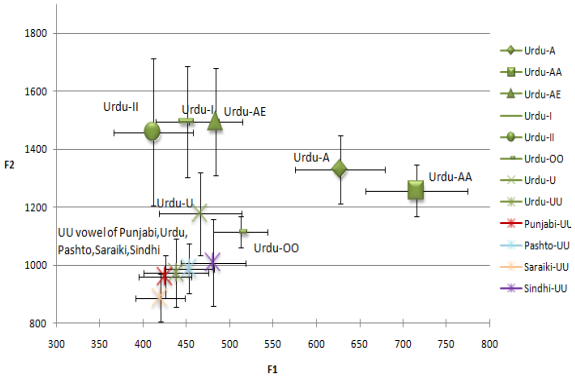

Figure 8: Comparison of UU vowel of 5 accents and all vowels of Urdu

In Experiment 2, distances calculated by Euclidian formulas have been used to classify accents. Table 4 shows the distance between different accents, computed by summing respective distances of these accents over all vowels.

Table 4: Distance between different accents of Urdu language

Accents	MFCC	Formants
Punjabi-Urdu	0.637308	0.1137
Punjabi-Pashto	0.713092	0.287106
Punjabi-Saraiki	1.47057	0.426182
Punjabi-Sindhi	1.394903	0.392127
Urdu-Pashto	0.648211	0.38703
Urdu-Saraiki	1.514395	0.402592
Urdu-Sindhi	1.324604	0.499803
Pashto-Saraiki	1.613643	0.39677
Pashto-Sindhi	1.317765	0.300893
Saraiki-Sindhi	2.263357	0.670466

See Appendix for distance between different accents of each vowel.

5. Discussion

The figures 1 to 8 of experiment 1 show that using F1 and F2, all vowels of the Urdu accent can be classified, with an exception of short vowel I which is being confused with long vowels II and AE of Urdu language (shown in all figures 1 to 8).

But on the other hand it is very clear from the above graphs that using F1 and F2, it is difficult to classify Punjabi, Urdu, Pashto, Saraiki and Sindhi accents of A, AA, AE, I, II, OO, U and UU vowels.

Table 4 of experiment 2 shows that distance calculated between two accents using MFCCs is always greater than the distance calculated using formant frequencies. Therefore, it can be concluded that on the basis of distance calculated using MFCCs, the probability of an accent to get confused with other accents is minimal. As given in table 4, the calculated distance between "Urdu and Saraiki" accents based on formant frequencies values is 0.402592 while based on MFCC values is 1.514395 which is almost three times greater than that calculated using formant frequencies. Therefore, MFCCs can be used to classify Punjabi, Urdu, Pashto, Saraiki and Sindhi accents of Urdu language.

6. Conclusions

Above mentioned results of the two experiments show that two dimensional formant features F1 and F2 are not sufficient to classify Punjabi, Urdu, Pashto, Saraiki and Sindhi accents of Urdu language spoken in different geographical regions of Pakistan. Therefore, there is a need to explore more dimensions of speech data. This need has been accomplished by using twelve dimensional Mel-frequency cepstral features, that retain accent related information of the speaker, to classify above mentioned five accents. Our results show that MFCC vectors can be used to classify Punjabi, Urdu, Pashto, Saraiki and Sindhi accents of Urdu language.

7. Acknowledgement

This work has been conducted through the project, Enabling Information Access for Mobile based Urdu Dialogue Systems and Screen Readers supported through a research grant from ICTRnD Fund, Pakistan.

8. References

[1] M. Lincoln, S. Cox and S. Ringland, "A comparison of two unsupervised approaches to accent identification", ICSLP, Dec , 1998.

[2] T. Rahman, "Language, ideology and power : language learning among the Muslims of Pakistan and North India", Oxford University Press, Karachi, 2002.

[3] S. M. Saleem, S. Anjum and J. Jalibi, "Oxford Urdu English Dictionary", Oxford University Press, Lahore, 2013.

[4] A. Ali, "Pakistani Pronunciation of RP vowels: An Exploratory Study ", ELT, NUML, Lahore.

[5] IPA, "Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet", Cambridge University Press, Cambridge, July 8, 1999.

[6] G. Salvi, "Accent Clustering in Swedish Using the Bhattacharyya Distance", in proc. International Congress of Phonetic Sciences (ICPhS), Barcelona, 2003.

[7] LIU Wai Kat and Pascale FUNG, "Fast Accent Identification and Accented Speech Recognition", IEEE, 1999.

[8] T. Chen, C. Huang, E. Chang and J. Wang, "Automatic Accent Identification using Gaussian Mixture Models", in *Proc. ASRU*, 2001.

[9] P. Chen, J. Lee and J. Neidert, "Foreign Accent Classification", CS 229, Fall 2011.

[10] D. Stantic and H. Jo, "Accent Identification by Clustering and Scoring Formants", World Academy of Science, Engineering and Technology, Vol: 63, Mar. 25, 2012.

[11] C. Huang, T. Chen and E. Chang, "Accent Issues in Large Vocabulary Continuous Speech Recognition", International Journal of Speech Technology 7, Pg: 141-153, 2004.

9. Appendix

The distance between "A" vowel of Punjabi, Urdu, Pashto, Saraiki and Sindhi accents of Urdu language calculated using formants and MFCCs is given in table 5.

Table 5: Distance between "A" vowel of different accents

Accents	MFCC	Formants
Punjabi-Urdu	0.06856	0.0151051
Punjabi-Pashto	0.09128	0.020927132
Punjabi-Saraiki	0.11807	0.049725722
Punjabi-Sindhi	0.25810	0.062430506
Urdu-Pashto	0.11247	0.035993314
Urdu-Saraiki	0.15598	0.037388876
Urdu-Sindhi	0.23204	0.075529297
Pashto-Saraiki	0.17023	0.0678984
Pashto-Sindhi	0.27159	0.047876065
Saraiki-Sindhi	0.29642	0.112110215

Table 6 contains the distance between "AA" vowel of Punjabi, Urdu, Pashto, Saraiki and Sindhi accents of Urdu language.

Table 6: Distance between "AA" vowel of different accents

Accents	MFCC	Formants
Punjabi-Urdu	0.068566	0.0151051
Punjabi-Pashto	0.091284	0.020927132
Punjabi-Saraiki	0.11807	0.049725722
Punjabi-Sindhi	0.258101	0.062430506
Urdu-Pashto	0.112472	0.035993314
Urdu-Saraiki	0.155985	0.037388876
Urdu-Sindhi	0.232043	0.075529297
Pashto-Saraiki	0.170231	0.067898458
Pashto-Sindhi	0.27159	0.047876065
Saraiki-Sindhi	0.296429	0.112110215

Table 7 contains the distance between "AE" vowel of Punjabi, Urdu, Pashto, Saraiki and Sindhi accents of Urdu language.

Table 7: Distance between "AE" vowel of different accents

Accents	MFCC	Formants
Punjabi-Urdu	0.06856	0.0151051
Punjabi-Pashto	0.09128	0.020927132
Punjabi-Saraiki	0.11807	0.049725722
Punjabi-Sindhi	0.25810	0.062430506
Urdu-Pashto	0.11247	0.035993314
Urdu-Saraiki	0.15598	0.037388876
Urdu-Sindhi	0.23204	0.075529297
Pashto-Saraiki	0.17023	0.067898458
Pashto-Sindhi	0.27159	0.047876065
Saraiki-Sindhi	0.29642	0.112110215

The distance between "I" vowel of Punjabi, Urdu, Pashto, Saraiki and Sindhi accents of Urdu language is given in table 8.

Table 8: Distance between "I" vowel of different accents

Accents	MFCC	Formants
Punjabi-Urdu	0.08632	0.013677
Punjabi-Pashto	0.08784	0.044865
Punjabi-Saraiki	0.22327	0.055401
Punjabi-Sindhi	0.12412	0.040967
Urdu-Pashto	0.06215	0.05581
Urdu-Saraiki	0.20928	0.058085
Urdu-Sindhi	0.12569	0.054643
Pashto-Saraiki	0.22059	0.038615
Pashto-Sindhi	0.10059	0.031453
Saraiki-Sindhi	0.27481	0.066827

The distance between "II" vowel of Punjabi, Urdu, Pashto, Saraiki and Sindhi accents of Urdu language is given in table 9.

Table 9: Distance between "II" vowel of different accents

Accents	MFCC	Formants
Punjabi-Urdu	0.086322	0.013677
Punjabi-Pashto	0.087848	0.044865
Punjabi-Saraiki	0.223272	0.055401
Punjabi-Sindhi	0.12412	0.040967
Urdu-Pashto	0.062159	0.05581
Urdu-Saraiki	0.209288	0.058085
Urdu-Sindhi	0.125695	0.054643
Pashto-Saraiki	0.22059	0.038615
Pashto-Sindhi	0.100599	0.031453
Saraiki-Sindhi	0.274814	0.066827

The distance between different accents of long vowel "OO" is given in table 10.

Table 10: Distance between "OO" vowel of different accents

Accents	MFCC	Formants
Punjabi-Urdu	0.08632	0.013677
Punjabi-Pashto	0.08784	0.044865
Punjabi-Saraiki	0.22327	0.055401
Punjabi-Sindhi	0.12412	0.040967
Urdu-Pashto	0.06215	0.05581
Urdu-Saraiki	0.20928	0.058085
Urdu-Sindhi	0.12569	0.054643
Pashto-Saraiki	0.22059	0.038615
Pashto-Sindhi	0.10059	0.031453
Saraiki-Sindhi	0.27481	0.066827

The distance between different accents of short vowel "U" is given in table 11.

Table 11: Distance between "U" vowel of different accents

Accents	MFCC	Formants
Punjabi-Urdu	0.08632	0.013677
Punjabi-Pashto	0.08784	0.044865
Punjabi-Saraiki	0.22327	0.055401
Punjabi-Sindhi	0.12412	0.040967
Urdu-Pashto	0.06215	0.05581
Urdu-Saraiki	0.20928	0.058085
Urdu-Sindhi	0.12569	0.054643
Pashto-Saraiki	0.22059	0.038615
Pashto-Sindhi	0.10059	0.031453
Saraiki-Sindhi	0.27481	0.066827

The distance between different accents of long vowel "UU" is given in table 12.

Table 12: Distance between "UU" vowel of different accents

Accents	MFCC	Formants
Punjabi-Urdu	0.08632	0.013677
Punjabi-Pashto	0.08784	0.044865
Punjabi-Saraiki	0.22327	0.055401
Punjabi-Sindhi	0.12412	0.040967
Urdu-Pashto	0.06215	0.05581
Urdu-Saraiki	0.20928	0.058085
Urdu-Sindhi	0.12569	0.054643
Pashto-Saraiki	0.22059	0.038615
Pashto-Sindhi	0.10059	0.031453
Saraiki-Sindhi	0.27481	0.066827