Urdu Speech Corpus and Preliminary Results on Speech Recognition

Hazrat Ali and Nasir Ahmad and Abdul Hafeez

Abstract Language resources for Urdu language are not well developed. In this work, we summarize our work on the development of Urdu speech corpus for isolated words. The Corpus comprises of 250 isolated words of Urdu recorded by ten individuals. The speakers include both native and non-native, male and female individuals. The corpus can be used for both speech and speaker recognition tasks. We also report our results on automatic speech recognition task for the said corpus. The framework extracts Mel Frequency Cepstral Coefficients along with the velocity and acceleration coefficients, which are then fed to different classifiers to perform recognition task. The classifiers used are Support Vector Machines, Random Forest and Linear Discriminant Analysis. Experimental results show that the best results are provided by the Support Vector Machines with a test set accuracy of 73%. The results reported in this work may provide a useful baseline for future research on automatic speech recognition of Urdu.

1 Introduction

Urdu is the national language of Pakistan understood by approximately 75% population of the country. Globally, Urdu speakers accumulate to around 70 million speakers [1]. Urdu language shares its vocabulary with many other Asian languages

Hazrat Ali (corresponding author)
Department of Electrical Engineering, COMSATS Institute of Information Technology Abbottabad
e-mail: hazratali@ciit.net.pk

Nasir Ahmad
Department of Computer Systems Engineering, University of Engineering and Technology Peshawar e-mail: n.ahmad@uetpeshawar.edu.pk

Abdul Hafeez
Department of Computer Systems Engineering, University of Engineering and Technology Peshawar e-mail: abdul.hafeez@uetpeshawar.edu.pk

including Arabic, Farsi, and Turkish. A framework for automatic speech recognition of Urdu can be helpful to contribute towards speech recognition of other similar languages. Unfortunately, for Urdu, lack of standard corpora and baseline approaches have been the bottleneck to make advancements on speech recognition research of Urdu.

Recently, there has been some work reported on the automatic speech recognition of Urdu. While these works have their own significance, either the corpus used in the work has not been specified or it is too limited to be generalized for diverse set of speakers. For example, Sarfraz et al. [2] has presented an Urdu corpus covering speakers only from a single city. Similarly, another speech corpus for Urdu has been presented in [3] however, it is not clear if the corpus is available for public use. Akram et al [4] have presented a continuous speech recognition system for Urdu however, the corpus used in the work is not identified. Information on training and test sets size is also missing. Besides, the accuracy reported by [4] does not exceed 54%. For Urdu digits recognition, a multilayer perceptron has been used by Ahad et al [5], presenting a framework for speech recognition of digits from 0 to 9. However, the work in [5] is based on speech data from a single speaker and thus, cannot be generalized for a diverse set of speakers. Another work reported for Urdu digits recognition is by Hasnain et al [6] with higher accuracy performance. It is not clear if the accuracy measures in [6] are reported for training set only or for unknown test set. The use of hidden markov models for Urdu speech recognition has been reported in [7]. The model used in [7] treats every single word as a single phoneme. This may work for words of shorter duration but may undergo degradation if the words have longer duration.

For the Urdu dataset presented in this work, previous work has used features from discrete wavelet transform with linear discriminant analysis (LDA) [8], MFCC features with LDA [9], [10]. In this work, we describe the Urdu corpus for the general understanding of the reader, and make it freely available for academic research use. Further, we report results on speech recognition task for this corpus with three different classifiers namely; Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and Random Forests (RF). The rest of the paper is organized as follows: In Section 2, we describe the development of the corpus and the way the audio files are organized. In Section 3, we discuss the extraction of MFCC features as well as the three classifiers used on the features. The results obtained are provided in Section 4. Finally, the paper is concluded in Section 5.

2 The Corpus

2.1 Corpus Development

The words recorded for this corpus are selected from the most frequently used words in Urdu literature, as summarized by the center of language engineering (CLE) [11].

These words include those which are used in everyday life, and digits from 0 to 9. Wherever possible, an attempt has been made to include antonyms or synonyms of various words. These words were then recorded by ten speakers with Sony Linear PCM Recorder. Any mistake in recording process was compensated by re-recording. The recording was accomplished in multiple sessions. Speakers coming for recording vary in age, origin and first language, ensuring that a diversity is achieved in the corpus. The recorded files are stored with sampling rate of 16000 Hz in *.wav* format. Average duration for each recording is half a second.

2.2 Corpus Organization

The master directory in this corpus contains ten sub-directories and each sub-directory corresponds to the individual speaker. Each sub-directory contains 250 audio files in *.wav* format. The information about each individual speaker is available in the sub-directory name. For example, the sub-directory named AKMNG2 corresponds to speaker AK (speakers are represented by combination of two letters, thus ranging from AA to AK and can be extended as well). The speaker gender information is contained in the third letter M (M corresponds to male and F corresponds to female). The fourth letter N in the sub-directory name denotes that the speaker is a non-native speaker (N represents that the speaker is non-native while Y represents that the speaker is a native speaker). The last two letters comprising of a character and a number correspond to the age of the speaker. Age ranges are from G1 (20 25 years) through G2 (26 30 years). Each file name provides information on speaker as well as the word number. The words are numbered from 001 to 250, appended to the sub-directory name to form the file name. An overview of the corpus organization is shown in Figure 1. Access to the corpus can be requested by writing email to the first author.

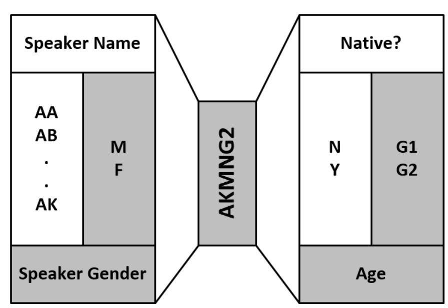

Fig. 1 Speakers are named from AA to AK. Speaker gender is defined by M for male and F for female. In the native field, N represents that speaker is non-native speaker and Y represents that speaker is native. Speakers belong to age group G1 or G2.

3 Experimental Setup

3.1 Features Extraction

For the dataset, we randomly divide the audio files into training and test sets with a ratio of 7:3. We then calculate the mel frequency cepstral coefficients (MFCC) for each audio file. The mel frequency cepstral coefficients have been in wide use by the speech processing community both for speech and speaker recognition applications [10], [12], [13], [14]. The MFCCs are based on mel-scale, a non-linear scale with logarithmic behavior [12]. Frequency mapping on a mel scale is given by equation:

$$f_{mel} = 2595 \times \log(1 + \frac{f}{700Hz}) \tag{1}$$

where, f_{mel} is the mel-scale frequency and f is the linear frequency in Hz. Different methods for calculation of the MFCCs can be seen in [12], [13], [14]. For MFCC calculation in this work, the Malcom's implementation has been used, as also used in [10],[21]. The steps involved in MFCC features extraction are demonstrated through algorithm shown in Figure 2. For each audio file, 12 coefficients are computed followed by concatenation of delta and delta-delta coefficients. Thus, each file is represented by 36 features set.

3.2 Support Vector Machines

Support Vector Machine (SVM) is a kernel based algorithm. SVMs are popularly used for discriminative classification. SVMs can be traced back to the work Boser et al [15]. They were used for automatic recognition of handwritten characters [16] and thus, became popular. In SVMs, the data of different classes is separated by hyper planes such that the distance for data of each class is maximized (for binary classification, the distance of samples of both the classes from the hyper plane will be maximized). Thus, SVMs are classifiers with large-margin boundary. For SVMs, the important feature is the kernel function used. The kernel function might be linear, polynomial or Gaussian. The strength of SVMs lie in the fact that they do not suffer the problem of local optima. However, attention is required to select the suitable kernel function. For SVMs, the function is given by sums of the kernel function $K(x_m, x_n)$:

$$f(x) = \sum_{m=1}^{N} \alpha_m t_m K(x_m, x_n) + d \tag{2}$$

where t_m denotes the ideal outputs, $\sum_{m=1}^{N} \alpha t_m = 0$ and α_m is greater than zero. Ideally, the outputs are $+1$ or -1 representing the corresponding class to which the data sample belongs. The output class for any data sample is decided by comparison of

Algorithm 1 Algorithm for MFCC calculation

for $i = 0$ to *No. of Frames* **do**
  Calculate Power Spectrum
end for
for $i = 0$ to *No. of Filter Coefficients* **do**
  Mel Filter Bank Calculation
  Apply the filter bank to the spectrum
  $sumE \leftarrow \sum$ *the energy in each filter*
  $logE \leftarrow log(sumE)$
end for
Discrete Cosine Transformation for the $logE$
Retain N coefficients
if $D \neq 0$ **then**
  repeat
    $Coeff(j) = Coeff(j) - Coeff(j-1)$ {calculate delta coefficients}
    $j \leftarrow j-1$
  until $j = 0$
else
  $Coeff \leftarrow Coeff$
end if
if $DD \neq 0$ **then**
  repeat
    $Coeff(j) = Coeff(j) - Coeff(j-1)$ {calculate delta-delta coefficients}
    $j \leftarrow j-1$
  until $j = 0$
else
  $Coeff \leftarrow Coeff$
end if

Fig. 2 MFCC Calculation (as in [10],[21])

value of $f(x)$ with a threshold value. Generally, the onv-vs-all approach is used if we have more than two classes of data (i.e., a multi-class problem). In our work on the use of SVM, we utilize the libSVM library [17]. We use the Gaussian RBF kernel, which for two data points, can be defined as below:

$$K(x_m, x_n) = exp(\gamma(\|x_m - x_n\|)^2) \tag{3}$$

We run a grid search and choose the γ and regularization constant C (hyper-parameters) after running the experiment over multiple iterations.

4 Random Forest

In computer vision, decision trees have been remarkable and successful for classification as well as regression tasks. Decision trees have previously been used as stand-alone approach. When an ensemble of multiple decision trees is used for decision making, they form a random forest classifier (or random decision forest classifier). RF has been successfuly used on hand-written digits recognition task as reported in [18], Other work on the use of RF classification is reported in [19]. For classification through RF classifier, the process involves training of the trees with features selected randomly. In order to make a final prediction, average is then calculated for the posteriors of each class output. To perform speech recognition using a RF classifier, we feed the MFCCs to train the classifier comprising of 300 trees.

4.1 Linear Discriminant Analysis

Linear Discriminant Analyis (LDA) [20] is popular for dimensionality reduction as well as for classification tasks. When LDA is applied to a data, it transforms the data into a matrix Θ. *"LDA tends to maximize the ratio between the inter-class variance and intra-class variance"* [10]. Classification is achieved such that for each test example, calculation of Euclidean distance is performed. So, for a particular problem, if we have n distinct classes, there will be n number of Euclidean distances to be calculated over each test example. The class is predicted for the prediction for which the corresponding distance is the smallest. LDA transformation can be represented by $S(\Theta)$;

$$S(\Theta) = \frac{|\Theta^T \Psi \Theta|}{|\Theta^T W \Theta|} \tag{4}$$

where, the within-class variance is given by W and variance matrix is given by Ψ, $|.|$ is the value of the determinant. For the speech recognition task, we use LDA with the MFCC features and compare the results with those obtained for RF and SVM classifiers.

5 Experimental Results

Once the recognition is performed, the prediction results are put into into a confusion matrix for the test data. For N number of words, the size of the confusion matrix is $N \times N$ matrix. $ConfM$ provides a general representation of the confusion matrix.

$$ConfM = \begin{matrix} c_{11} & c_{12} & c_{13} \cdots & c_{1N} \\ c_{21} & c_{22} & c_{23} \cdots & c_{2N} \\ c_{31} & c_{32} & c_{33} \cdots & c_{3N} \\ . & . & .... & . \\ . & . & .... & . \\ c_{N1} & c_{N2} & c_{N3} \cdots & c_{NN} \end{matrix} \tag{5}$$

In the above confusion matrix, $ConfM$, correct word recognition is shown by the values in the diagonal entries i.e., c_{ij} for $i = j$. Conversely, the number of false predictions for a test word is provided by the enteries in the non-diagonal position of the matrix, i.e., c_{ij} for $i \neq j$. The SVM classifier has resulted in an overall test

Table 1 Recognition Accuracy in Percentage

S. No	Word Number	Recognition Rate (SVM classifier)	Recognition Rate for RF	Recognition Rate for LDA
1	001	100%	66.67%	100%
2	002	66.67%	33.33%	33.33%
3	003	66.67%	100%	100%
4	004	100%	66.67%	66.67%
5	005	66.67%	66.67%	66.67%
6	006	33.33%	100%	66.67%
7	007	66.67%	66.67%	66.67%
8	008	100%	66.67%	0%
9	009	66.67%	33.33%	33.33%
10	010	66.67%	33.33%	100%

accuracy of 73%. Compared to this, the overall accuracy obtained by the random forest classifier as well as the LDA classifier is 63%. Figure 3, Figure 4 and Figure 5 show the confusion matrix plots for the three classification methods namely, SVM classification, Random Forest classification and LDA classification respectively. For each digit, the corresponding recognition rates for SVM classifier, LDA classifier and Random Forest classifier are shown in Table 1. It is obvious from the results that accuracy achieved by LDA classifier is same as the accuracy for RF classifier, i.e., an overall accuracy of 63%. From the confusion matrix, it can be noted that for the word number 7, the LDA classifier has resulted in 0% accuracy (as the empty 7th column can be seen in Figure 5).

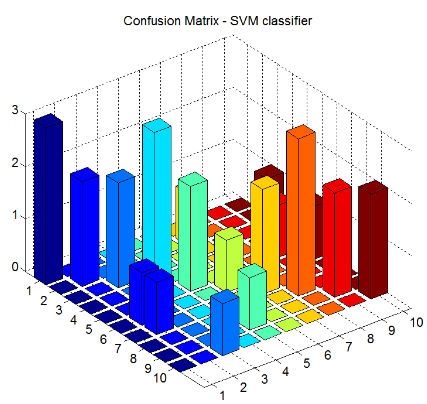

Fig. 3 Confusion matrix plot (For SVM classifier

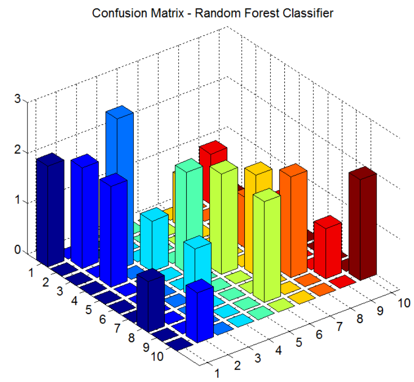

Fig. 4 Confusion matrix plot (for Random Forest classifier

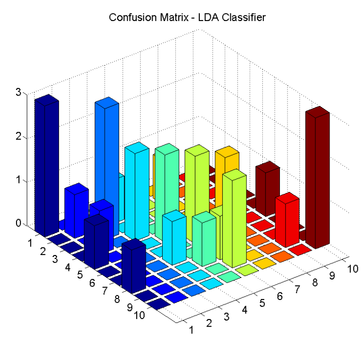

Fig. 5 Confusion matrix plot (for LDA classification)

6 Conclusion

In this paper, we have reported our work on the development of Urdu corpus comprising of 250 words spoken by ten speakers. We further reported our results for a speech recognition task with MFCC features extracted from the audio data. For classification purpose, we have used three classifiers namely; SVM, RF and LDA and reported percentage accuracy for each classifier. Experimental results have shown that SVM has performed well on this particular dataset with a 73% recognition accuracy compared with the 63% accuracy for RF and LDA. These results can serve as a reference baseline for further advancement on the Urdu dataset. The dataset is available for academic/research use and thus, a direct comparison of results is conceivable. For future work, firstly, the corpus can be extended by including more recordings and extending the list of words thus, covering a more diverse range of dialects, speakers age and vocabulary. Secondly, more robust speech recognition models can be used on the Urdu data set, such as Hidden Markov Model and deep learning approaches as these can arguably be more robust providing much higher accuracy. Thirdly, an ensemble model which combines classification scores from different classifiers can also be explored for this data, for example, a late fusion approach as used in [22].

References

1. “Ethnologue.” [Online]. Available: http://www.ethnologue.com/show_country.asp?name=PK.
2. H. Sarfraz et al., “Speech Corpus Development for a Speaker Independent Spontaneous Urdu Speech Recognition System,” in Proceedings of the O-COCOSDA, Kathmandu, Nepal, 2010. doi: 10.1109/ivtta.1994.341535
3. A. A. Raza, S. Hussain, H. Sarfraz, I. Ullah, and Z. Sarfraz, “Design and development of phonetically rich Urdu speech corpus”, in Proceeding of International Conference on Speech Database and Assessments, COCOSDA, 2009, pp. 38-43. doi: 10.1109/icsda.2009.5278380
4. Akram M U, and Arif M., “Design of an Urdu Speech Recognizer based upon acoustic phonetic modeling approach”. In: Proceedings of 8th International Multitopic Conference (INMIC) 2004,. Dec 2004, 91-96. doi: 10.1109/inmic.2004.1492852
5. Ahad A, Fayyaz A, Mehmood T., “Speech recognition using multilayer Perceptron”. In: Proceedings. IEEE Students Conference, ISCON 02. Aug 2002, pp. 103-109. doi: 10.1109/iscon.2002.1215948
6. Hasnain S, Awan M. “Recognizing spoken urdu numbers using fourier descriptor and neural networks with matlab”. In: Second International Conference on Electrical Engineering, (ICEE 2008). March 2008, pp. 1-6. doi: 10.1109/icee.2008.4553937
7. Ashraf J, Iqbal N, Sarfraz Khattak N, Mohsin Zaidi, “A. Speaker independent Urdu speech recognition using HMM”. In: The 7th International Conference on Informatics and Systems (INFOS 2010). March 2010, pp. 1-5. doi: 10.1007/978-3-642-13881-2_14
8. Ali H, Ahmad N, Zhou X, Iqbal K, Ali S M. “DWT features performance analysis for automatic speech recognition of Urdu”. SpringerPlus, 2014, 3(1): 204. doi: 10.1186/2193-1801-3-204
9. H. Ali, N. Ahmad, X. Zhou, “Automatic Speech Recognition of Urdu Words using Linear Discriminant Analysis”, Journal of Intelligent and Fuzzy Systems, vol. 28, no. 5, pp. 2369 - 2375, June 2015. doi: 10.3233/ifs-151554
10. H. Ali, A. Jianwei, K. Iqbal, “Automatic Speech Recognition of Urdu Digits with Optimal Classification Approach”, International Journal of Computer Applications, vol. 118, no. 9, May 2015. doi: 10.5120/20770-3275
11. “Center for Language Engineering.” [Online]. Available: www.cle.org.pk.
12. Molau S, Pitz M, Schluter R, Ney H., “Computing mel-frequency cepstral coefficients on the power spectrum”. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 01). 2001, pp. 73-76. doi: 10.1109/icassp.2001.940770
13. Han W, Chan C F, Choy C S, Pun K P., “An efficient MFCC extraction method in speech recognition”. In: Proceedings. IEEE International Symposium on Circuits and Systems, ISCAS 2006. May 2006. doi:10.1109/iscas.2006.1692543
14. Kotnik B, Vlaj D, Horvat B. “Efficient noise robust feature extraction algorithms for distributed speech recognition (dsr) systems”. International Journal of Speech Technology, 2003, vol. 6 np. 3, pp. 205-219
15. Boser B E, Guyon I M, Vapnik V N., “A training algorithm for optimal margin classifiers”. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT 92. 1992, pp. 144-152. doi: 10.1145/130385.130401

16. Bottou L, Cortes C, Denker J, Drucker H, Guyon I, Jackel L, LeCun Y, Muller U, Sackinger E, Simard P, Vapnik V. Comparison of classifier methods: a case study in handwritten digit recognition. In: Proceedings of the 12th IAPR International. Conference on Pattern Recognition, Oct 1994, pp. 77-82. doi: 10.1109/icpr.1994.576879
17. Chang C C, Lin C J., "LIBSVM: A library for support vector machines". ACM Transactions on Intelligent Systems and Technology, 2011, vol. 2 pp. 27:1-27:27. doi: 10.1145/1961189.1961199. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm
18. Ho T K. "Random decision forests". In: Proceedings of the Third International Conference on Document Analysis and Recognition. Aug 1995, vol 1. pp. 278-282. doi: 10.1109/icdar.1995.598994
19. Caruana R, Karampatziakis N, Yessenalina A. "An empirical evaluation of supervised learning in high dimensions". In: Proceedings of the 25th International Conference on Machine Learning, ICML 08. 2008, pp. 96-103. doi: 10.1145/1390156.1390169
20. Balakrishnama S, Ganapathiraju A. Linear discriminant analysis; a brief tutorial. http://www.music.mcgill.ca [Online] Accessed: 10 February, 2016
21. H. Ali, X. Zhou, and S. Tie, "Comparison of MFCC and DWT features for automatic speech recognition of Urdu". In International Conference on Cyberspace Technology (CCT 2013) pp. 154-158, November 2013, Beijing, China. doi:10.1049/cp.2013.2112
22. H. Ali, A. S. d'Avila Garcez, S. N. Tran, X. Zhou and K. Iqbal, "Unimodal late fusion for NIST i-vector challenge on speaker detection," Electron. Lett., vol. 50, no. 15, pp. 1098-1100, Jul. 2014. doi: 10.1049/el.2014.1207