International Journal of Computer Applications (0975 8887)
Volume 118 - No. 9, May 2015

Automatic Speech Recognition of Urdu Digits with Optimal Classification Approach

Hazrat Ali
School of Computer
and Communication Engineering
University of Science
and Technology Beijing
Beijing 100083, China

An Jianwei
School of Computer
and Communication Engineering
University of Science
and Technology Beijing
Beijing 100083, China

Khalid Iqbal
Department of Computer Science
COMSATS Institute of
Information Technology
Attock, Attock, Pakistan

ABSTRACT

Speech Recognition for Urdu language is an interesting and less developed task. This is primarily due to the fact that linguistic resources such as rich corpus are not available for Urdu. Yet, few attempts have been made for developing Urdu speech recognition frameworks using the traditional approaches such as Hidden Markov Models and Neural Networks. In this work, we investigate the use of three classification methods for Urdu speech recognition task. We extract the Mel Frequency Cepstral Coefficients, the delta and delta-delta features from the speech data and train the classifiers to perform Urdu speech recognition. We present the performance achieved by training a Support Vector Machine (SVM) classifier, a random forest (RF) classifier and a linear discriminant analysis classifier (LDA) for comparison with SVM. Consequently, the experimental results show that SVM gives better performance than RF and LDA classifiers on this particular task.

General Terms:

Automatic Speech Recognition

Keywords:

Linear Discriminant Analysis, Mel-Frequency Cepstral Coefficients, Random Forest, Support Vector Machines, Urdu

1. INTRODUCTION

Speech processing is an interesting area of research. One of the useful application of speech processing is automatic speech recognition (ASR) which is an effective tool for human-machine interaction. An ASR system enables a person to talk to a machine. Such applications are used in voice operated activities, automated credit card activation, security and surveillance facilities. More recent integration of ASR systems have enabled users talking to *OK Google* and *Apple Siri* using smart phones, tablets and phablets. Similarly, isolated words recognition or spoken digits recognition are widely popular in data entry automation, PIN code operated devices/applications/services, banking automation and voice dialing applications. The field of ASR has observed tremendous developments during the past few decades (for developed languages, in particular). These include ASR systems for English, French, Mandarin and Japanese [1–5]. For languages which are not rich in terms of available computational linguistics resources, the research is still less mature. This holds true for Urdu language as well, having insufficient progress for ASR development even though it is one of the largest languages of the world with approximately 60-70 million speakers around the world [1]. It is the national language of Pakistan, widely spoken and understood in India and has several hundred-thousand speakers in Bengladesh, Nepal, United Kingdom, South Africa, and United States [2]. It also shares significant vocabulary with Arabic and Persian, for example, *dunya (meaning world)* and *kitab (meaning book)*.

In this paper, we present our work on speech recognition of Urdu digits using Support Vector Machines (SVMs), Random Forests (RF) and Linear Discriminant Analysis (LDA) classifiers. To the best of our knowledge, these approaches (excluding LDA) have not been reported for Urdu ASR before. We expect that this work will serve as a baseline for future research work on Urdu ASR as it is based on a more diverse Urdu speech corpus, unlike previous attempts which were based on mono-speaker database only (we outline this in Section 2). The remaining of this paper is organized as follows; In section 2, we summarize different existing approaches and results as available for Urdu ASR. In section 3, we provide an explanation of the features namely Mel-Frequency Cepstral Coefficients and the classification algorithms namely SVMs, RF and LDA. We then explain our experiment and present the results in Section 4. Finally, we conclude our paper in section 5 and outline future directions as well.

2. RELATED WORK

Ghai et al [6] have briefly presented several developments on Urdu ASR in their work on ASR of Indo-Aryan languages such as Hindi, Punjabi, Bengali and Gujarati. Akram et al [7] have presented an ASR system for continuous Urdu speech. However it lacks identification of a standard corpus used in the experimentation. It also misses information on the size of training and test sets and the mechanism for selecting the two sets. The reported accuracy is limited to 54%. In similar work on Urdu digits recognition, Ahad

[1] http://www.ethnologue.com/language/urd
[2] [Online resource] http://www.cle.org.pk/

et al [8] have reported the use of a multilayer perceptrons (MLP) for Urdu digits ASR from 0 to 9, however the work is limited to single speaker corpus and thus limited to speaker-specific applications only as they have not utilized a multiple speaker corpus. Hasnain & Awan [9] have also reported Urdu ASR for digits using a neural network model with two hidden layers for classification, however, their work is based on the use of fourier descriptors and correlations coefficients and not on MFCCs. Much higher accuracy rate has been reported, though, it is not clear whether the reported values are for training set only or have been obtained for an explicit test set. More sophisticated work on speech recognition use Hidden Markov Models (HMM) [10]. The HMM have also been used by [11] for Urdu, presenting a speaker-independent speech recognition system. The framework utilized by [11] is the open source Sphinx-4. They represent each word as a single phoneme which means that the system performance may suffer degradation for longer words. More recent work for continuous speech Urdu ASR is by Huda et al [12] using Sphinx open source toolkit. We are not making direct comparisons to Huda's work [12] as the data scale as well as application scale differ than our work. Besides, the speech data used in their experimentation is limited to one particular accent only. Ali et al [13] have reported results on the same dataset of Urdu, as used in the work reported in this paper, and have analyzed the use of Discrete Wavelet Transform (DWT) features and compared the performance with MFCCs. They conclude their discussion in favor of using MFCCs. But the classifier is essentially the same to the one used in their previous work i.e. Linear Discriminant Analysis as in [14–16]. Their work does not involve making investigation of performance of more robust classifiers. We, on the other hand, suggest the use of SVM and compare it with a random forest classifier too [17]. In this work, we limit our experimentation to Urdu digits recognition.

3. METHODOLOGY

3.1 Mel Frequency Cepstral Coefficients

The Mel Frequency Cepstral Coefficients (MFCC) are popular and widely used within the speech processing community. MFCCs are based on the mel scale (a logarithmic scale) and attempt to mimic the human ear response. The Mel scale is directly related to the mechanism behind the human ear response to frequencies in an audio signal (*this response is non-linear*). The Mel-scale possess a linear spacing for frequencies below 1000 Hz and a non-linear spacing for frequencies above 1000 Hz [18]. The relation between the Mel scale frequencies and the Hertz frequencies can be represented by Eq. 1;

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

Several methods for MFCC extraction have been proposed by [18–20]. For MFCC calculation, we use the Malcom's implementation [3]. We calculate 12 coefficients for each audio file and combine them with the delta and delta-delta coefficients. Delta and delta-delta coefficients are the time derivatives of the original coefficients and are helpful in retaining the dynamic information in the speech data. A basic step wise approach to calculate the MFCC and the delta, delta-delta MFCCs has been summarized in Algorithm 1.

[3] `https://engineering.purdue.edu/~malcolm/interval/1998-010/`

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

3.2 Support Vector Machines

Support Vector Machine (SVM) is a popular kernel based discriminative classification algorithm. SVMs were introduced first by Boser et al [21] and got their popularity for handwritten character recognition [22]. SVMs are regarded to be the large-margin boundary classifier as they attempt to separate the data with hyper planes by maximizing the distance from the data of both classes (for binary classification case). SVM is characterized by the underlying kernel function, say linear, polynomial and Gaussian. SVMs are usually easy to train and does not suffer the local optima problem. However, the selection of kernel function is critical. SVMs have been used for various machine learning task, *for example*, hand-written character recognition [23–25], object recognition [26], speaker recognition and language recognition [27].[4] SVM is a binary classification algorithm and is comprised of sums of kernel function $k(x_i, x_j)$.

$$f(x) = \sum_{i=1}^{N} \alpha_i t_i K(x_i, x_j) + d \quad (2)$$

where t_i represents the ideal outputs, $\sum_{i=1}^{N} \alpha t_i = 0$ and $\alpha_i > 0$. The ideal outputs are either $+1$ or -1 depending upon the class the data sample belongs to. The value of $f(x)$ compared against a threshold decides the output class of a particular test sample. For multi-class problems, the one-vs-all approach is adapted usually to achieve classification. For SVM, we use the libSVM library [29], which can handle the problem of multi-class data. We train the SVM with

[4] A nice tutorial on SVM is available from Burges [28].

the Gaussian RBF kernel, which for two data points x_i and x_j is given by Eq. 3;

$$K(x_i, x_j) = exp(\gamma(\|x_i - x_j\|)^2) \quad (3)$$

We select the optimal hyper-parameters γ and regularization constant C for the SVM after multiple iterations on the training and test data.

3.3 Random Forests

Decision Trees have been popular in computer vision applications where decision trees have been used as stand-alone entities. Random Decision Forest or Random Forest (RF) is the ensemble version of decision trees where the trees are randomly trained [17]. The use of RF was first reported by [30] for handwritten digit recognition. In an RF classifier, the trees are trained with randomly selected features and the average of the class posteriors from the trees is calculated for final prediction. RF classifiers are reported to be very successful in a variety of machine learning tasks [31]. In our work, we use an RF classifier on the task of speech recognition and our results for the RF classifier are based on 300 trees.

3.4 Linear Discriminant Analysis

The Linear Discriminant Analysis (LDA) [32] is both a classification and a dimensionality reduction technique. The LDA transforms the data to find a matrix $\mathbf{\Theta}$ and seeks to maximize the ratio between the inter-class variance and the intra-class variance, thus achieving greater separability. For classification, the Euclidean distance is calculated over each test example. Generally, for n classes, n Euclidean distances are calculated for each test example. The smallest distance then determines the class for the test example. The objective function for LDA transformation can be defined by $J(\Theta)$;

$$J(\Theta) = \frac{|\Theta^T \Psi \Theta|}{|\Theta^T W \Theta|} \quad (4)$$

where W is the average within-class variance, Ψ represents the total variance matrix, $|.|$ is the determinant value. We use LDA classification besides SVM and RF classifiers and compare the results on the digit recognition task.

4. EXPERIMENT

4.1 Dataset

This experiment is based on the dataset of Urdu digits recorded by ten speakers. The speakers include native/non-native, male/female speakers of different age. We normalize our data with zero-mean and variance of 1. We train our classifier with randomly selected training data. We choose a 7:3 ratio for training and test set distribution.

4.2 Confusion Matrix

We summarize the results of correct recognition of the test instances in a confusion matrix. For P number of words, we get a $P \times P$ matrix. The general representation has been shown by CM

$$CM = \begin{matrix} m_{11} & m_{12} & m_{13}... & m_{1P} \\ m_{21} & m_{22} & m_{23}... & m_{2P} \\ m_{31} & m_{32} & m_{33}... & m_{3P} \\ . & . & .... & . \\ . & . & .... & . \\ m_{P1} & m_{P2} & m_{P3}... & m_{PP} \end{matrix} \quad (5)$$

For this confusion matrix, CM, the diagonal entries show the correct word recognition i.e. m_{ij} for $i = j$. Similarly, the number of mis-judgments of a test word against a particular word are shown by the non-diagonal entries, i.e. m_{ij} for $i \neq j$. We report an overall test accuracy of 73% for the SVM classifier. We also report the results for a random forest classifier and LDA classifier, both resulting in an accuracy of 63%. The confusion matrix plots for SVM based classification, Random Forest based classification and LDA classification have been shown in Figure 1, Figure 2 and Figure 3 respectively. We show the recognition rate for each digit as achieved by SVM classifier, RF classifier and LDA classifier in Table 1. For this particular task, the accuracy achieved for LDA classifier is not different than the accuracy for RF classifier as both give an overall accuracy of 63%. However, by looking at the confusion matrix graph for the LDA classifier, it turns out that LDA classifier is the only classifier generating 0% accuracy for the word number 7 (See the empty 7th column in Figure 3).

5. CONCLUSION AND FUTURE WORK

In this paper, we have presented the use of SVM, RF and LDA classifiers for speech recognition of Urdu dataset. We trained the SVM classifier with MFCC features and observed a recognition rate of 73% on the test set. We also trained a random forest classifier and an LDA classifier with the same data and reported an overall accuracy of 63% for both. We observed better performance for the SVM classifier[5]. The experimentation are based on speech data of ten speakers including both native/non-native, male/female speakers of different age. We expect these results to serve as a baseline for future experimentation. We believe that these results will prompt further research on Urdu ASR. We suggest that the use of more robust classifier can help with achieving better performance. Similarly, exploring the use of SVM-HMM hybrid model is also an interesting dimension for future experimentation. Besides, an interesting and useful task would be to extend the work to continuous speech recognition of Urdu. We will also be exploring the use of deep learning models for Urdu ASR as they have been highly successful on large vocabulary speech recognition tasks.

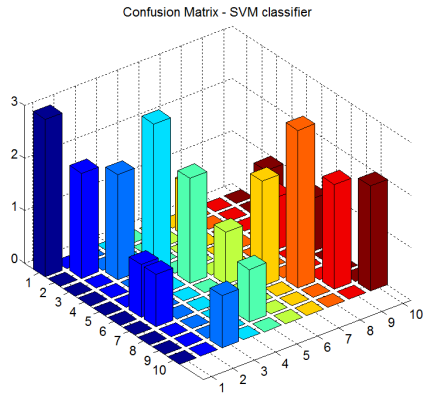

Fig. 1. A visualisation of confusion matrix with SVM classification

[5]This, of course, does not imply that one classifier is better than the other on any data in general, as the choice of classification algorithm vary from case to case.

Table 1. Recognition Rate for Digits

S. No	Word Number	Recognition Rate for SVM	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%

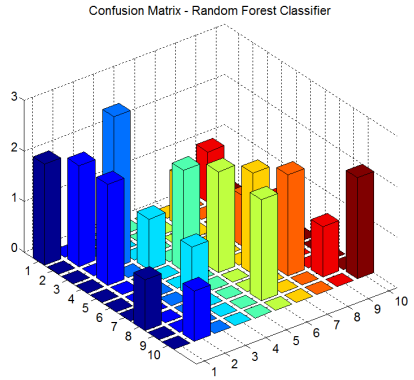

Fig. 2. A visualisation of confusion matrix with Random Forest classification

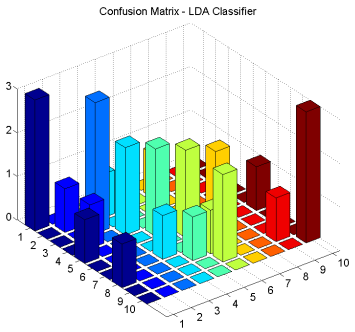

Fig. 3. A visualisation of confusion matrix with LDA classification

6. REFERENCES

[1] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, Feb 1978.

[2] L. Gagnon, S. Foucher, F. Laliberte, and G. Boulianne, "A simplified audiovisual fusion model with application to large-vocabulary recognition of french canadian speech," *Canadian Journal of Electrical and Computer Engineering*, vol. 33, no. 2, pp. 109–119, 2008.

[3] T. Shimizu, Y. Ashikari, E. Sumita, J. Zhang, and S. Nakamura, "NICT/ATR Chinese-Japanese-English Speech-to-Speech Translation System," *Tsinghua Science and Technology*, vol. 13, no. 4, pp. 540 – 544, 2008.

[4] J.-J. Mao, Q. lin Chen, F. Gao, R. Guo, and R.-Z. Lu, "STIS: a Chinese spoken dialogue system about Shanghai transportation information," in *Proceedings. 2003 IEEE Intelligent Transportation Systems*, vol. 1, Oct 2003, pp. 65–68.

[5] K. Ohtsuki, T. Matsuoka, T. Mori, K. Yoshida, Y. Taguchi, S. Furui, and K. Shirai, "Japanese large-vocabulary continuous-speech recognition using a newspaper corpus and broadcast news," *Speech Communication*, vol. 28, no. 2, pp. 155 – 166, 1999.

[6] W. Ghai and N. Singh, "Analysis of Automatic Speech Recognition Systems for Indo-Aryan Languages : Punjabi A Case Study," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 2, no. 1, pp. 379–385, 2012.

[7] M. U. Akram and M. Arif, "Design of an Urdu Speech Recognizer based upon acoustic phonetic modeling approach," in *Proceedings of 8th International Multitopic Conference (INMIC) 2004,*, Dec 2004, pp. 91–96.

[8] A. Ahad, A. Fayyaz, and T. Mehmood, "Speech recognition using multilayer perceptron," in *Proceedings. IEEE Students Conference, ISCON '02.*, vol. 1, Aug 2002, pp. 103–109.

[9] S. Hasnain and M. Awan, "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.

[10] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb 1989.

[11] J. Ashraf, N. Iqbal, N. Sarfraz Khattak, and A. Mohsin Zaidi, "Speaker independent urdu speech recognition using hmm," in *The 7th International Conference on Informatics and Systems (INFOS 2010)*, March 2010, pp. 1–5.

[12] H. Sarfraz, S. Hussain, R. Bokhari, A. A. Raza, I. Ullah, Z. Sarfraz, S. Pervez, A. Mustafa, I. Javed, and R. Parveen, "Large vocabulary continuous speech recognition for urdu," in *Proceedings of the 8th International Conference on Fron-*

tiers of Information Technology. New York, USA: ACM, 2010, pp. 1:1–1:5.

[13] H. Ali, N. Ahmad, X. Zhou, K. Iqbal, and S. M. Ali, "DWT features performance analysis for automatic speech recognition of Urdu," *SpringerPlus*, vol. 3, no. 1, p. 204, 2014.

[14] H. Ali, N. Ahmad, X. Zhou, M. Ali, and A. Manjotho, "Linear discriminant analysis based approach for automatic speech recognition of urdu isolated words," in *Communication Technologies, Information Security and Sustainable Development*, ser. Communications in Computer and Information Science, F. K. Shaikh, B. S. Chowdhry, S. Zeadally, D. M. A. Hussain, A. A. Memon, and M. A. Uqaili, Eds. Springer International Publishing, 2014, vol. 414, pp. 24–34.

[15] H. Ali, N. Ahmad, and X. Zhou, "Automatic speech recognition of Urdu words using linear discriminant analysis," *Journal of Intelligent and Fuzzy Systems*, 2015, pre-print.

[16] H. Ali, A. d'Avila Garcez, S. Tran, X. Zhou, and K. Iqbal, "Unimodal late fusion for NIST i-vector challenge on speaker detection," *Electronics Letters*, vol. 50, no. 15, pp. 1098–1100, July 2014.

[17] A. Criminisi and J. Shotton, "Classification forests," in *Decision Forests for Computer Vision and Medical Image Analysis*, ser. Advances in Computer Vision and Pattern Recognition, A. Criminisi and J. Shotton, Eds. London: Springer, 2013, pp. 25–45.

[18] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 1, 2001, pp. 73–76.

[19] W. Han, C.-F. Chan, C.-S. Choy, and K.-P. Pun, "An efficient mfcc extraction method in speech recognition," in *Proceedings. IEEE International Symposium on Circuits and Systems, ISCAS 2006.*, May 2006.

[20] B. Kotnik, D. Vlaj, and B. Horvat, "Efficient noise robust feature extraction algorithms for distributed speech recognition (dsr) systems," *International Journal of Speech Technology*, vol. 6, no. 3, pp. 205–219, 2003.

[21] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, ser. COLT '92. New York, NY, USA: ACM, 1992, pp. 144–152.

[22] L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Muller, E. Sackinger, P. Simard, and V. Vapnik, "Comparison of classifier methods: a case study in handwritten digit recognition," in *Proceedings of the 12th IAPR International. Conference on Pattern Recognition,*, vol. 2, Oct 1994, pp. 77–82.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[24] B. Schölkopf, C. Burges, and V. Vapnik, "Extracting support data for a given task," in *First International Conference on Knowledge Discovery & Data Mining, Menlo Park*. AAAI Press, 1995, pp. 252–257.

[25] ——, "Incorporating invariances in support vector learning machines," in *Proceedings of the 1996 International Conference on Artificial Neural Networks*, ser. ICANN 96. Verlag: Springer, 1996, pp. 47–52.

[26] V. Blanz, B. Schölkopf, H. H. Bülthoff, C. Burges, V. Vapnik, and T. Vetter, "Comparison of view-based object recognition algorithms using realistic 3d models," in *Proceedings of the 1996 International Conference on Artificial Neural Networks*, ser. ICANN 96. Verlag: Springer, 1996, pp. 251–256.

[27] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, pp. 210–229, 2006.

[28] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[29] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[30] T. K. Ho, "Random decision forests," in *Proceedings of the Third International Conference on Document Analysis and Recognition*, vol. 1, Aug 1995, pp. 278–282 vol.1.

[31] R. Caruana, N. Karampatziakis, and A. Yessenalina, "An empirical evaluation of supervised learning in high dimensions," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: ACM, 2008, pp. 96–103.

[32] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis; a brief tutorial," http://www.music.mcgill.ca/~ich/classes/mumt611_07/classifiers/lda_theory.pdf, [Online] Accessed: 10 November, 2014.