Journal of Intelligent & Fuzzy Systems 28 (2015) 2369–2375
DOI:10.3233/IFS-151554
IOS Press

Automatic speech recognition of Urdu words using linear discriminant analysis

Hazrat Ali[a,*], Nasir Ahmad[b] and Xianwei Zhou[a]
[a]*School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, P.R. China*
[b]*Department of Computer Systems Engineering, University of Engineering and Technology Peshawar, Peshawar, Pakistan*

Abstract. Urdu is amongst the five largest languages of the world and possess a very important role as it shares its vocabulary with languages as Arabic, Persian, Hindi and several other languages of the Indo-Pak. The Automatic Speech Recognition task of Urdu has not been addressed significantly. This paper presents the statistical based classification technique to achieve the task of Automatic Speech Recognition of isolated words in Urdu. The proposed approach is based on calculation of 52 Mel Frequency Cepstral Coefficients for each isolated word. The classification has been achieved with Linear Discriminant Analysis. The successful or incorrect matches have been presented in the Confusion Matrix. As a prototype, the framework has been trained with audio samples of seven speakers including male/female, native/non-native and speakers with different ages. The test set comprises of audio data of three speaker. For each isolated, percentage error has been calculated. It was found that majority of the words are recognized with percentage error less than 33%. Some words suffer 100% error and were referred to be the bad words. This work may provide a baseline for further research on Urdu Automatic Speech Recognition.

Keywords: Urdu automatic speech recognition, mel frequency cepstral coefficients, linear discriminant analysis

1. Introduction

User friendly and natural interaction between man and machine has always been a complementary part of technological development. Speech is the most effective medium of communication between human and same is envisaged to be applicable for human-machine interaction. Therefore, Automatic Speech Recognition (ASR) has significantly attracted researchers for the last five decades and has attained considerable success in noise-free environments. Successful ASR enables the computers to exhibit human-like behavior by understanding the voice input to them. Such hearing systems have been developed in various languages such as English, French, Japanese, Chinese and Arabic [1–5], and have wide-spread application ranging from data entry to security and surveillance. The research on ASR has enabled the communities with lower level of literacy to interact with machines, and similarly facilitated the interaction of blind and disabled people with the computers [6].

Despite the development of ASR systems in these languages, there has been no significant contribution to ASR of Urdu language, which is one of the largest languages of the world, with approximately 70 million across the globe[1]. Wiqas [7] has summarized the research work conducted on the ASR of the languages of the Indo-Pak, including the research work on Urdu ASR. A continuous speech ASR system for

*Corresponding author. Hazrat Ali, School of Computer and Communication Engineering, University of Science and Technology Beijing, 10083, Beijing, P.R. China. Tel: +8613261071049; E-mail: engr.hazratali@yahoo.com.

[1] [Online] http://www.ethnologue.com/language/urd.

Urdu language has been presented in [8], however, no information about the corpus has been provided. The recognition rate for the continuous speech ASR is reported to be 54%. Furthermore, it lacks the information about the use of number of words/sentences and the training/test data. Azam [9] has proposed an Artificial Neural Networks (ANN) based Urdu speech recognition system however; this work is limited to digits recognition only. Moreover, the application of the system is limited to single speaker only. Ahad et al. [10] has used a different class of ANN called multilayer perceptrons (MLP) however; they have achieved recognition of Urdu digits from 0 to 9 for mono-speaker database only. Hasnain et al. [11] has made yet another effort to achieve the task of digits recognition for 0 to 9, based on the use of feed-forward neural network models developed in Matlab. A more recent contribution to isolated words recognition has been made by [12], developing a Hidden Markov Model (HMM) [13] based speaker-independent speech recognition system for Urdu. In this work the open source framework Sphinx-4 has been used for the classification. A "wordlist" grammar language model was adopted where each word was represented as a single phoneme instead of dividing into sub-units. An apparent limitation of this approach is that this may be applicable to shorter words but for longer words, the performance may degrade drastically. Huda [14] has used a relatively larger data set for the training purpose, however, the system developed is for continuous speech recognition task and thus different than the work reported in our paper. Besides, the recognition results are yet modest and are limited to one particular accent only.

Research on ASR can be targeted at small, medium or large vocabulary applications; it may be for digits only, isolated words only or continuous speech applications. The applications of isolated words recognition are well known including the automated banking applications, automatic data and PIN codes entry applications, e-health monitoring and voice dialing phone applications etc. In this paper the ASR task for medium vocabulary isolated words has been undertaken containing 100 isolated words of Urdu.

The three important components of an ASR system are the corpus i.e. the database of speech data, the features extraction and the classification. In Section 2 of this paper, the corpus used for this work has been discussed briefly. The features extraction approach and the major steps involved in the extraction of these features have been presented in Section 3. The classification of the different words based upon the features obtained for each word, has been discussed in Section 4. Finally, the results have been summarized in Section 5.

2. Corpus selection

The use of a standard corpus forms the most important component of an ASR framework. A standard corpus should cover a range of acoustic variations and different aspects of a language. These include session and speaker variations. In this work, the corpus developed by Ali et al. [15], has been used. This corpus contains 250 isolated words selected from the list of most frequently used words, developed by the Center for Language Engineering [16]. Audio files for one hundred isolated words have been selected from the corpus and used in the training and testing of the system. The one hundred words used contain the digits from 0 to 9, names of seasons, days of the week and the names of months. Few of the words are also accompanied by their corresponding antonyms. The words are available in separate audio files with an average length of 500 milliseconds and stored in mono format with. *wav* extension. Based upon the attributes such as age, gender and origin, this corpus provides a balanced distribution. The files include the words uttered by both male and female speakers of different ages. Similarly, a variety of accents has been covered by including the audio recordings by both native and non-native speakers

Table 1a
Sample of representation of the speech data (as in [15])

Speaker identification	Age group	Gender	Native non-native
AAMNG1	G1	Male	Non-Native
ABMNG1	G1	Male	Non-Native
ACMNG2	G2	Male	Non-Native
AEFYG1	G1	Female	Native
AFFYG1	G1	Female	Native
AGMNG1	G1	Male	Non-Native
AHMNG1	G1	Male	Non-Native

Note: G1 represents age group of 20–25 years, G2 represents age group of 26–30 years.

Table 1b
Sample of representation of the speech data for test set

Speaker identification	Age group	Gender	Native non-native
AIMYG2	G2	Male	Native
AJMNG2	G2	Male	Non-Native
AKFNG1	G1	Female	Non-Native

originating from different areas. For example, Pashto speakers who origin from different locations in Pakistan have variations in their pronunciations of the same Urdu words. Thus, data from these speakers provide a variety of samples for training and testing purpose. A sample representation of the attributes of the speakers has been shown in Table 1.

3. Features selection

Feature Extraction is one of the most important modules of an Automatic Speech Recognition System. For continuous speech recognition, the feature extraction is typically aimed to capture the distinguishing characteristics of the phonemes i.e. the smallest unit of sound. However, for isolated words recognition, each word is usually split into equal number of segments and features are extracted from each of the segments. In this work, each word is split into four segments and the Mel Frequency Cepstral Coefficients (MFCC) based features have been obtained for each segment. This also ensures that the feature vectors for both the training and test sets data have same dimensions.

3.1. Mel frequency cepstral coefficients

The MFCC features are the most commonly used features for ASR applications. The MFCCs closely resembles the mechanism of human hearing. The Mel scale is based on the fact that the frequency response of the human's ear to the audio signal is not a linear function of frequency. This response can be best modeled on a Mel scale where the spacing between frequencies above 1000 Hz is logarithmic [17]. The relation between the Mel scale frequencies and the Hertz frequencies can be represented by the following equation:

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

The Mel Frequency Cepstrum is the power spectrum of a speech signal for short term and is based upon a linear cosine transform of a log power spectrum on the Mel scale. The Mel Frequency Cepstrum comprises of the MFC coefficients. Several methods for MFCC extraction have been proposed by [17–19]. The major steps in the extraction of MFCC are shown in Fig. 1.

In the pre-processing step, the segmentation of the words and noise removal have been achieved by using Adobe Audition Software. The sampling rate was set to 16000 Hz and the audio samples were saved as. wav files in mono format before being input to the algorithm. The Adobe Audition software has also been utilized for amplification or attenuation of the audio signal, as necessary, to obtain a uniform *db* level for all the samples. The pre-processing stage also includes the Pre-emphasis of the signal to increase the energy of the higher frequency contents. The pre-emphasis is achieved using filter of the form, $H(z)$:

$$H(z) = 1 - 0.97z^{-1} \tag{2}$$

The pre-processing is followed by the windowing of the speech signal. A rectangular window as defined by equation for $w(n)$ in Equation (3) has been used. For speech processing applications, hamming window is more commonly used to avoid information loss, however for isolated words processing, rectangular window is equally beneficial.

$$w(n) = \begin{cases} 1 & 0 \le n \le M-1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

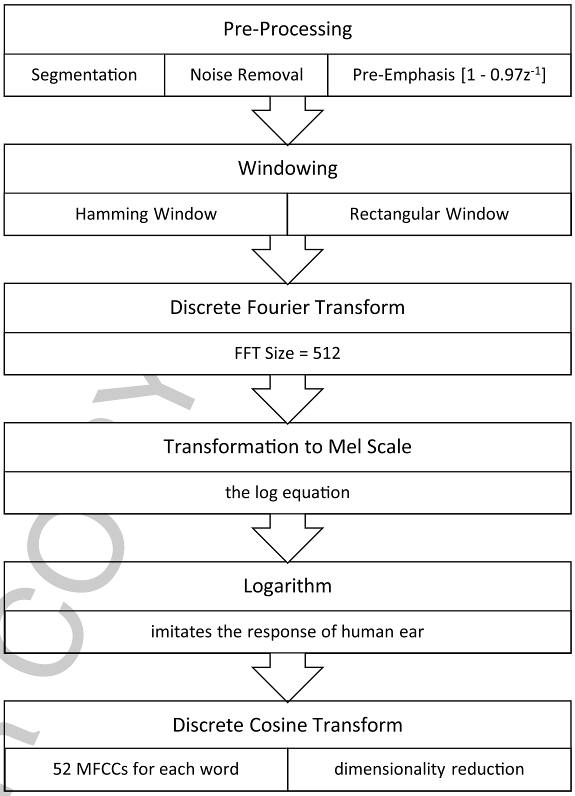

Fig. 1. Extraction of mel frequency cepstral coefficients.

where $M = 128$. Fast Fourier Transform [20, 21] is applied to the windowed frame of the signal. The size of FFT is $N = 512$. The spectrum, thus obtained, is transformed to the Mel scale, as defined by the equation for f_{mel}. To imitate the logarithmic response of human ear, the output of the mel scale filters bank is subjected to *base 10 Logarithmic* function. Finally, the application of Discrete Cosine Transform (DCT) [22] generates the MFCCs, i.e. 52 MFCCs for each isolated word.

4. Classification

The recognition on the basis of MFCCs requires a supervised classification technique for which Linear Discriminant Analysis is a strong candidate [23, 24]. The classification includes; 1) Training of the system and 2) Testing of the system. 70% percent of the data has been used for training the ASR system and the remaining 30% data has been used for testing of the system.

4.1. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a classification as well as dimensionality reduction technique. LDA can be class-dependent or class-independent, based upon maximization of the ratio of between class variance to within class variance or maximization of the ratio of overall variance to within class variance, respectively [23–26].

4.2. Training and testing data

To evaluate the performance of the ASR system, the MFCCs of a total of hundred words have been used for training and testing of the system. As a simple case, the training and testing has been done with the speech data of first ten speakers. The training set contains data from both native and non-native speakers of Urdu. Similarly, it also contains male as well as female speakers, as shown in Table 1.

4.3. Confusion matrix

The number of correct matches from the testing data with the training data has been summarized in a Confusion Matrix. The confusion matrix is of size $N \times N$ for N number of words. It can be represented as shown by $\boldsymbol{M_c}$.

$$M_c = \begin{matrix} \boldsymbol{m_{11}} & \boldsymbol{m_{12}} & \boldsymbol{m_{13}} & \ldots & \boldsymbol{m_{1N}} \\ \boldsymbol{m_{21}} & \boldsymbol{m_{22}} & \boldsymbol{m_{23}} & \ldots & \boldsymbol{m_{2N}} \\ \boldsymbol{m_{31}} & \boldsymbol{m_{32}} & \boldsymbol{m_{33}} & \ldots & \boldsymbol{m_{3N}} \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ \boldsymbol{m_{N1}} & \boldsymbol{m_{N2}} & \boldsymbol{m_{N3}} & \ldots & \boldsymbol{m_{NN}} \end{matrix} \tag{4}$$

The number of correct matches for a word i has been shown by the diagonal entries of the confusion matrix, i.e. $\boldsymbol{m_{ij}}$ for $i = j$. Number of confusions of word i with word j has been shown by non-diagonal entries, i.e. $\boldsymbol{m_{ij}}$ for $\boldsymbol{i \neq j}$.

5. Results

The error in the recognition of any isolated word is calculated from the confusion matrix. For an isolated word i, the diagonal entry $\boldsymbol{m_{ii}}$, divided by the sum of all the entries in row $\boldsymbol{i}$, gives the fraction of test data correctly matched. The sum of all the entries in a row is always equal to the number of test signals. This ratio can be defined mathematically as;

$$\boldsymbol{Correct\ Match,\ C} \overset{\text{def}}{=} \frac{\boldsymbol{m_{ij}}}{\boldsymbol{m_{i1} + m_{i2} + \ldots m_{iN'}}} \qquad \boldsymbol{for\ i = j,\ \ j = 1,\ 2,\ 3 \ldots N} \tag{5}$$

Thus, the error is measured by using the following equation;

$$\boldsymbol{\%\ error = (1 - C) \times 100} \tag{6}$$

5.1. Results for first ten words

Figure 2 shows the confusion matrix graph for the first ten words. The x-axis and y-axis represent the indexes for the words i.e. 001 to 010. The number of successful or incorrect matches is represented by the height of the bars. As already mentioned, the maximum possible height is 3 as the number of test signals used here is 3. The percentage error and number of fraction of test signals correctly recognized has been summarized for the first ten words in Table 2. As shown in this table, the first word gives 66% correct match, also depicted by the confusion matrix graph, by the first bar having a height of 2. The test signals for word 004 has undergone a 0% error and the bar for this word has a height of 3. Similarly, the results for other words are

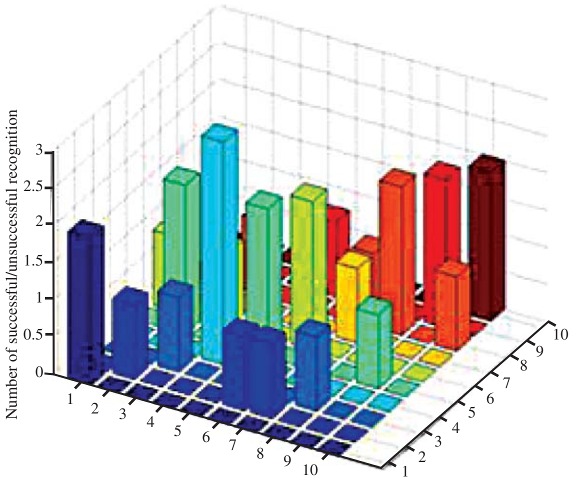

Fig. 2. Confusion matrix graph for first ten words. Note: The horizontal axes can be read as values from 1 through 10. The vertical axes shows the number of successful/unsuccessful recognition.

Table 2
Percentage error for words 001 to 010

S. No	Word number	Value of C	% Error
1	001	0.6667	33.33%
2	002	0.3333	66.67%
3	003	0.3333	66.67%
4	004	1.0	0%
5	005	0.6667	33.33%
6	006	0.6667	33.33%
7	007	0.3333	66.67%
8	008	0.6667	33.33%
9	009	0.6667	33.33%
10	010	0.6667	33.33%

obvious from the confusion matrix graph and the corresponding table.

5.2. *Results for words 031 to 040*

As a second sample of the result, confusion matrix graph for word 031 to 040 has been shown in Fig. 3. The corresponding fractional values for correct matches and percentage error have been summarized in Table 3. The results shown in Table 3 are very important and needs to be discussed. As shown in Table 3, it is obvious that there is a zero percent error for words 032 through word 034. On the other hand, a complete mismatch exists for word 031, resulting in a 100% error.

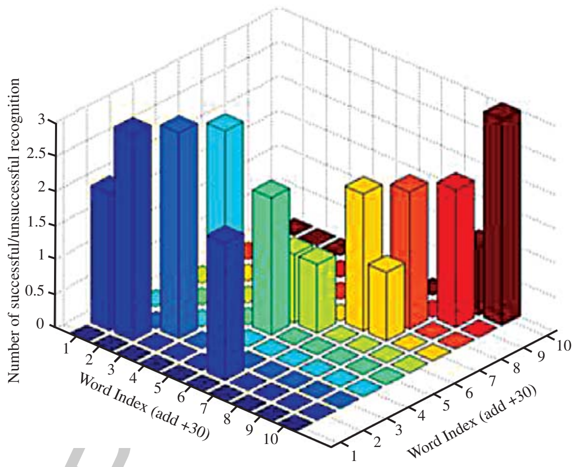

Fig. 3. Confusion matrix graph for words 031 to 040. Note: The horizontal axes can be read as values from 31 through 40. The vertical axes shows the number of successful/unsuccessful recognition.

Table 3
Percentage error for words 031 to 040

S. No	Word number	Value of C	% Error
1	031	0	100%
2	032	1.0	0%
3	033	1.0	0%
4	034	1.0	0%
5	035	0.6667	33.33%
6	036	0.3333	66.67%
7	037	0.6667	33.33%
8	038	0.6667	33.33%
9	039	0.6667	33.33%
10	040	1.0	0%

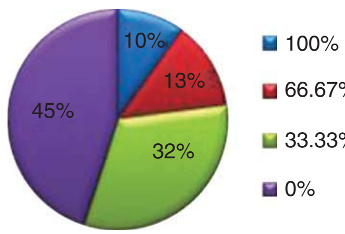

Fig. 4. Percentage of Test Data having different percentage error.

5.3. *Overall percentage error*

Figure 4 shows the proportion of the words with 100%, 66.67%, 33.33% and 0% error, respectively. The analysis shows that the percentage error is either zero or 33.33% for majority of the words. However, for few of the words, the value is larger approaching the maximum possible value i.e. 100%. The overall error, $\boldsymbol{E}$, can be measured as;

$$\frac{100\%\ of\ (10\times 3)+66.67\%\ of\ (13\times 3)+33.33\%\ of\ (32\times 3)+0\%\ of\ (45\times 3)}{(100\times 3)}$$

From this calculation, $E = 29.33\%$. These results have been previously reported in [27]. This is comparable with so many existing ASR systems as developed for other languages with audio-only based features. This value, however, can be reduced further by increasing the amount of training data.

5.4. Bad words

The words having a 100% error rate are referred to be the Bad Words. The primary reason for such a poor performance of the ASR system for these words, is the poor quality of recording which was determined through manual analysis of the audio files. Besides this, as each word has been divided into four segment, there is a possibility that more than one segment are matching exactly with segments of other words and the ASR framework is confused.

6. Future work

This ASR system has been developed for speech recognition of isolated words only. This is a medium vocabulary application limited to a hundred words and can be extended to several thousand words. However, in that case, an even larger data for the training of the system will be required. Thus, there is need to increase the corpus size. This paper can serve as a baseline for future research on ASR of Urdu language and can be extended to Continuous Speech Recognition of Urdu. This is an audio-only based feature extraction for ASR. The system can be evaluated by using audio-visual features which should result in the enhancement of the performance. Furthermore, a more promising development recently made in ASR is based on deep learning models. These deep learning models can be explored and investigated further for Urdu ASR.

Acknowledgments

The authors are thankful to the supporting staff of the Department of Electrical Engineering, University of Engineering and Technology, Peshawar, Pakistan. The authors also extend their gratitude to Salman Ilahi and Irfan Ahmad for their valuable input and suggestions throughout this work. The authors are highly grateful to the anonymous reviewers for giving useful comments on this work.

References

[1] H. Sakoe and S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, *IEEE Transactions on Acoustic, Speech and Signal Processing* **26**(1) (1978), 43–49.

[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* **33**(2) (2008), 109–119, Spring.

[3] S. Morii, K. Niyada, S. Fujii and M. Hoshimi, Large vocabulary speaker-independent Japanese speech recognition system, *in IEEE International Conference on Acoustics, Speech and Signal Processing*, 1985, pp. 866–869.

[4] T. Shimizu, Y. Ashikari, E. Sumita and J. Zhang, NICT/ATR Chinese-Japanese-English speech-to-speech translation system, *Tshingua Science and Technology* **13**(4) (2008), 540–544.

[5] M. Jiaju, C. Qiulin, G. Feng, G. Rong and L. Ruzhan, SHTQS: A telephone-based Chinese spoken dialogue system, *Journal of Systems Engineering and Electronics* **16**(4) (2005), 881–885.

[6] S. Khadivi and S. Ney, Integration of speech recognition and machine translation in computer-assisted translation, *IEEE Transactions on Audio, Speech and Language Processing* **16**(8) (2008), 1551–1564.

[7] 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)* **2**(1) (2012), 379–385.

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

[9] S.M. Azam, Z.A. Mansoor, M. Shahzad Mughal and S. Mohsin, Urdu Spoken Digits Recognition Using Classified MFCC and Backpropgation Neural Network, *in Computer Graphics, Imaging and Visualization, CGIV'07*, 2007, pp. 414–418.

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

[11] S.K. Hasnain and M.S. Awan, Recognizing spoken Urdu numbers using fourier descriptor and neural networks with Matlab, *in Second International Conference on Electrical Engineering, (ICEE 2008)*, 2008, pp. 1–6.

[12] J. Ashraf, N. Iqbal, N.S. Khattak and A.M. Zaidi, Speaker Independent Urdu speech recognition using HMM, *in The 7th International Conference on Informatics and Systems (INFOS 2010)*, 2010, pp. 1–5.

[13] L.R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, *Proceedings of the IEEE* **77**(2) (1989), 257–286.

[14] H. Sarfraz, et al., Large Vocabulary Continuous Speech Recognition for Urdu, *in 8th International Conference on Frontiers of Information Technology, (FIT'10)*, 2010.

[15] H. Ali, N. Ahmad, K.M. Yahya and O. Farooq, A Medium Vocabulary Urdu Isolated Words Balanced Corpus for Automatic Speech Recognition, *in 2012 International Conference on Electronics Computer Technology (ICECT 2012)*, 2012, pp. 473–476.

[16] Center for Language Engineering. (2012) [Online]. http://www.cle.org.pk/

[17] S. Molau, M. Ptiz, 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)*, 2001, pp. 73–76.

[18] W. Han, C.-F. Chan, C.-S. Choy and K.-P. Pun, An efficient MFCC extraction method in speech recognition, *in IEEE International Symposium on Circuits and Systems, (ISCAS 2006)*, 2006.

[19] B. Kotnik, D. Vlaj and B. Horvat, Efficient noise robust feature extraction algorithms for distributed speech recognition (DSR) systems, *International Journal of Speech Technology* **6**(3) (2003), 205–219.

[20] J.G. Proakis and D.G. Manolakis, *Digital Signal Processing*, Principles, Algorithms & Applications, 4th ed., Pearson Education, Inc, 2007.

[21] V.K. Ingle and J.G. Proakis, Digital Signal Processing Using Matlab, 3rd ed. Standford, USA: Cengage Learning, 2010.

[22] D. Salomon, *Data Compression*, The Complete Reference, 4th ed. London, United Kingdom, Springer, 2007.

[23] S. Balakrishnama, A. Ganapathiraju and J. Picone, Linear discriminant analysis for signal processing problems, *in Proceedings of the IEEE Southeastcon*, 1999, pp. 36–39.

[24] S. Balakrishnama and A. Ganapathiraju. (Accessed: 2012, March) Linear Discriminant Analysis; A Brief Tutorial. [Online]. http://www.music.mcgill.ca/∼ich

[25] N. Jakovljevic, D. Miskovic, M. Janev, M. Secujski and V. Delic, Comparison of linear discriminant analysis approaches in automatic speech recognition, *Elektronika Ir Elektrotechnika* **19**(7) (2013), 76–79.

[26] M. Katz, H.-G. Meier, H. Dolfing and D. Klakow, Robustness of Linear Discriminant Analysis in Automatic Speech Recognition, *Proceedings of 16th International Conference on Pattern Recognition, (ICPR 2002)*, 2002, pp. 30371.

[27] H. Ali, N. Ahmad, X. Zhou, M. Ali and A. Asghar, Linear Discriminant Analysis Based Approach for Automatic Speech Recognition of Urdu Isolated Words, in Communication Technologies, Information Security and Sustainable Development, in Springer CCIS series, vol. 414, 2014, pp. 24–34.