Automatic Speech Recognition - A Literature Survey on Indian languages and Ground Work for Isolated Kannada Digit Recognition using MFCC and ANN

S B Harisha 1, S Amarappa 2, Dr. S V Sathyanarayana 3

1,2 Department of Telecommunication Engineering
Jawaharlal Nehru National College of Engineering, Shimoga - 577204
sbhjnnce@yahoo.co.in
2amarappas@yahoo.com
3 Department of Electronics and Communication Engineering
Jawaharlal Nehru National College of Engineering, Shimoga - 577 204
Email: vvs.ice@gmail.com

Abstract - In this research paper, we present an overview of Automatic Speech Recognition (ASR) in Indian Languages. Research in ASR on Indian languages started way back in 1990's. In this literature review a detailed survey of ASR in Indian languages is projected. Also we developed and implemented an efficient algorithm for “Isolated Kannada Digit Recognition” using Mel Frequency Cepstral Coefficients (MFCC) as feature vector and Artificial Neural Network (ANN) as classifier. In Neural network, feed forward with back propagation algorithm is used. The system is developed by considering the performance of a voice controlled machine in Kannada language. System Performance is evaluated for isolated Kannada digits with satisfactory results.

Keywords: Speech Recognition, Kannada Language, Digit recognition, Artificial Neural Network (ANN), Mel Frequency Cepstral Coefficients (MFCC).

I. INTRODUCTION

Automatic speech Recognition (ASR) is a process by which, a computer takes a speech signal as its input and converts it into text [1]. ASR is also known as “Speech Recognition”, “Computer Speech Recognition”, “Speech to Text”, or “STT”. Speech recognition can also be defined as the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of textual words. Keyboard, although a popular medium is not very convenient, as it requires a certain amount of skill for effective usage. A mouse on the other hand requires a good hand eye co-ordination. Physically challenged peoples find computer, difficult to use. Partially blind people find reading from a monitor difficult. To eliminate all these constraints speech interface gives the best solution by trapping human voice in a digital computer and decoding it into corresponding text [2].Although speech is the most natural and effective method of communication between human beings, it is not easy to quickly review, retrieve, and reuse speech documents if recorded as audio signals. Therefore, transcribing speech is expected to become a crucial capability for the coming IT era [3].A speech interface would support many valuable applications like telephone directory assistance, spoken database querying for novice users, “hands busy” applications in medicine or fieldwork, office dictation devices and for controlling electronic devices [4]. Kannada is a language spoken in India predominantly in the state of Karnataka, making it the 27th most spoken language in the world. It is the official and administrative language of state Karnataka. Developing ASR for Kannada is interesting and challenging. Basically there are three approaches for speech recognition.

i) Acoustic Phonetic Approach
ii) Pattern Recognition Approach
iii) Artificial Intelligence Approach

The rest of the paper is organized as follows. Literature survey of ASR on Indian languages is presented in section II, motivation for Kannada ASR is expressed in section III, proposed Isolated Kannada Digit Recognition
using MFCC as feature vector and ANN is explained in section IV. Experimental results and discussions are presented in section V. Concluding remarks are given in section VI.

II. EXISTING WORK (ASR – OVERVIEW OF INDIAN LANGUAGES)

According to Census 2011, India has 122 major languages and 2371 dialects (a particular form of a language which is peculiar to a specific region or social group). Linguistic diversity is rich in India, Out of 122 languages 22 are given the grade of National languages. Kashmiri(5.5 Million speakers as per census 2001), Hindi (258-422), Punjabi(29), Bengali (83), Nepali(2.9), Assamese (13), Manipuri(1.5), Oriya(32), Marathi(72), Gujarati(46), Konkani(2.5), Kannada(38), Telugu(74), Malayalam(33), Tamil(61), Urdu(52), Sanskrit(0.001), Sindhi(2.5), Santali(6.5), Maithili(12.2), Bodo (1.4) and Dogri (2.3).In India, some languages have many scripts and some language have single script. More interestingly, accent is not uniform within the same language speaking society hence it is a major hurdle to develop the ASR system for Indian Languages.

In the early 1920s machine recognition came into existence. The first machine to recognize speech to any significant degree was commercially named, Radio Rex (toy, manufactured in 1920) [5]. Research in speech technology began in early 1936 at Bell Labs. In 1939, Bell Labs demonstrated a speech synthesis machine (which simulates talking) at the World Fair in New York. The earliest attempts to devise systems for ASR by machine were made in 1950s, when various researchers tried to exploit the fundamental ideas of acoustic phonetics.

In English and other foreign languages, lots of work has been found in this field [6]. Michael Price, James Glass, Anantha P. Chandrakasan (2015) [7] described an IC that provides a local speech recognition capability for a variety of electronic devices. The chip performs 5,000 word recognition tasks in real-time with 13.0% word error rate, 6.0 mW core power consumption, and a search efficiency of approximately 16 nJ per hypothesis.

Assamese Language


Bengali Language

Automatic Speech Recognition - A Literature Survey on Indian languages and Ground Work for Isolated Kannada Digit Recognition using MFCC and ANN

for trained and tested sentences with fewer mixture components in HMMs. Manjunath, K.E. et al. (2013) [28] developed a Consonant-Vowel Recognition System (CVRS) for determining a sequence of Consonant-Vowel (CV) units present in a given speech utterance. Anukriti et al. (2013) [29] proposed a method to build up a speech recognition system which could interpret in any language (tested with Hindi and Bengali language) and implement it on Windows 7 platform which could be speaker independent. Malhotra, K. et al. (2008) [30] gives an approach to identify gender and accent of a speaker using Gaussian mixture modeling technique. In this approach spectral features have been incorporated in the form of mel frequency cepstral coefficients (MFCC). Babi, K.N. et al. (2012) [31], Hassan, F. et al. (2012) [32], Kotwal, M.R.A. et all (2011) [33], Asfak-Ur-Rahman, M et al. (2012) [34] and Hassan, F et al. (2011) [35] presented a Bangla ASR by suppressing gender effects. Ahamed, B. et al. (2013) [36] analyzed variation in performance of Bangla ASR based on accent. Toma, T.T. et al. (2013) [37] describes the effect of Bengali accent on English vowel recognition. It is seen that spectral characteristics of different English vowel sounds are largely influenced by the Bengali-accented speech.

Gujarati Language
Malde, K.D. et al. (2013) [38] described the development of speech corpora in Indian languages (viz., Gujarati and Marathi from remote villages) for the task of phonetic transcription. Undhad, A.G. et al. (2014) [39] proposed use of novel vowel landmark detection (VLD) algorithm for low resourced language, viz., Gujarati, an Indian language. The proposed VLD algorithm has detection rate of 78.92 %, 76.40 % and 73.89 %, which is 8.79 %, 7.23 % and 7.17 % more as compared to loudness-based method in lecture, spontaneous and read mode, respectively. Patil, H.A. et al. (2012) [40] addressed phonetic transcription related issues in Gujarati and Marathi (Indian Languages). The anusvara in both of these languages are produced based on the immediate following consonant. Implication for this finding for the problem of phonetic transcription is presented.

Hindi Language
Sinha, S. et al. (2013) [63] and Kumar, A. et al. (2014) [64] estimated number of Gaussian mixture component for Hindi database based upon the size of vocabulary. The experimental results show that the maximum performance of the proposed system is achieved when we use four component Gaussian mixtures HMM model. Singhvi, A. et al. (2008) [65] developed a phoneme segmentation and phoneme recognition for Hindi speech recognition. Pandey, B. et al. (2010) [66] proposed a recognition system for identification of the speaker, language and the words spoken. The system makes use of Adaptive Neuro-Fuzzy Inference paradigm for the same. Sakti, S. et al. (2009) [67] outlined the first Asian network-based speech-to-speech translation system developed by the Asian Speech Translation Advanced Research (A-STAR) consortium. Ali, M. et al. (1976) [68] applied adaptive pattern recognition theory for the recognition of speech patterns initially in the form of sonograms of CVC Hindi syllables spoken in isolation on the IBM 1130 computer. Agarwal, A. et al. (2010) [69] proposed an algorithm to detect word boundaries in continuous speech of Hindi language. The proposed algorithm is based mainly on two prosodic parameters, pitch and intensity. Upadhyaya, P. et al. (2013) [70] presented a study about the low dimension visual (LDV) space features and investigates the improvement in audio visual automatic speech recognition using different set of visual features. Dileep, A.D. et al. (2013) [71] addressed the issues in the design of an intermediate matching kernel (IMK) for classification of sequential patterns using support vector machine (SVM) based classifier for tasks such as speech recognition. They propose the HMM based IMK for matching sequential patterns of varying length. Roy, S. et al. (2014) [72] finds the dependence of Prominence on the phonetic cues (F0, Amplitude, Duration) in the form of mathematical function. They relate to the phonetic cues of F0, Amplitude, Duration, and Pauses. Vijayalakshmi, P. et al. (2011) [73] based on the analysis of coarticulation effect in consonant-vowel (CV) and vowel-consonant (VC) combinations, triphones that capture negligible coarticulation effect are replaced by the base monophone, in the triphone-based speech recognition system for Hindi. Anukriti, et al. (2013) [74] proposed a speech recognition system which could interpret in any language (tested with Hindi and Bengali language) and implement it on Windows 7 platform which could be speaker independent. Mittal, T. et al. (2013) [75] extracted spectral features from speech to perform speaker classification based on their age. Dileep, A.D. et al. (2013) [76] proposed a novel approach to design a pyramid match kernel (PMK) using hidden Markov model. They study the performance of the SVM-based classifiers using the proposed PMK for recognition of isolated utterances of E-set in English alphabet and recognition of consonant-vowel segments of speech in Hindi and compare with that of the SVM-based classifiers using score-space kernels and alignments kernels. Mohan, A. et al. (2013) [77] discussed the problem of building acoustic models for ASR using speech data from multiple languages. Goel, S. et al. (2010) [78] intend to provide an insight into the theoretical and implementation details of a speech based Dialog query system over Asterisk PBX server. Imran, A. et al. (2011) [79] developed a natural language Hindi speech interface to enable Hindi speaking population access market prices of commodities. Joshi, S. et al. (2013) [80] pronunciation assessment of vowels of Indian English uttered by speakers with Gujarati L1 using confidence measures obtained by automatic speech recognition. It is observed that Indian English speech is better represented by Hindi speech models for vowels common to the two languages rather than by American English models.

Kashmiri/Manipuriodia/nepali Languages

Malhotra, K. et al. (2008) [81] gave an approach to identify gender and accent of a speaker using Gaussian mixture modeling technique. The proposed approach is text independent and identifies accent among four regional Indian accents in spoken Hindi and also identifies the gender. The accents worked upon are Kashmiri, Manipuri, Bengali and neutral Hindi. Malhotra, K. et al. (2013) [82] brought out the impact of native language on accent in regional Indian Languages and also highlights the extent of the effect of native accent of a speaker on utterances spoken in non-native languages. Kashmiri affected the accent much more than Bengali or Manipuri. Regmi, B.N. et al. (2013) [83] studied multimodality in Own Communication Management (OCM) focusing on how linguistic communication involves gestures in order to manage communication. Some of the main findings from the study are that about 66% of all OCM expressions involve gestures, and that the distribution of choice and change function of OCM is about 90% to 10%. Allwood, J. et al. (2012) [84] presented a brief account of the principles, methodology, current status, and preliminary findings, based on an incrementally growing and multimodal activity based spoken language corpus of Nepali. Rahul, L. et al. (2013) [85] discussed the implementation of phoneme based Manipuri Keyword Spotting System (MKWSS). The performance of the system depends on the ability of detection of the keywords. An overall performance of 65.24% is obtained from the phoneme based MKWSS. Sunil Kumar,
Automatic Speech Recognition - A Literature Survey on Indian languages and Ground Work for Isolated Kannada Digit Recognition using MFCC and ANN

S.B. et al. (2013) [86] introduces a speech corpus in Indian languages namely Bengali and Odia, which provided phonetic and prosodic information. Phonetics and prosody are vital parameters in human speech perception.

Malayalam Language

Anand A.V. et al. (2012)[87] described the development of state-of-the-art large vocabulary continuous speech recognition (LVCSR) system with an application for visually challenged. The best configuration of the system achieved word accuracy of 75% in average. Accuracy of the system is further increased up to 80% in average, by implementing speaker adaptation technique. Kurian, C. et al. (2009) [88] and Renjith, S. et al. (2013) [89] presented speaker independent speech recognition system for Malayalam digits. Digit speech recognition is important in many applications such as automatic data entry, PIN entry, voice dialing telephone, automated banking system, etc. The system employs MFCC as feature for signal processing and ANN for recognition. Mohammed, M. et al. (2012) [90] discussed regarding an attempt to develop a digit recognition system using the HMM Toolkit (HTK). Thasleema, T.M. et al. (2007) [91] presented a novel and accurate feature extraction technique for recognizing Malayalam spoken vowels based on LPC method and compared the result with wavelet packet decomposition method. Raji Sukumar, et al. (2010) [92] discussed a novel technique for recognition of the isolated question words from Malayalam speech query. They have created and analyzed a database consisting of 500 isolated question words. Fast Fourier Transform (FFT) and Discrete Cosine transform (DCT) is used for the feature extraction purpose and ANN is used for classification and recognition. Kurian, C. et al. (2010) [93] developed a speech recognition of Malayalam isolated digit is created by using MFCC and SVM. Sangwan, A. et al. (2010) [94] proposed a language analysis and classification system that leverages knowledge of speech production. The proposed scheme automatically extracts key production traits that are strongly tied to the underlying language structure.

Marathi Language

Patil, P.P. et al. (2014) [95] presented Marathi connected word speech recognition system. The system is implemented using MFCC feature extraction technique and Continuous Density HMM (CDHMM). Kamble, V.V. et al. (2014) [96] analyzed emotion recognition from Marathi speech signals. In this paper several method are extracting feature from speech signal to estimation of energy, intensity and pitch contour using MFCC. Gaussian mixture Models (GMM) is used to develop Emotion classification model. Gaikwad, S. et al. (2013) [97] described the collection of audio corpus for Marathi language. The size of corpus collected is 28420 isolated words and 17470 sentences from around 500 speakers. The corpus is transcribed as well as annotated and is available for recognition system.Patil, P.P. et al. (2014) [98] described the ASR system that recognizes Marathi phoneme using CDHMM. In this paper MFCC for feature extraction. Baum Welch algorithm is used for re-estimating the parameters. Finally, the Viterbi algorithm is used to recognize the phoneme. Mohan, A. et al. (2013)[99] build an acoustic models for ASR using speech data from multiple languages. Techniques for multi-lingual ASR are developed in the context of the subspace GMM (SGMM). These techniques are applied to Hindi and Marathi language data obtained for an agricultural commodities dialog task in multiple Indian languages. Godambe, T. et al. (2013) [100] described the development of a continuous speech database. Speech data was collected from about 1500 literate speakers from 34 districts of Maharashtra, with a variety of characteristics such as age group, gender, mother tongue and educational qualification.

Punjabi Language

Singh, P. et al. (2011) [101] presented the analysis of vowel phonemes of Punjabi. It shows formant analysis of vowels produced by speakers of Punjabi as a first language from Punjab. Comparison and analysis have been given with ten vowel phonemes of English. Mean of the respective formant frequencies are different. Kaur, A. et al. (2010) [102] presented an effective method for segmentation of Punjabi Speech into syllable like Basic units for ASR systems. It is a well known fact Syllable based systems perform better than phoneme or word based systems.

Tamil Language

Rojathai, S et al. (2014) [103] proposed a Tamil speech word recognition system with Phase Autocorrelation (PAC). Ram, C.S. et al. (2014) [104] described Effective Automatic Speech emotion recognition on Human
computer Interaction for Tamil speech using SVM. Radha, V. et al. (2010) [105] proposed an improved Voice Activity Detector (VAD) for Automatic Tamil Speech Recognition System (ATSRS). The performance evaluation is done with the speech quality measures like MSE and PSNR. Waidyanatha, N. et al. (2012) [106] applied the International Telecommunication Union (ITU) recommended R800 Mean Opinion Score (MOS) and Difficulty Score (DS) voice quality evaluation methods. Harish, S. et al. (2011) [107] highlighted the importance of the segmented speech, language model and co-articulation effect which influences the speech production. Monophone and triphone based speech recognition systems are developed and their performance shows the importance of the above mentioned parameters. Plauche, M. et al. (2006) [108] presented an inexpensive approach for gathering the linguistic resources needed to power a simple spoken dialog system. Ganesh, A.A. et al. (2013) [109] introduced a new algorithm named Varied-Length Maximum Likelihood algorithm that is used for identifying the boundary of each character and explains the proposed process of speech recognition system for Tamil language. Nijusekar, C. et al. (2010) [110] proposed speech interactivity embedded module that is quite simplified and suitable for programmable anthropomorphic dialogue and menu driven recognition applications. Kumar, J.C. et al. (2010) [111] described the details of a bilingual speech recognition system. Kumar, J.C. et al. (2010) [112] proposed a novel idea for using two different feature streams in a continuous speech recognition system. In this paper, instead of concatenation, they build separate sub word HMMs for each of the feature streams during training.

**Telugu Language**

Beke, A. et al. (2012) [113] intended to automatically predict vowel duration in spontaneous speech based on three different methods. Nagamani, M. et al. (2005) [114] improved Intelligent tutor for Telugu language learning –for adult education. INTTELL is based on cognitive psychology in human computer interaction. Venkateswarlu, R.L.K. et al. (2012) [115] developed an efficient speech recognition system for Telugu letter recognition. Each telugu word ends with vowels. So there is a scope for research about Telugu vowels recognition rate. Sreenu, G. et al. (2004) [116] described the interaction with the system through telugu speech. This system works using only microphone. Ramamohan, S. et al. (2006) [117] proposed a sinusoidal model for characterization and classification of different stress classes (emotions) in a speech signal. Sinusoidal features perform better compared to the linear prediction and cepstral features. Amol T., et al. (2014) [118] recognized six basic emotions viz. Anger, Disgust, Fear, Happiness, Neutral and Sadness using selective features of speech signal of different languages like German and Telugu using SVM. Koolagudi, S.G. et al. (2010) [119] proposed epoch parameters extracted from LP (Linear Prediction) residual and zero frequency filtered speech signal for recognizing the emotions present in speech. Average emotion recognition of 61% and 58% is observed respectively for the above models. Rani, N.U. et al. (2013) [120] described the reduction of confusion pairs to recognize more number of words of different rates of speech. In this paper, training is performed on different speech rates (Normal, Slow and Fast) and testing also done on different rates of speech.

**Kannada Language**

Shridhara, M.V. et al. (2013) [121] explained collection of speech data in Kannada language for prosodically guided phonetic search engine and the issues involved in transcription. Parameswarappa, S. et al. (2012) [122] introduced the Kannada Corpus tool, a suite of Perl (Program Extraction and Reporting Language) programs implementing an iterative procedure to build Kannada corpora from the web. P. Punitha et al. (2014) [123] described the design of an algorithm to recognize continuous Kannada speech using HMM Method in the speaker dependent mode. LPC coefficients are used as features. K-means procedure is performed on the feature vectors to obtain the observation sequence. Muralikrishna, H. et al. (2013) [124] has implemented Kannada isolated digit recognition system using MFCC as feature vector and HMM as pattern recognizer. Performance of the system is evaluated and compared based on the MFCC along with its first and second order derivatives. Antony, P.J. et al. (2010) [125] presented the development of a part-of-speech tagger that can be used for analyzing and annotating Kannada texts. Sadanandam M. et al.(2012) [126] described a text independent language recognition system using new features derived from MFCC feature of speech signal with a common code book and discrete HMM to achieve very good LID recognition performance with less computation time comparing with that of a state of art phone based systems available in literature. Kashyap, K.H. et al. (2003) [127] described Hybrid neural network architecture for age identification of ancient Kannada scripts. Rajashekararadhya, S.V. et al. (2009) [128] proposed a zone-based hybrid feature extraction system. They obtained 97.85 %, 96.8 %, 95.1 % and 95 % recognition rates for Kannada, Telugu, Tamil and Malayalam numerals respectively, using support
Automatic Speech Recognition - A Literature Survey on Indian languages and Ground Work for Isolated Kannada Digit Recognition using MFCC and ANN

vector machine. Nagaraja, B.G. et al. (2012) [129] demonstrated speaker identification in the context of Monolingual and Cross-lingual for Indian languages with the constraint of limited data. The languages considered for the study are English, Hindi and Kannada. Sangwan, A. et al. (2010) [130] proposed a language analysis and classification system that leverages knowledge of speech production. The proposed scheme automatically extracts key production traits (or “hot-spots”) that are strongly tied to the underlying language structure. Parameswarappa, S. et al. (2011) [131] developed a WSD system for target words in Kannada language.

III. MOTIVATION

The above survey is done to the best of our knowledge and based on this it is observed that a lot of work on ASR is done in English and other foreign languages. As far as Indian languages are concerned, some works related to ASR are found in Hindi, Bengali, Assamese, Tamil, Telugu, Marathi, and Malayalam Languages. In Kannada Natural Language processing (NLP) some works are reported on Named Entity Recognition (NER) [132], Parts of Speech Tagger, speaker identification, but the ASR work is not explored as much as other major spoken Indian languages. This motivated us to take up ASR in Kannada as the research topic.

IV. PROPOSED WORK: ISOLATED KANNADA DIGIT RECOGNITION USING MFCC AND ANN

The Mel-Frequency Cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. ANN is a computational model that tries to simulate the structure and/or functional aspects of biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach [133]. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Neural networks are non-linear statistical data modeling tools [134].

The proposed speech recognition system can generally be divided in different components as illustrated in Figure 1. Speech was recorded in normal office environment using a microphone of 20Hz to 16000 Hz frequency range. The recording is done with 8 kHz sampling frequency quantized by 16 bits.

A. Methodology:

Kannada Speech recognition is processed in various steps such as: Framing, Voiced part extraction, Pre-emphasis, Feature extraction and Artificial Neural Network classifier. The algorithm is developed on MATLAB platform using Signal Processing Toolbox, Statistical Toolbox and Neural Network Toolbox.

Step 1: Framing
Speech is a highly non-stationary signal and hence speech analysis must be carried out on short segments, across which the speech signal is assumed to be stationary. For this the speech signal is divided into frames of small durations, typically 10 to 30ms with an overlap of 5 to 15ms. Here the frames of 10 ms consisting 80 samples without overlapping are considered.

Let \( x[n] \) be a speech signal with a sampling frequency of \( fs \), and is divided into \( P \) frames each of length \( N \) samples such that \( \{ x_1(n), x_2(n), x_3(n), \ldots, x_P(n) \} \), where \( x_i(n) \) denotes the \( i^{th} \) frame of the speech signal \( x[n] \) and is given by

\[
    x_i(n) = \{ x[i \ast N + n] \}_{n=0}^{N-1}
\]

(1)

The speech signal \( x[n] \) is represented in a matrix notation of size \( N \times P \), where \( N=80 \) and \( P=100 \).
Step 2: Voiced part extraction

The problem of locating the beginning and end of a speech utterance in a background of noise is of importance. The selection of speech signal that correspond to speech will eliminate the significant computation. The voiced part is extracted based on two simple time domain measurements: energy and zero crossing rates [135]. The amplitude of the speech signal varies appreciably with time and the amplitude of unvoiced segments is generally much lower than the amplitude of voiced segments. The short-time energy of the speech signal provides a convenient representation that reflects these amplitude variations. The short-time energy is calculated using rectangular window as given in equation (2).

$$E_n = \sum_{m=-\infty}^{\infty} [x(m) w(n-m)]^2; \text{ Where } w(n) \text{ is a window function}$$ (2)

The model for speech production suggests that the energy of voiced speech is concentrated below 3 kHz because of the spectrum falloff introduced by the glottal wave, whereas for unvoiced speech, most of the energy is found at highest frequencies. In general if the zero-crossing rate is high, the speech signal is unvoiced else it is voiced [136].

Step 3: Pre-emphasis

As high frequency components of speech have lesser amplitude, Pre-emphasis improves its SNR. The transfer function of the pre-emphasis filter is given in equation (3):

$$H(z) = 1 - az^{-1} \quad 0.9 \leq a \leq 1.0$$ (3)

Where ‘a’ are the filter coefficient and is chosen as 0.9375. The output signal y(n) after pre-emphasis is given in equation (4) [137]:

$$y(n) = x(n) - a \ast x(n-1)$$ (4)

Step 4: Feature extraction using MFCC

It is a process of extracting features from the input signal by reducing the dimension of the input-vector still maintaining the uniqueness of the signal. The outline of the computation of MFCC is shown in Figure 2.

![Figure 2: Computation of Mel Frequency Cepstral Coefficients (features)](image)

The MFCC features extraction is done in following steps:

(i) Windowing: Multiply pth frame x(n) is with a hamming window function given in equation (5)

$$w(n) = \begin{cases} 0.54-0.46\cos\left(\frac{2\pi n}{N}\right) & \text{for } 0 \leq n \leq N-1 \\ 0 & \text{otherwise} \end{cases}$$ (5)

(ii) DFT: calculate DFT of each frame to get spectrum.

(iii) Magnitude: Calculate the modulus of Fourier transform |X|.

(iv) Mel frequency filter banks: |X| is warped according to the Mel scale [138].

- For any given frequency f, measured in Hz, Mel is calculated by the equation (6)

$$\text{Mel}(f) = 2595 \ast \log_{10}\left(1 + \frac{f}{700}\right)$$ (6)

- [X] is segmented into a number of critical bands by means of a Mel filter bank which typically consists of a series of overlapping triangular filters defined by their center frequencies.

- The parameters that define a Mel filter bank are (a) number of Mel filters, F (b) minimum frequency, fmin and (c) maximum frequency, fmax.

- For speech, in general, it is suggested in [10] that fmin> 100 Hz. Furthermore, by setting fmin above 50-60Hz, we get rid of the hum resulting from the AC power, if present. In [139]
Automatic Speech Recognition - A Literature Survey on Indian languages and Ground Work for Isolated Kannada Digit Recognition using MFCC and ANN

it is also suggested that fmax be less than the Nyquist frequency. It is known that, there is no much information above 6.8 KHz in human speech.

(v) Take log: The logarithm of the filter bank outputs is taken.
(vi) DCT: Finally, DCT is taken to get MFCC features. Here each frame size is a vector of length 20.

Step 5: Classification by Using Artificial Neural Network
In this work ANN is used for pattern classification and recognition. The role of a recognizer is to assign the feature vector provided by the feature extractor to a category. It aims at measuring the similarity between an input speech and a reference pattern or model which is obtained during training. A Feed forward neural network with a single hidden layer is used for classification as shown in figure 3. MFCC features calculated are 500 and hence input layer of 500 nodes, hidden layer of 20 nodes and the output layer of 10 nodes constitute the network.

B. Database Creation:
In order to facilitate the training and testing of the recognizer, speech database is required. A variety of speech samples were obtained from different speakers to form the speech database. The collected database includes speech samples from 5 different speakers aged between 20 to 35 years. 70% of the collected data is used for training and remaining is used for testing. Apart from this, a small database is created for testing. The digits must be spoken clearly so that it avoids general variations and confusions.

C. Algorithm:
Based on the components dealt in previous section the following are the step by step procedures followed in programming our task.

1. Record a speech sample of a digit for 1 second. Let the sampling frequency be 8000 Hz (Nyquist criterion).
2. Frame the signal such that each frame consists of 10ms speech (80 samples / frame).
3. Extract the voiced part in the speech signal using energy and zero crossing rates.
4. Pre-emphasize the voiced part of a speech signal.
5. Extract the MFCC features of the speech sample.
6. Give the features to trained ANN to classify the signal.

V. RESULTS AND DISCUSSION
The digit database consists of ten utterances “zero” to “nine” collected from 5 male speakers. The database is divided into training and testing. Training set is used to train the neural network. Testing set is used to test the performance of neural network. The trained network is tested for 10 samples of each digit collected from speakers. This recognizer can be very well adapted for voice dialing. The histogram of digit recognition accuracy is shown in Figure 4.

B. Database Creation:
In order to facilitate the training and testing of the recognizer, speech database is required. A variety of speech samples were obtained from different speakers to form the speech database. The collected database includes speech samples from 5 different speakers aged between 20 to 35 years. 70% of the collected data is used for training and remaining is used for testing. Apart from this, a small database is created for testing. The digits must be spoken clearly so that it avoids general variations and confusions.

C. Algorithm:
Based on the components dealt in previous section the following are the step by step procedures followed in programming our task.

1. Record a speech sample of a digit for 1 second. Let the sampling frequency be 8000 Hz (Nyquist criterion).
2. Frame the signal such that each frame consists of 10ms speech (80 samples / frame).
3. Extract the voiced part in the speech signal using energy and zero crossing rates.
4. Pre-emphasize the voiced part of a speech signal.
5. Extract the MFCC features of the speech sample.
6. Give the features to trained ANN to classify the signal.

V. RESULTS AND DISCUSSION
The digit database consists of ten utterances “zero” to “nine” collected from 5 male speakers. The database is divided into training and testing. Training set is used to train the neural network. Testing set is used to test the performance of neural network. The trained network is tested for 10 samples of each digit collected from speakers. This recognizer can be very well adapted for voice dialing. The histogram of digit recognition accuracy is shown in Figure 4.
As shown in table 1, the digits 0, 3, 6 and 9 are recognized with an accuracy of 90% whereas as remaining digits 1, 2, 4, 5, 7 and 8 are recognized with an accuracy of 100%. The system developed works with an average accuracy of 96% when it is tested with the test database. When it works in normal office environment, with a different group of people, the system accuracy reduces. The speech waveform of digit 2 is shown in figure 5 along with its energy, zero crossing rate and voiced part.

### Table 1: Recognition Accuracy

<table>
<thead>
<tr>
<th>Kannada digits (kannada digits in words)</th>
<th>% 0</th>
<th>% 1</th>
<th>% 2</th>
<th>% 3</th>
<th>% 4</th>
<th>% 5</th>
<th>% 6</th>
<th>% 7</th>
<th>% 8</th>
<th>% 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kannada digits</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>90</td>
<td>100</td>
<td>100</td>
<td>90</td>
<td>100</td>
<td>100</td>
<td>90</td>
<td>100</td>
<td>100</td>
<td>90</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

VI. CONCLUSION

In this research paper an overview of Automatic speech recognition in Indian Languages is presented and also an attempt is made to build a model for Kannada isolated digit recognition using MFCC as feature vector and ANN as classifier. ANN promises to be a successful and powerful approach for effective classification. Digit recognition finds applications in voice dialing systems and it needs to be a speaker independent system. This system acts as a basis for real time speech recognition products, where input will never be a single word. In this work promising recognition accuracy is achieved for offline testing. The model has to be tested in a real time environment in future. The accuracy increases with the increase in training data, as a result, memory needed also increases. Hence there should be a tradeoff between database size and accuracy.

REFERENCES

Automatic Speech Recognition - A Literature Survey on Indian languages and Ground Work for Isolated Kannada Digit Recognition using MFCC and ANN


[16] Patgir,C. ; Sarma, M. ; Sarma, K.K., “RNN and SOM based classifier to recognize assamese fricative sounds designed using frame based temporal feature sets”, International Joint Conference on Neural Networks (IJCNN), pp. 3496 – 3502, 2014


[18] Shahnawazuddin, S ; Thotappa, Deepak ; Sarma, B D ; Deka, A ;Prasanna, S R M ; Sinha, R., “Assamese spoken query system to access the price of agricultural commodities”, National Conference on Communications (NCC) , pp. 1 – 5, 2013


Kumar, A.; Dua, M.; Choudhary, A. “Implementation and performance evaluation of continuous Hindi speech recognition” International Conference On Electronics And Communication Systems (ICECES), Page(s): 1 – 5, 2014


Mehta, K.; Anand, R.S. “Robust front-end and back-end processing for feature extraction for hindi speech recognition”, IEEE International Conference on Communication Intelligence and Computing Research (ICCC), Pp. 1 – 4, 2010


Automatic Speech Recognition - A Literature Survey on Indian languages and Ground Work for Isolated Kannada Digit Recognition using MFCC and ANN


[92] Raji Sukumar, A.; Sarin Sukumar, A.; Firoz Shah, A.; Anto P, B." Key-Word Based Query Recognition in a Speech Corpus by Using Artificial Neural Networks” Second International Conference on Computational Intelligence, Communication Systems and Networks (ICCCSN), Pp.33 – 36, 2010

[113] Green, J.G.; Srinivasan, M.; Prasad, M.N.; Girija, P.N.," Improving reading and writing skills with intelligent tutor for Telugu language learning - INTELL” International Conference on Intelligent Sensing and Information Processing, Pp.426 - 49b, 2005

ISSN 2277-1956/V4 N1-91-105
Automatic Speech Recognition - A Literature Survey on Indian languages and Ground Work for Isolated Kannada Digit Recognition using MFCC and ANN


[129] Nagaraja, B.G. ; Jayanna, H.S.” Mono and Cross lingual speaker identification with the constraint of limited data” International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME), Pp.439 - 443, 2012


[134] Raji Sukumar A, Sarin sukumar A, Firoz Shah A, Babu Anto P.” Key-Word Based Query Recognition In a Speech Corpus By Using Artificial Neural Networks”, 2nd International Conference on Computational Intelligence, Communication Systems and Networks, pp.33-36, 2010


