Isolated Word Recognition System for Malayalam using Machine Learning

Maya Moneykumar
IIITM-K, Trivandrum
maya.moneykumar@iiitmk.ac.in

Elizabeth Sherly
IIITM-K , Trivandrum
sherly@iiitmk.ac.in

Win Sam Varghese
IIITM-K , Trivandrum
sam.varghese@iiitmk.ac.in

Abstract

Automatic Speech Recognition (ASR) has received greater level of acceptance as it creates speech recognition by the human machine interface. This paper focuses on developing a syllable based speech recognition system for Malayalam language. The proposed system consists of three different phases such as preprocessing, segmentation and classification. The preprocessing is performed for noise reduction, DC component removal, pre-emphasis and framing. The segmentation process implemented using Syllable Segmentation Algorithm segments the word utterances into syllables, that are inturn fed into the system for feature extraction. In the feature extraction step, we have proposed a novel approach by adding energy and zero crossing, along with MFCC features. The classification is done using Artificial Neural Network and is also compared with HMM classifier. Experiments are carried out with real-time utterances of 100 words, and obtained 96.4 % accuracy in ANN, which outperformed HMM.

1 Introduction

Automatic Speech Recognition (ASR) system, especially speech to text conversion is one of the most challenging tasks in nowadays. The ultimate aim of ASR is to understand spontaneous speech using a computer. However, a word-level identification, which is a complex task in itself is also a major chore in any ASR system. In this paper, a word-level identification is performed to make a computer identify the isolated words spoken by a person and to convert it into corresponding text. ASR is basically a pattern recognition problem which also involves a number of technologies and research areas like Signal Processing, Natural Language Processing, Statistics and Cognitive Science. Different factors like gender, emotional state, accent and pronunciation make speech recognition a complex task. The mode of articulation, nasality, pitch, volume, and speed variability in speech also make speech recognition a difficult task. Even when speech is recognized, the accuracy rate can be less due to various behaviors of speech, usage of words, higher variability, background noise and the linguistic features of language and speech. This can seriously affect the system performance. There exists works which attempt for speaker dependent and speaker independent recognition. Our attempt is to achieve Speaker independence, which is difficult to achieve because in order to recognize the speech patterns, these models should be trained with speech data of a large group of people.

Developing an efficient speech recognizer that can exhibit natural capabilities of every human possesses, along with language capability is much harder. As far as Malayalam language is considered, which has a rich set of vocabulary and the modern Malayalam alphabet has 15 vowel letters, 41 consonant letters, and a few other symbols. Malayalam is one of the richest languages in terms of number of alphabets and also one of the toughest languages while considering speech. This is because of the variations in pronunciation. Speech recognition in Malayalam language is still in its infancy stage and the dream of a system which can interact with the users in native language is still in its early stage. Hence, developing an efficient speech recognition system in Malayalam has great relevance. An isolated word identification system aims at identifying a spoken word from the trained vocabulary.

The work is explained in this paper in 6 sections. The first section is the abstract of the work. The second section gives an introduction where the

D S Sharma, R Sangal and E Sherly. Proc. of the 12th Intl. Conference on Natural Language Processing, pages 158–165,
Trivandrum, India. December 2015.

problem and the relevance of the problem in the current scenario is explained. The third section gives a review of already existing methods and the comparative study about the advantages and disadvantages of the existing methods. The fourth section is the core part of this work which introduces the methodology adopted to solve the problem, and the implementation of the algorithm for the same. The results are analyzed in the fifth section which also contains a detailed study about the results obtained. The future scope of the work is included in the last section.

2 Literature Review

ASR technology, during the yesteryears has made advancement to a great extent and has reached the point where this facility is used in various fields by millions. Speech recognition works, in foreign languages, advanced a lot since 1920. The first work in speech recognition was the development of a toy named Radio Rex, a celluloid dog,which was developed in 1920. Even though researches were carried out since 1936, the first successful speech recognizer was developed in the year 1952 by David et.al, which could recognize digit utterances by a single speaker. The system used spectral energy and formant frequencies for recognition purpose. Another notable development in this field was the implementation of phoneme based speech recognizer in the year 1959. In 1960s Japan made their first leap into the field of speech recognition and developed a special purpose hardware to improve the computational speed of the then systems. They also developed a hardware phoneme recognizer in 1962 followed by a digit recognizer in 1963 (Nnamdi Okomba S et al. , 2015).

The ASR works were extended to the field of isolated word utterance recognition in 1970s, during which prominent works were carried out by Velichko and Zagoruyko in Russia , Itakura in United State and Cakoe and Chiba in Japan. Meantime, CMU also played their role in the ASR field. Several other systems such as HEARSAY II(1975), HARPY(1976), HWIM (1977) and KEAL (1977) were implemented during this period. A major shift in the ASR technology happened in 1980, where the template based approach got replaced by statistical modeling methods like Hidden Markov Model (HMM). Different international languages, including English, French and Pashto implemented speech recognition systems with HMM and proved successful. An automatic dictation system in French implemented using HMM gave an accuracy of 76.2%. A digit recognition system for English implemented using HMM gave an accuracy rate of 88%. It was in 1980s itself, that Support Vector Machine and Neural Network approach got popularity despite of its faded entry in the early 1950s. This period also witnessed the progress in speech recognition works carried out with continuous speech. ASR using Artificial Neural Networks (ANN) achieved excellent results in tasks such as voiced/unvoiced discrimination in 1989 as well as phoneme recognition and spoken digit recognition in 1989. Peeling and Moore applied Multi Layer Perceptron (MLP) to digit recognition and thereby obtained excellent results. Various foreign languages, including Malay, Indonesian and Arabic, later implemented ASR using ANN. The work carried out in Malay language using MLP considered utterances of 4 different speakers and obtained an accuracy of 95% in identifying words within the trained vocabulary (Ahmed et al., 2012; Sheila D Apte, 2012; Nnamdi Okomba S et al. , 2015).

Automatic Speech Recognition has tremendous potential in the Indian scenario, as well. Since common man depends on the internet and its services, they will be interacting with machines every time. In order to enjoy the benefits completely and to bridge the digital divide, the communication should happen in their local languages. Keeping this fact in mind, most of the Indian languages have worked and are still working with ASR tasks. Text- to speech conversion has been developed for the visually challenged section of society, for Indian languages such as Hindi, Bengali, Marathi, Tamil, Telugu and Malayalam. Isolated Word speech recognition system is built for most spoken Indian languages namely Telugu, Hindi, Urdu, Kannada, Marathi, Tamil, Malayalam, Bengali and Oriya using Hidden Markov Model tool kit (HTK). It works as text dependent speaker recognition mode. A notable work in Tamil - 'A syllable based isolated word recognizer for Tamil handling OOV words', uses a subword based continuous speech recognizer for word identification. Notable works were carried out during the period 2000 to 2005 in languages including Hindi, Tamil and Telugu. Some of the works include Speaker Independent Continuous speech recog-

nizer for Hindi(2000), Speech recognizer for Specific Domain in Tamil (2001), Speech Recognition of Isolated Telugu Vowels Using Neural Networks (2003) and Digit Recognizer for Hindi (2006). Several other Indian languages like Punjabi, Marathi, Gujarathi, Assamese and Kannada successfully imprinted their footsteps in the field of ASR during early 2000s (Kurian Cini, and Kannan Balakrishnan. , 2012; Akila A and E. Chandra, 2013; Akila A and E. Chandra, 2014).

Eventhough Malayalam is still in the budding stage while considering ASR, there exist some noteworthy works and the significant one was from CDAC, Trivandrum where they developed a speech recognition system for visually impaired people. MFCC method was used as a front-end to extract acoustic features from the input signal and a hybrid model integrating rule based and statistical method was used to handle pronunciation variations in the dictionary. The system achieved word accuracy of 75%. A speaker independent continuous speech recognizer based on PLP Cepstral Coefficient, was also developed for Malayalam which employs Hidden Markov Model for pattern recognition. The system got trained with 21 male and female speakers and obtained a word accuracy of 89% when tested with continuous speech data. While considering LPC features for classification, Malayalam speech recognition system obtained an accuracy rate of 81.2%, which is indeed a notable work. Vimal Krishnan et.al developed a small vocabulary (5 words) speech recognition where Artificial neural network technique (ANN) is used for classification and recognition purpose and achieved a recognition rate of 89%. Raji Kumar et.al presented recognition of the isolated question words from the Malayalam speech query using DWT and ANN and recognition accuracy of 80% has been reported.

3 Methodology

The word identification system designed for Malayalam language uses a syllable based segmentation approach. Instead of training the system with independent words, we use syllable combinations for training and identification. This eliminates the problem of maintaining a large set of training data, as different words share common syllables as well as phone segments. New words can be added to the vocabulary without building new models for the existing syllables and phonemes corresponding to the word. Here, we design a word identification system based on two popular classifier techniques which are Hidden Markov Model (HMM)and Artificial Neural Network (ANN). The various phases involved in speech recognition include data preparation, syllabification and classification. The general system architecture of the speech recognition system implemented in this work is given below.

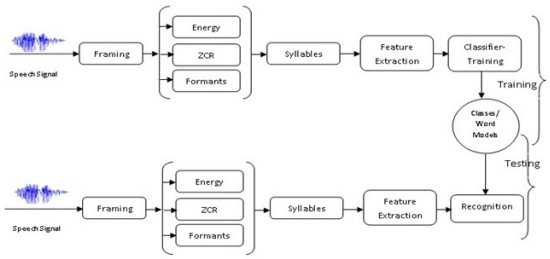

Figure 1: General System Architecture

3.1 Data Preparation

The work maintains a database of utterances which consists of a total of 100 words used in the agriculture domain. Multiple utterances of these words by 9 different speakers(6 male & 3 females) were recorded using an audio tool Audacity with a sampling frequency of 16000Hz. A speech corpus was in turn developed which consists of these audio files along with the syllable transcription.

In audio recording, a DC offset is an undesirable characteristic of a recording sound which occurs during sound capturing. The offset causes the center of the waveform to not be at 0, but at a higher value. This can cause two problems:

- Either the loudest part of the signal will be clipped prematurely, since the base of the waveform has been moved up
- Inaudible low frequency distortion will occur.

The signal, hence should undergo a DC component removal which is done as

$$x(n) = a(n)avg, \quad avg = mean(a); \quad (1)$$

where a is the speech signal

The pre-processing stage consists of noise

reduction and filtering. Pre-emphasis refers to a process designed to increase the magnitude of some (usually higher) frequencies with respect to the magnitude of other (usually lower) frequencies in order to improve the overall signal-to-noise ratio. In speech processing, a pre-emphasis is applied by high pass filtering the signal to smooth the spectrum to achieve a uniform energy distribution spectrum.

Pre-emphasis is done by applying the formula

$$y(n) = x(n) - 0.9955 * x(n-1) \qquad (2)$$

In speech processing, it is often advantageous to divide the signal into frames to achieve stationarity. A frame based analysis is essential for speech signal. Framing is the process of decomposing the speech signal into a series of overlapping/non-overlapping frames. The speech signal is stationary within windows of 20 to 30 ms duration and hence is divided into non-overlapping frames with a duration of 30 ms (Sheila D Apte, 2012) .

3.2 Syllabification & Feature Extraction

3.2.1 Syllabification

In this work, syllabification plays an important role and hence every recorded utterance which becomes the input to the system should eventually undergo syllabification. Instead of using commonly used approaches for syllabification including group delay method, we have developed a novel approach to syllabify utterances which make use of short time energy, zero crossing rate as well as formant frequencies. In syllabification, each and every utterance is segmented into respective syllables irrespective of the number of syllables it consist of. The Syllable Segmentation Algorithm introduced in this work is implemented using Matlab and has obtained an accuracy of 95%.

Syllable Segmentation Algorithm

The algorithm segments syllables, based on the concept that the energy measure for a vowel (voiced segment) is much higher than the silence part and the energy measure of the consonant(unvoiced segment) is lower than the voiced segment(vowel) but higher than the silence part. The zero crossing rate is higher for the silence part and lower for the voiced part of a signal.

Algorithm 1: Algorithm for Syllable Segmentation

Input: *Signal of Isolated Word utterance*: I
Output: *.wav files of Syllables*

1 *Signal Acquisition*
2 *Preprocessing*
3 *Apply a high pass filter and calculate ZCR*
4 *Apply low pass filter and calculate Energy*
5 *Calculate the formants F1 & F2*
6 *calculate the peaks*
7 *Calculate the Residual Energy*
8 *Find the approximate syllable boundaries using the calculated values.*
9 *Attenuate the syllables based on their energy to get the accurate syllable boundary.*
10 *Apply window function*
11 *Store the segmented syllables as wav files*

After obtaining the energy and ZCR values of frames, it is necessary to smooth the signal to find the energy and ZCR peaks. Smoothing means that we even out a signal, by mixing its elements with their neighbors. In order to obtain an accurate measure, we are considering two different energy calculations. The first one is the energy of the signal which is the short time energy and the second is the energy of the signal filtered with low pass filter with a cutoff frequency of 1100Hz. This energy calculation will help us in discarding the mis identified syllable boundaries.

The preprocessed signal then undergoes the syllabification process by detecting the syllable boundaries. Energy and Zero crossing

$$E = \sum_{n=0}^{N-1} s^2(n) \qquad (3)$$

and zero crossing rate as

$$Z = \sum_{n} |sgn(s(n)) - sgn(s(n-1))| \qquad (4)$$

where $sgn(s(n)) = 1, if s(n) \geq 0$,
$sgn(s(n)) = -1, if \quad s(n) < 0$

The syllable segmentation for a word with its speech signal is depicted below in Fig. 2.

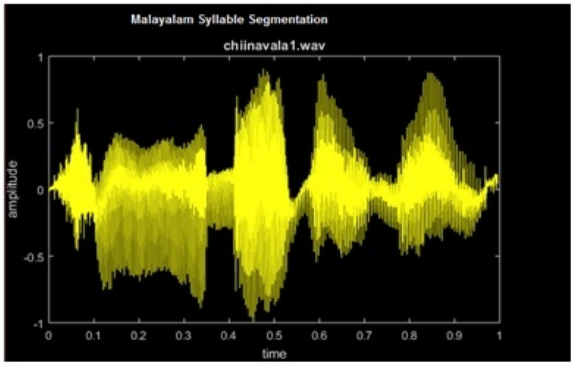

Figure 2: Speech signal for utterance chiinavala

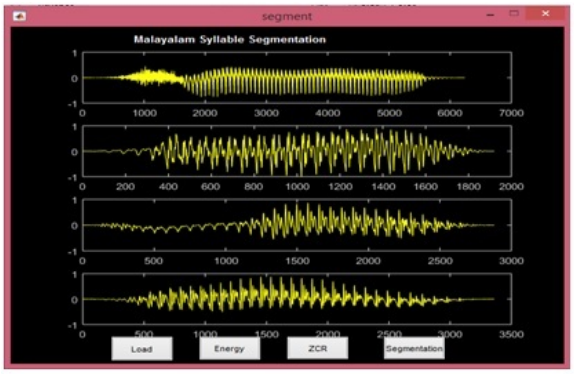

Figure 3: Result of Syllabification

3.2.2 Feature Extraction

The syllables which are stored as .wav files are then fed into the feature extraction process. The syllabified speech signal is divided into non-overlapping frames of 30 ms length by keeping in mind the concept that spectral evaluation are reliable in the case of a stationary signal and that a speech signal is stationary within a window of duration 20-30 ms. For the purpose of feature extraction, spectral analysis algorithm such as MFCC is used. Cepstral analysis has been widely used as the main analysis method for speech recognition. The mel-frequency cepstral coefficients (MFCC) appear to furnish a more efficient representation of speech spectra than other analysis methods. Human auditory system is assumed to process speech signal in a non-linear fashion. Lower frequency components of speech signal contain more information. The mel-frequency cepstrum (MFC) can be defined as the short-time power spectrum of a speech signal which is calculated as the linear cosine transform of the log power spectrum. Conversion from normal frequency to mel-frequency is given by

$$m = 2595log10(1 + (f = 700))(4.7) \quad (5)$$

where f indicates normal frequency and m is the corresponding mel-frequency. In order to improve the recognition accuracy, later two additional parameters, energy and zero crossing rate are also considered along with MFCC.

3.3 HMM Classification System

After preprocessing, syllabification and feature extraction, HMM is used to recognize the speech and training. The system is being trained using task grammar, acoustic models and lexical models using HTK toolkit.

For training, a 5 state HMM prototype model was created with the first and last non- emitting states. The main step in HMM training is defining a prototype model as a model structure. A train file is also created which will redirect HTK to the location where the feature vector files (mfcc files) are stored. Realigning the training data is done next where the word-to-phone mapping operation is performed. In this case, all pronunciations for each word is considered and then output the pronunciation that best matches the acoustic data. In order to recognize the word using phoneme combinations, triphone files are also generated (Pammi S. C and V. Keri., 2005).

The testing phase which is responsible for recognizing the utterances also follows the same set of steps till feature extraction, as in the training phase. The testing signals were also converted into a series of feature vectors. During the testing phase, recognition happens where the decoder compiles the recognition network using the task level word network. The word level transcriptions are constructed from the recognition network, as the next step.

3.3.1 ANN Based Speech Recognition System

The speech recognition system based on ANN was implemented using Multilayer Perceptron (MLP) which is a popular form of neural network. The input vectors are 12 MFCC values, along with two additional features energy and zero crossing

rate. The architecture used for ANN based ASR is given below.

Table 1: ANN Architecture

No: of input neurons	60
No: of neurons in the hidden layer	100
No: of hidden layers	1
No: of output neurons	80
Activation function	Sigmoid

The system is trained using Backpropagation algorithm with learning rate 0.3 and number of epochs 200. The training and testing were performed with various scenarios. The system is trained only with MFCC and MFCC, along with energy and zero crossing. A notable greater performance is shown in later case.

4 Result & Discussions

The goal of this work was to design a speech recognition system for Malayalam. In order to identify a method which gives better recognition accuracy, we used two well known approaches which will solve pattern recognition problems. The system was trained with both Artificial Neural Network and Hidden Markov Model so that if a syllable is given, it will identify the class to which it belongs to. To test the performance of both ANN and HMM, various test data were given where the accuracy in recognizing words within the vocabulary and out of vocabulary words, by the same as well as different speakers. The test data is categorized as follows:

Exp	Test Data	Speaker
Test1	OOV	Same Speaker
Test2	OOV	Diff. Speaker same gender
Test3	within the vocabulary	Same Speaker
Test4	within the vocabulary	DIff. Speaker same gender
Test5	OOV	Diff. gender
Test6	within the vocabulary	Diff. gender

Table 2: Test Data Set

The system was provided a data set with a total of 564 instances out of which 70% of the data was taken for training and rest 30% for testing. By comparing overall performance of both HMM and ANN, it was found that ANN outperformed HMM in almost all the test cases except for that of a different gender utterances. It was also found that the speech recognition system based on ANN performs better while adding more features other than the conventional MFCC features.

Given, are the results of recognition by both HMM and ANN, while using only MFCC features for training and testing.

Here, ANN has shown a better recognition accuracy in especially different speaker utterances and out of vocabulary utterances. But the performance of both HMM and ANN were not satisfactory in the case of different gender utterances, even though HMM showed slightly improved result.

Exp	HMM	ANN
Test1	87.5%	89.8%
Test2	62%	75%
Test3	91.67%	95.83%
Test4	93%	85.7%
Test5	79.17%	62.5%
Test6	80%	64.2%

Table 3: Results comparison HMM & ANN

Figure 4: Results comparison HMM & ANN

In order to improve the performance rate, along

with the MFCC parameters, energy measure as well as zero crossing rate of the signal was also considered as attributes for ANN. This choice was proved correct by the recognition accuracy resulted. Given below, is the table and graph depicting the difference in performance by the system trained with ANN where the parameters selection alone is different. The total instances used for training and testing and also the test conditions were the same.

Exp	ANN (only mfcc)	ANN (mfcc,energy,zcr)
Test1	89.8.5%	95%
Test2	75%	81.25%
Test3	95.83%	99.2%
Test4	85.7%	96.4%
Test5	62.5%	62.5%
Test6	64.2%	65%

Table 4: ANN Results Comparison

Eventhough, addition of new parameters helped in improving recognition accuracy of Test 1 to Test 4, Test 5 and Test 6 didnt show any improvement. The system didnt successfully identify a different gender utterance.

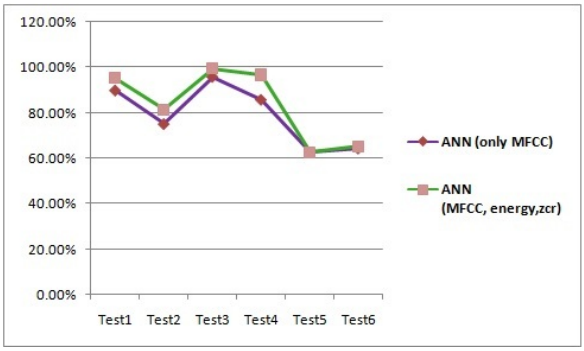

Figure 5: ANN Result Comparison

The diagram below depicts the overall performance comparison of the system performance.

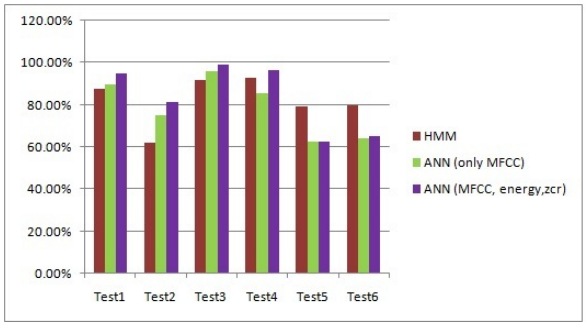

Figure 6: Overall Performance Comparison

5 Conclusion & Future Scope

This work addresses the problem of automatic speech recognition for isolated words in Malayalam language. A new algorithm was proposed with a feature set of formant frequencies, energy measure and zero crossing rate. An ANN based and HMM based models have been created and testing was performed words within the vocabulary and out of vocabulary. Based on performance evaluation, it is shown that Artificial Neural Network is well suited for speech recognition system of isolated words. However, the system performance is not satisfactory for gender specific.

5.1 Future Directions

In this research work, a speech recognition system with a moderate degree of accuracy is designed. The emphasis was on speech recognition for isolated words which finds applications in different areas where there is a man- machine interface. By considering the various factors, including noisy environment and gender specific, a deep leaning approach may be considered as future scope. The system shall also be extended for continuous speech recognition using syllable based approach. Attempts should be made to identify the features which will help the system identify utterances irrespective of the gender of the speaker.

References

. Ahmed, Irfan, Nasir Ahmad, Hazrat Ali, and Gulzar Ahmad. 2012. *The development of isolated words pashto automatic speech recognition system.*, In Automation and Computing (ICAC), 18th International Conference on IEEE, 1-4

. Al-Qatab Bassam AQ, and Raja N. Ainon. 1983. *Arabic speech recognition using hidden Markov model*

toolkit (HTK).. In Information Technology (ITSim), International Symposium on IEEE, vol. 2, 557-562

. Kurian Cini, and Kannan Balakrishnan. 2012. *Continuous speech recognition system for Malayalam language using PLP cepstral coefficient*. Journal of Computing and Business Research, vol. 3.1

. Akila A and E. Chandra. 2013. *Isolated Tamil Word Speech Recognition System Using HTK*. International Journal of Computer Science , vol. 3.02, 30-38

. Akila A and E. Chandra. 2014. *Performance enhancement of syllable based Tamil speech recognition system using time normalization and rate of speech*. CSI Transactions on ICT , vol. 2.2, 77-84

. Khetri G. P., Padme S. L., Jain D. C., Fadewar H. S., Sontakke B. R., and Pawar V. P. 2012. *Automatic Speech Recognition for Marathi Isolated Words*. Application or Innovation in Engineering & Management (IJAIEM) , vol.1(3)

Resch Barbara 2003. *Automatic Speech Recognition with HTK*. Signal Processing and Speech Communication Laboratory. Inffeldgase. Austria , Disponible en Internet: http://www. igi. tugraz. at/lehre/CI .

Sheila D Apte 2012. *Speech and audio processing*. Wiley publication , Feb 2012

Sunny Sonia, S. David Peter, and K. Poulose Jacob 2013. *Performance of Different Classifiers in Speech Recognition*. International Journal of Research in Engineering and Technology , vol 2

Kurian Cini and Kannan Balakrishnan 2011. *Malayalam Isolated Digit Recognition using HMM and PLP cepstral coefficient*. International Journal of Advanced Information Technology (IJAIT) , vol 1.5

Kurian Cini and Kannan Balakrishnan 2011. *Malayalam Isolated Digit Recognition using HMM and PLP cepstral coefficient*. International Journal of Advanced Information Technology (IJAIT) , vol 1.5

Pammi S. C., and V. Keri 2005. *A package for automatic segmentation.*.

Zegers Pablo 1998. *Speech recognition using neural networks*. Diss. University of Arizona

Gaikwad, Santosh K. Bharti W. Gawali and Pravin Yannawar 2010. *A review on speech recognition technique*. International Journal of Computer Applications , vol 10.3

Sunny Sonia, D. Peter, and K. Jacob 2013. *Combined Feature Extraction Techniques And Naive Bayes Classiffier For Speech Recognition.* CS & IT-CSCP

Jurafsky D. and Martin J 2000. *Speech and Language Processing: An Introduction to Natural Language Processing*. Computational Linguistics and Speech Recognition. Delhi, India , Pearson Education

Nnamdi Okomba S., Adegboye Mutiu Adesina, and Candidus O. Okwor. 2015. *Survey of Technical Progress in Speech Recognition by Machine over Few Years of Research*. IOSR Journal of Electronics and Communication Engineering Vol 10, Issue 4, Ver. I (Jul - Aug .2015), PP 61-67