Study of Speech Recognition Technology and its Significance in Human-Machine Interface

Jay V. Vyas
Department of Electronics & Communication Engineering
L.J Institute of Engineering & Technology, Ahmedabad, India

Dr. Anil C. Suthar
Department of Electronics & Communication Engineering
L.J Institute of Engineering & Technology, Ahmedabad, India

Abstract

This paper gives brief information regarding Speech Recognition Technology by including various speech parameters, different approaches to speech recognition, basics of acoustic models, language models, complex algorithms and feature extraction techniques. The objective of paper precisely focusses on speech processing and its applicability in technologies like Human-Machine-Interface (HMI) and various day to day life applications.

Keywords: HMI, Feature Extraction, Acoustic Models, Language models, HMM, DTW

I. INTRODUCTION

Speech is the primary mode of communication among human beings. It is the most natural and efficient way of exchanging information. Now, before getting into the depth of topic, let us first understand what is Speech Recognition? Its answer is- Speech Recognition is the inter-disciplinary sub-field of computational linguistics that develops methods and technologies that enables recognition and translation of spoken language into the text by computers. Now, let us look at the brief history of Speech Recognition technology. Research in automatic speech recognition by machine has been done for almost four decades. In 1960's several fundamental ideas in speech recognition surfaced and were published. However, the decade started with several Japanese laboratories entering the recognition arena and building special-purpose hardware as a part of their systems. [2] In 1960's three key research projects were initiated that have had major application on the research and development of speech recognition for the past 20 years. [2] A final achievement of note in 1960's was the pioneering research of Reddy in the field of continuous speech recognition by dynamic tracking of phonemes. In 1970's speech-recognition achieved a number of significant mile stones [2].

Finally, at AT & T Bell labs, researchers began a series of experiments aimed at making speech recognition systems that were truly speaker independent. To achieve this goal a wide range of sophisticated clustering algorithms were used to determine the number of distinct patterns required to represent all variations of different words across a wide user population. This research has been refined over a decade so that the techniques for creating speaker-independent patterns are now widely used. Speech research in the 1980's was characterized by a shift in technology from template-based approaches to statistical modelling methods especially the Hidden Markov Model approach [2]. Although the methodology of HMM was well understood. Finally, the 1980's was a decade in which a major impetus was given to large vocabularies, continuous speech recognition systems by the Defence Advanced Research Projects Agency community, which sponsored a large research program aimed at achieving high word accuracy for a 1000-word, continuous speech recognition, database management task. DARPA program has continued into the 1990's, with emphasis shifting to natural language front ends to the recognizer, and the task shifting to retrieval of air travel information.[2]

II. PROCESS OF SPEECH PRODUCTION AND PERCEPTION IN HUMANS

Figure 1 shows a schematic diagram of the speech-production/ speech-perception process in human beings. The production (speech- generation) process begins when the talker formulates a message in his mind that he wants to transmit to the listener via speech. The machine counterpart to the process of message formulation is the creation of printed text expressing the words of the messages. The next step in the process is the conversion of message into a language code. This roughly corresponds to converting the printed text of message into a set of phoneme sequences corresponding to sounds that make up words, along with the prosody markers denoting duration of sounds, loudness of sounds, and pitch accent associated with the sounds. Once the language code is chosen, the talker must execute a series of neuromuscular commands to cause the vocal cords to vibrate when appropriate and to shape the vocal tract such that the proper sequence of speech sounds is created and spoken by the talker, thereby producing an acoustic signal as the final output.

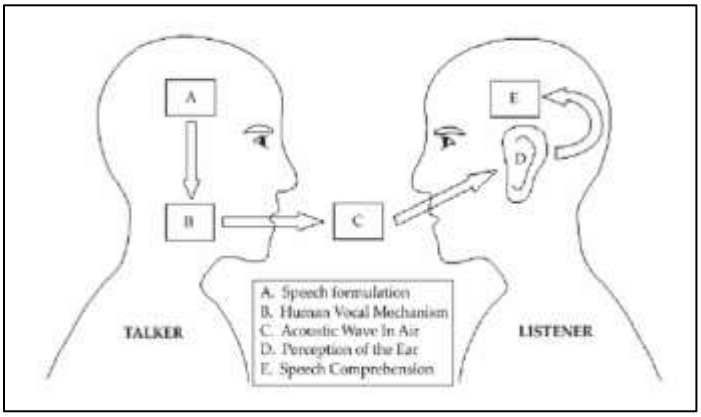

Fig. 1: Schematic diagram of speech-production/speech-perception process

Once the signal is generated and propagated to the listener, the speech-perception process begins. First the listener processes the acoustic signal along with the membrane in the inner ear, which provides a running spectrum analysis of the incoming signal. A neural transduction process converts the spectral signal at the output of basilar membrane into activity signals on the auditory nerves, corresponding roughly to feature extraction process. The neural activity along the auditory nerve is converted into a language code at the higher centre of processing within the brain, and finally message comprehension is achieved.

III. APPROACHES TO AUTOMATIC SPEECH RECOGNITION

In this section we provide an over view of several proposed approaches to Automatic Speech Recognition (ASR) by machine with the goal of providing some understanding as to essentials of each proposed method, and the basic strengths and weaknesses of first two approaches. Artificial Intelligence approach is out of this paper's scope. One can refer to book "Fundamentals of Speech Recognition- by Lawrence Rabiner" for more details.
Broadly speaking, there are three approaches to speech recognition namely:

1) The Acoustic-Phonetic Approach
2) The Pattern Recognition Approach
3) The Artificial Intelligence Approach.

A. Acoustic Phonetic Approach

Figure 3 shows a block diagram of the acoustic-phonetic approach to speech recognition. The first step in the processing is the speech analysis system, which provides an appropriate (spectral) representation of the characteristics of time-varying speech signal. The most common techniques of spectral analysis are the class of filter bank methods and the class of linear predictive coding methods. The next step in the processing is the feature-detection stage. The idea here is to convert the spectral measurements to a set of features that describe the broad acoustic properties of the different phonetic units. Among the features proposed for recognition are nasality (presence or absence of nasal resonance), frication (presence or absence of random excitation in the speech), formant locations, voiced-unvoiced classification, and the ratios of high- and low frequency energy.

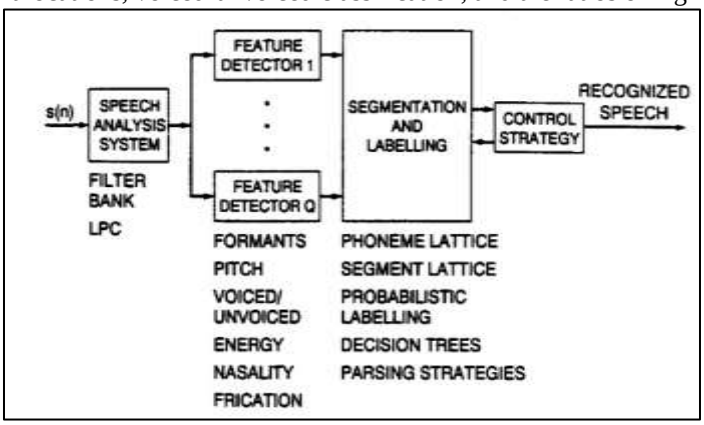

Fig. 2: Block Diagram of acoustic-phonetic speech recognition system

The third step in the procedure is the segmentation and labelling phase whereby the system tries to find stable regions and then the label the segmented region according to how well the features within that region match those of individual phonetic units. This stage is the heart of acoustic-phonetic recognizer and is the most difficult to carry out reliably; hence various control

strategies are used to limit the range of segmentation points and label probabilities. Issues in Acoustic Phonetic Approach: - Many problems are associated with the acoustic-phonetic approach to Speech Recognition. These problems, in many ways, account for the lack of success in practical speech recognition systems. Among these are the following:-

- The method requires extensive knowledge of the acoustic properties of phonetic units is assumed a priori in the acoustic-phonetic approach.
- The choice of features is made mostly based on ad hoc consideration. For most systems the choice of features is based on intuition and is not optimal inn a well-defined and meaningful sense.
- The design of sound classifiers is also not optimal. Ad hoc methods are generally used to construct binary decision trees.
- No well-defined automatic procedure exists for tuning the method on real, labelled speech. In fact, there is not even an idea way of labelling the training speech in a manner consistent and agreed on uniformly by a wide class of linguistics experts.

B. A Statistical pattern Recognition Approach

The pattern recognition paradigm has four steps, namely:-

- Feature measurement, in which a sequence of measurements is made on the input signal to define the "test pattern". For speech signals the feature measurements are usually the output of some type of spectral analysis technique such as filer bank analyser, LPC analysis, or DFT.
- Pattern Training, in which one or more test patterns corresponding to speech sounds of the same class are used to create a pattern representative of the features of that class. The resulting pattern, generally called reference pattern, can be an exemplar or template, derived from some type of averaging technique, or it can be a model that characterizes the statistics of the features of the reference pattern.
- Pattern classification, in which unknown test pattern is compared with each sound class reference pattern and measure is of similarity between the test pattern and each pattern is computed.
- Decision logic, in which the reference pattern similarity scores are used to decide which reference pattern or possibly sequence of reference patterns best matches the unknown test pattern.

The General Strengths and Weaknesses of Pattern Recognition Model: -

- The performance of the system is sensitive to the amount of training data available for creating sound class reference patterns; generally the more training, the higher the performance of system for virtually any task.

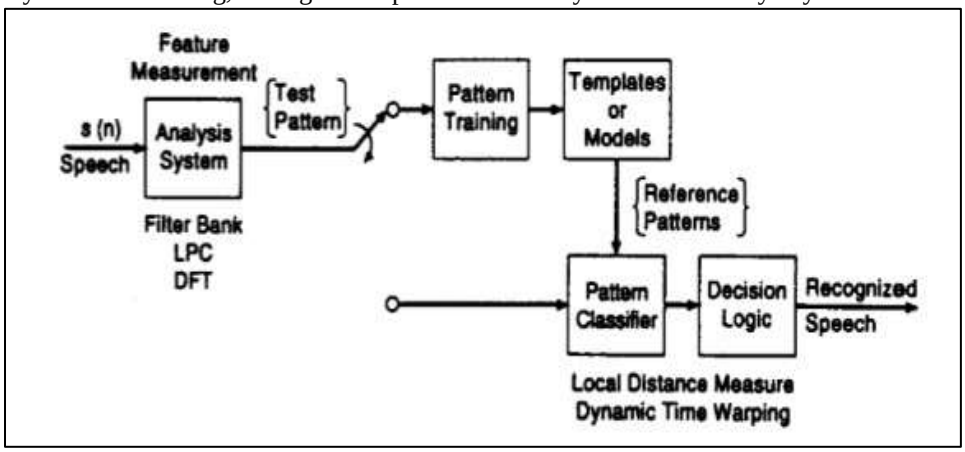

Fig. 3: Block Diagram of pattern-recognition speech recognizer

- The reference patterns are sensitive to speaking environment and transmission characteristics of the medium used to create the speech.
- No speech-specific knowledge is used explicitly in the system; hence system is insensitive to choice of vocabulary words, task, syntax, and task semantics.
- Because the system is insensitive to sound class, the basic techniques are applicable to a wide range of speech sounds, including phrases, whole words, and sub word units.
- It is relatively straight forward to incorporate syntactic constraints directly into the pattern recognition structure, thereby improving recognition accuracy and reducing computation.

IV. FEATURE EXTRACTION TECHNIQUES

This section covers various feature extraction techniques used in ASR. Feature extraction is the computation of sequence of feature vectors which provides compact representation of given speech signal. Usually it is performed in three stages. First stage is called speech analysis, which performs spectra-temporal analysis of speech signal and generates raw features describing the envelope of power spectrum of short speech intervals. Second stage compiles extended feature vectors composed of static and

dynamic features. Finally, the last stage transforms these extended feature vectors into more compact and robust form that are then supplied to the recognizer. Widely used speech features for auditory modelling are cepstral coefficients, which are obtained through Linear Predictive Coding (LPC). Other well-known speech feature extraction is based on Mel-Frequency Cepstral Coefficients (MFCC). Methods based on perceptual prediction which is good under noisy conditions are PLP and PLP-RASTA. There are some other speech extraction methods like RFCC, LSP etc.

V. CLASSIFICATION OF ASR

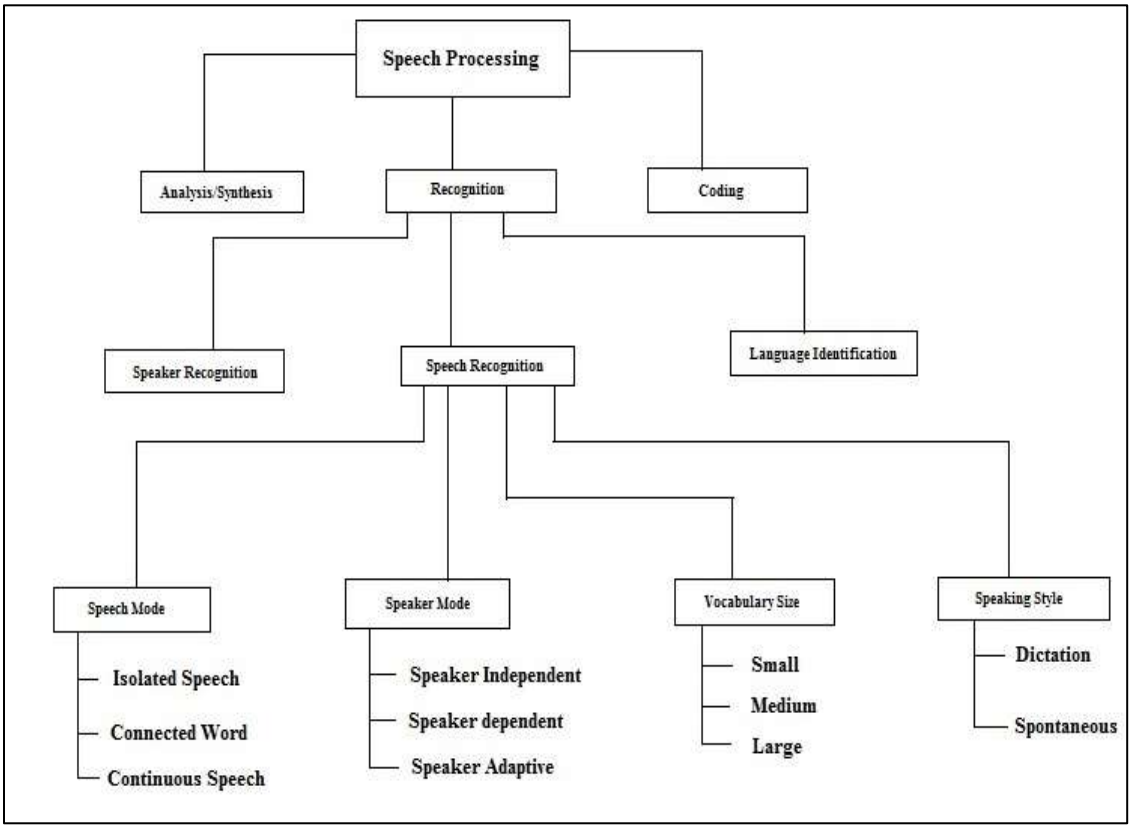

Fig. 4: Classification of ASR

VI. ACOUSTIC MODEL

An acoustic model is a file that contains statistical representations of each of the distinct sounds that makes up a word. Each of these statistical representations is assigned a label called a phoneme. The English language has about 40 distinct sounds that are useful for speech recognition, and thus we have 40 different phonemes. An acoustic model is created by taking a large database of speech (called a speech corpus) and using special training algorithms to create statistical representations for each phoneme in a language. These statistical representations are called Hidden Markov Models ("HMM"s). Each phoneme has its own HMM. [5]

For example, if the system is set up with a simple grammar file to recognize the word "house" (whose phonemes are: "hh aw s"), here are the (simplified) steps that the speech recognition engine might take: - The speech decoder listens for the distinct sounds spoken by a user and then looks for a matching HMM in the Acoustic Model. In our example, each of the phonemes in the word house has its own HMM. When it finds the matching HMM in the acoustic model, the decoder takes the note of the phoneme. The decoder keeps track of the matching phonemes until it reaches a pause in the user's speech. When pause is reached, the decoder looks up the matching series of phonemes it heard in its pronunciation dictionary to determine which word was spoken.

VII. LANGUAGE MODEL

A statistical language model is a probability distribution over sequences of words. Given such a sequence it assigns probability to the whole sequence. Having a way to estimate the relative likelihood of different phrases is useful in many NLP applications, especially one that generates text as output. Language model is used in speech recognition, machine translation, parsing, handwriting recognition and many other applications. In ASR, computer tries to match sounds with word sequences. The language model provides context to distinguish between words and phrases that sound similar. Language models are used in

information retrieval in the query likelihood model. Here a separate language model is associated with each document in a collection. Data sparsity is the major problem in building language models.

VIII. DYNAMIC TIME WARPING (DTW) ALGORITHM

Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed. For instance, similarities in walking patterns would be detected, even if in one video, the person was walking slowly and if in another, he or she were walking more quickly, or even if there were accelerations and decelerations during the course of one observation. DTW has been applied to video, audio, and graphics indeed, any data which can be turned into a linear representation can be analysed with DTW. [1][2][3]

In general, DTW is a method that allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions. The sequences are "warped" nonlinearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This sequence alignment method is often used in the context of hidden Markov models. Dynamic time warping (DTW) is such a typical approach for a template based approach matching for speech recognition and also DTW stretches and compresses various sections of utterance so as to find alignment that results in best possible match between template and utterance on frame by frame basis. By "frame" we mean short segment (10-30ms) of speech signal which is basis of parameter vector computation, and "match" defined as sum of frame-by frame distances between template and input utterance. Template with closest match defined in manner chosen as recognized word

IX. MACTHING TECHNIQUES

Speech-recognition engines match a detected word to a known word using one of the following techniques (Svendsen et al., 1989) (I) Whole-word matching: The engine compares the incoming digital-audio signal against a pre-recorded template of the word. This technique takes much less processing than sub-word matching, but it requires that the user (or someone) prerecord every word that will be recognized - sometimes several hundred thousand words. Whole-word templates also require large amounts of storage (between 50 and 512 bytes per word) and are practical only if the recognition vocabulary is known when the application is developed. (II) Sub-word matching: The engine looks for sub-words – usually phonemes and then performs further pattern recognition on those. This technique takes more processing than whole-word matching, but it requires much less storage (between 5 and 20 bytes per word).

X. SIMULATION RESULTS

This section includes simulation results that were achieved based on several algorithms mentioned in some of the referred papers. Here X-axis refers to samples while y-axis refers to amplitude.

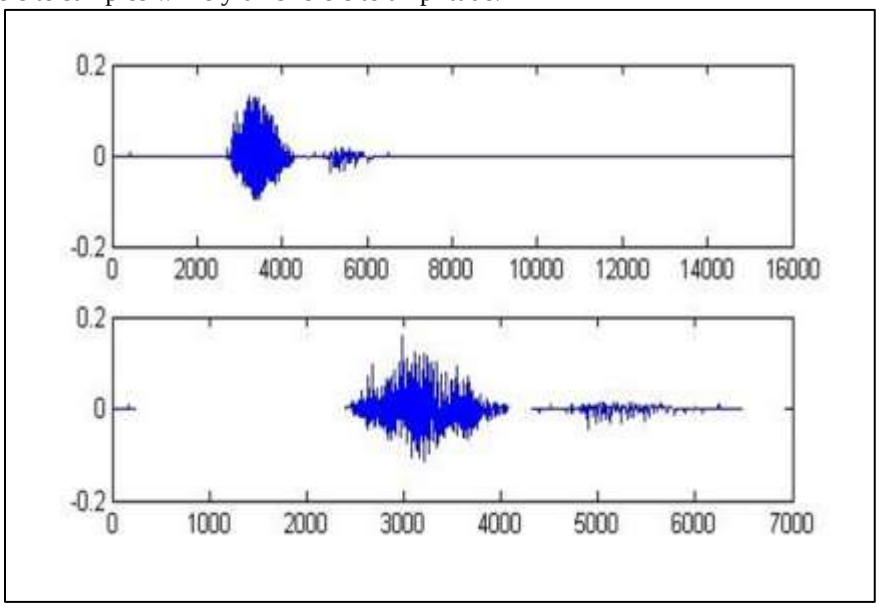

Fig. 5: End point detection of letter “A”

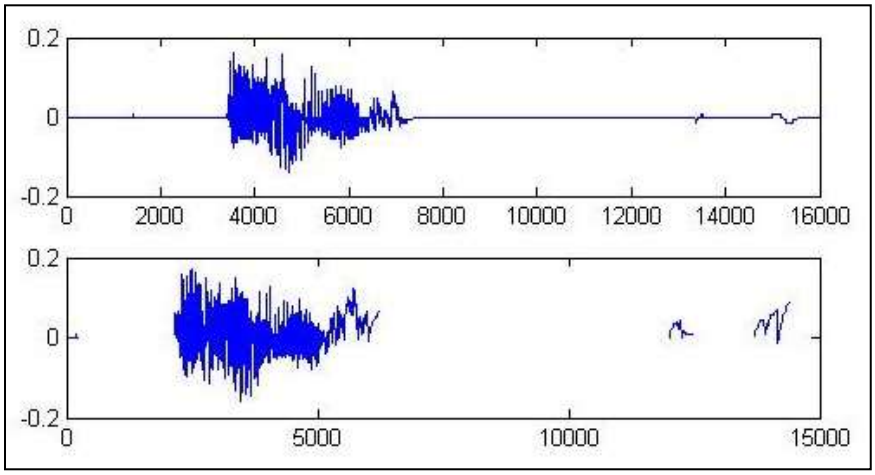

Fig. 6: End point detection of letter "B"

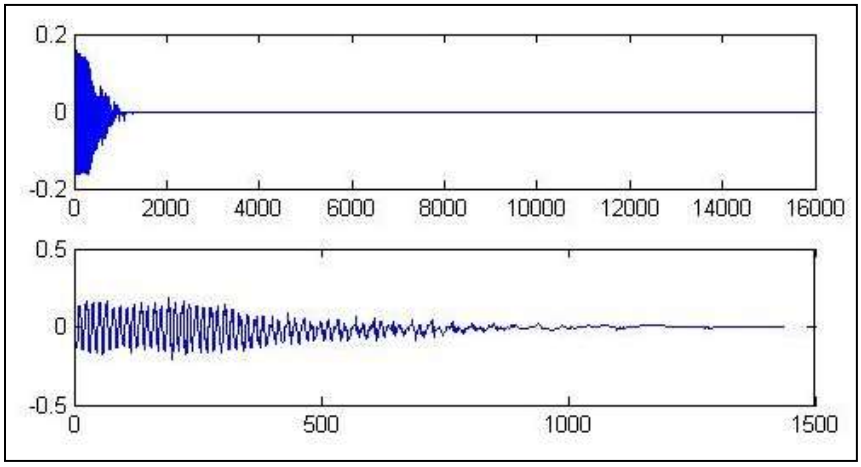

Fig. 7: Zero-crossing detection of letter "A"

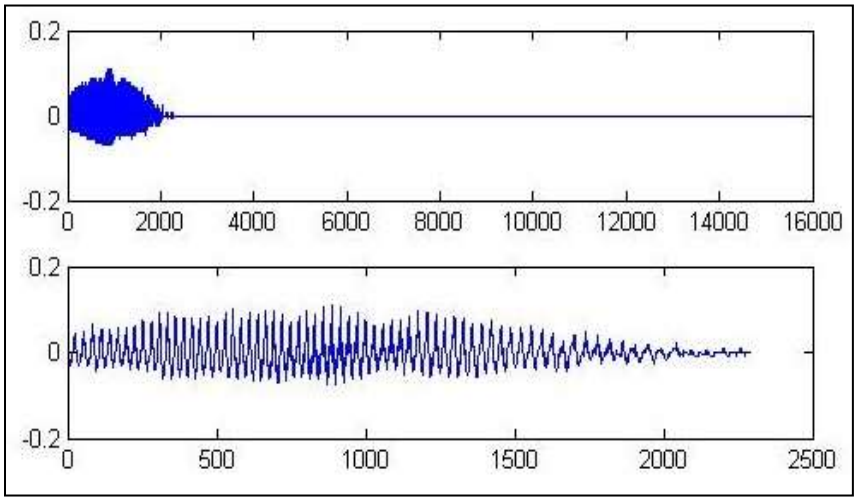

Fig. 8: Zero-crossing detection of letter "B"

REFERENCES

[1] S.K. Gaikwad et al.," A Review on Speech Recognition Technique," International Journal of Computer Applications, ISSN:2324-7569

[2] A. Chaudhari & R.Kshirsagar, "Process Speech Recognition System using AI technique", International Journal of Soft Computing & Engineering, ISSN:2231-2307,Vol-2,Issue-5, November 2012

[3] M. G. Gohil & S. J. Varmora," Artificial Intelligence for Speech Recognition", International Multidisciplinary Research Journal, ISSN: 2349-7637 (Online),Vol-1, Issue-2, September 2014

[4] S. P. Shinde & V. P. Dehmukh, "Implement "Implementation of Pattern Recognition techniques & overview of its applications in various areas of AI", International Journal of Advances in Engineering & Technology, ISSN : 2231-1963,Vol-1,Issue-4,pp. 127-137,September 2011

[5] G. H. Vardhan & G. H. Charan," Artificial Intelligence & its Applications for Speech Recognition", International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064,Vol-3, Issue-8, August 2014
[6] L. R. Rabinar and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances", The Dell System Technical Journal, Vol. 54. No. 2, February 1975
[7] R.G Bachu et.al, "Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal", Bulletin Of The Polish Academy Of Sciences, Vol. 62, No. 3, 2014
[8] Giuseppe Riccardi et.al, "Active Learning: Theory and Applications to Automatic Speech Recognition", IEEE Transactions On Speech And Audio Processing, Vol. 13, No. 4, July 2005
[9] R. Prashantkumar et.al, "Two Wheeler Vehicle Security System", International Journal of Engineering Sciences & Emerging Technologies, Volume 6, Issue 3, December 2013, pp. 324-334
[10] P. K. Kurzekar et.al, "A Comparative Study of Feature Extraction Techniques for Speech Recognition System", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, Issue 12, December 2014