International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

Research Article

A Systematic Analysis of Automatic Speech Recognition: An Overview

Taabish Gulzar[Ȧ*], Anand Singh[Ȧ] , Dinesh Kumar Rajoriya[Ḃ] and Najma Farooq[Ȧ]

[Ȧ]Department of Electronics and Communication, Dehradun Institute of Technology, Mussourie Diversion Road, Makkwala Dehradun, India
[Ḃ]Department of Electronics and Communication, Sagar Institute of Science, Technology and Engineering, Bhopal, M.P, India

Accepted 18 May 2014, Available online 01 June2014, **Vol.4, No.3 (June 2014)**

Abstract

Most high-flying and primary means of communication among humans is speech. Despite the researches and developments in the field of automatic speech recognition the accuracy of the said is still a research challenge. This paper reviews past work comparing modern speech recognition systems and humans to determine how far recent dramatic progress in technology has evolved towards the objective of human-like performance. An overview of sources of knowledge is introduced and the use of knowledge to create and verify hypotheses is discussed.

Keywords: *Automatic speech recognition, Feature Extraction, Utterance, Dynamic time wrapping, Matching.*

1. Introduction

From previous several decades human beings tried to create technologies that could recognize correct speech. While humans can differentiate speech very easily, they in fact make use of much acoustic, linguistic and contextual information. It has been seen that relation between physical speech signal and the corresponding words is so much complex and very hard to understand. Both the research areas of automatic speech recognition (ASR) and human speech recognition (HSR) observe the recognition process from the acoustic signal to a series of recognized units. For ASR, the objective is to automatically transcribe the speech signal in terms of a sequence of items as close as possible to a reference transcription (L. Rabiner et al, 1993; F. Jelinek, 1997). In HSR, the attention is on understanding how human listeners recognise spoken utterances. On the basis of advances in statistical modelling of speech, automatic speech recognition (ASR) systems find extensive application in tasks that make use of human-machine interface, such as automatic call processing in telephone networks and query-based information systems that provide updated travel information, stock price quotations, weather reports, embedded systems etc.

1.1 Definition and Basic Model of speech recognition

Speech Recognition also known as Automatic speech recognition (ASR) is defined as a process of converting a speech signal into a set of words by a certain algorithm that can be implemented as a system program or a process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words (V. Zue et al, 1996; Z. Mengjie, 2001). Automatic speech recognition (ASR) is one of the fastest growing areas in the framework of speech science and engineering. Research in speech processing and communication for the most part, was enthused by people's desire to build mechanical models to follow human verbal communication capabilities. The primary aim of ASR systems is to develop the new techniques and systems for speech input to machines. Mathematical representation of speech recognition system in straightforward equations which contain frontend unit, model unit, language model unit, and search unit is shown in Fig. 1.

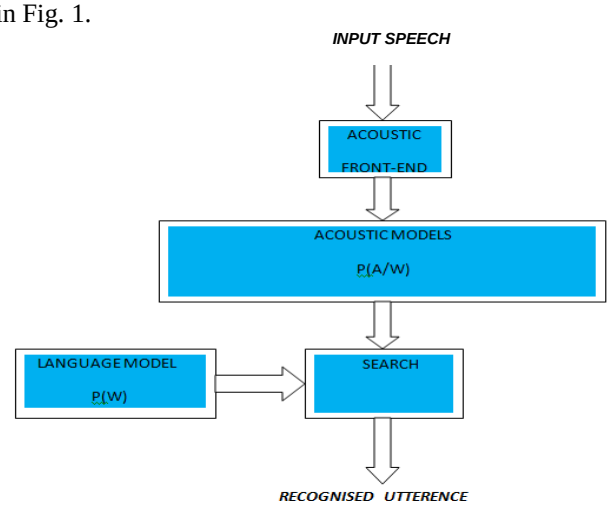

Fig 1 shows the basic model of speech recognition.

One of standard approach to large vocabulary continuous speech recognition is to presume a simple probabilistic model of speech production whereby a specified word set, W, generates an acoustic observation sequence Y, with probability P(W,Y). The objective is then to decode the

*Corresponding author: **Taabish Gulzar**

word string, based on the acoustic observation sequence; so that the decoded string has the maximum a posteriori (MAP) probability.

$$P(W/A) = \arg\max_w WP(W/A) \quad (1)$$

Using Baye s rule, equation (1) can be written as

$$P(W/A) = P(A/W) \times P(W)/P(A) \quad (2)$$

Since P(A) is independent of W, the MAP decoding rule of equation(1) is

$$W = \arg\max_w P(A/W) \times P(W) \quad (3)$$

The first term in equation (3) P(A/W), is generally called the acoustic model, as it estimates the probability of a sequence of acoustic observations, conditioned on the word string. Hence P(A/W) is calculated. For large vocabulary speech recognition systems, it is essential to develop statistical models for sub word speech units, developed word models from these sub word speech units, (using a lexicon to describe the composition of words), and then postulate word sequences and evaluate the acoustic model probabilities via standard concatenation methods. The second term in equation (3) P(W), is called the language model. It illustrates the probability related with a postulated sequence of words. Such language models can includes both syntactic and semantic constraints of the language and the recognition task.

2. Literature Survey

Machine recognition came into existence in early 1920's. The first machine (Radio Rex) that was being used for speech recognition was manufactured in 1920 (Aravind Ganapathiraju et al. 2001).In 1950's, most of the speech recognition systems investigated spectral resonances during the vowel portion of each speech which were extracted from output signals of an analogue filter bank and logic circuits (Sadaoki Furui, 2005). Davis, Biddulph, and Balashek developed a system for a single speaker to recognise the isolated digits that rely heavily on computing spectral resonances during the vowel region of each digit (K. H. Davis et al, 1952). In this period another effort was vowel recognizer that was constructed by Forgie and Forgie at MIT Licoln laboratories in 1959,in which 10 vowels embedded in a /b/-vowel/t/ format were recognized in a speaker independent manner (J. W. Forgie et al, 1959). The spectral information was provided using a filter bank analyser and a time varying approximation of the vocal tract resonances was made to choose which vowel was spoken. The major difficulty in speech recognition was in the nonuniformity of time scales in speech events.

In early 1960s, a set of elementary time-normalization methods, that was based on the capability to accurately detect speech starts and ends, that considerably reduced the variability of the recognition scores was developed by Martin and his colleagues at RCA Laboratories (T. B. Martin et al, 1964). Meanwhile, in the Soviet Union, the use of dynamic programming methods for time aligning a pair of speech utterances (generally known as dynamic time warping (DTW)), including algorithms for connected word recognition was put forth by (T. K. Vintsyuk, 1968). In 1960's various special purpose hardware's were developed because the computers were still not fast enough (J. Suzuki et al, 1961). Japanese system which was described by Suzuki and Nakata of the Radio Research Lab in Tokyo was a hardware vowel recognizer (T. Sakai et al, 1962). A sophisticated filter bank spectrum analyzer along with logic that links the outputs of each channel of the spectrum analyzer (in a weighted manner) to a vowel decision circuit came into existence, and majority decisions logic scheme was used to choose the spoken vowel. An achievement in the 1960s was the revolutionary research of Reddy in the area of continuous speech recognition by dynamic tracking of phonemes (D. R. Reddy, 1966). Eventually this research remains a highly successful speech recognition research program at Carnegie Mellon University and is still a world leader in continuous speech recognition systems.

In 1970's research by Velichko and Zagoruyko in Russia (V. M. Velichko et al, 1970), Cakoe and Chiba in Japan (H. Sakoe et al, 1978), and Itakura in the united States in the field of speech recognition of isolated word or discrete utterance recognition became a viable and practical technology. The advance use of pattern recognition ideas in speech recognition was put forth by Russian studies, Japanese researchers showed the application of dynamic programming methods. Itakura's work presents how the ideas of linear predictive coding (LPC) could be extended to speech recognition systems by means of an appropriate distance measure based on LPC spectral parameters (F. Itakura, 1975). At IBM labs, Large vocabulary speech recognition was investigated for three distinct tasks, namely the New Raleigh language for simple database queries (C. C. Tappert et al, 1971), the laser patent text language for transcribing laser patents (F. Jelinek et al, 1975), and the office correspondence task, called Tangora, for dictation of simple memos.

The problem of connected word speech recognition was the main focus of research in 1980's. The primary aim was to develop a robust system capable of recognizing a fluently spoken thread of words (eg., digits) based on matching a concatenated pattern of individual words. Moshey J. Lasry , first to develop a feature based speech recognition system in the beginning of 1980.His research consists of speech spectrograms of letters and digits (R. K. Moore,1994). A broad range of the algorithm based on matching a concatenated pattern of individual words were formulated and implemented, including the two level dynamic programming approach of Sakoe at Nippon Electric Corporation (NEC) (H. Sakoe, 1979), the one pass method of Bridle and Brown at Joint Speech Research Unit (JSRU) in UK (J. S. Bridle et al, 1979), the level building approach of Myers and Rabiner at Bell Labs (C. S. Myers et al, 1981), and the frame synchronous level building approach of Lee and Rabiner at Bell Labs (C. H. Lee et al, 1989). In some of the laboratories (Primarily IBM, Institute for Defense Analyses (IDA), and Dargon systems) the methodology of hidden Markov modelling

(HMM) was quite understood. In late 1970s, a method for speech recognition that make use of the combination of instantaneous cepstral coefficients and their first and second order polynomial coefficients, now called Δ and ΔΔcepstral coefficients, as fundamental spectral features for speech recognition (S. Furui, 1986). In the 1980s, the idea of applying neural networks to speech recognition was reintroduced. Neural networks were first introduced in the 1950s, but they suffer various difficulties because of practical problems. In the 1980s, a deeper consideration of the strengths and limitations of the technology was attained, as well as an understanding of the relationship of this technology to classical pattern classification methods (S. Katagiri, 2003; R. P. Lippmann, 1987; A. Weibel et al, 1989). Presently most of the speech recognition systems are based on the statistical work developed in 1980, and a significant extension have been made in 1990's. A number of innovations took place in the field of speech recognition in 1990s. The challenging problem of pattern recognition was transformed into an optimization problem relating minimization of the empirical recognition error (B. H. Juang et al, 2000). This basic paradigmatic change was caused by the recognition of the fact that the distribution functions for the speech signal could not be accurately chosen or defined, and that Bayes' decision theory becomes inapplicable under these circumstances. Essentially, the objective of a recognizer design should be to attain the least recognition error rather than provide the best fitting of a distribution function to the given (known) data set as advocated by the Bayes criterion. This error minimization concept produced a number of techniques, such as discriminative training and kernel-based methods. Many different techniques were investigated to enhance the robustness of speech recognition systems that are caused due to mismatching between training and testing conditions, because of the background noises, voice individuality, microphones, transmission channels, room reverberation, etc. The key techniques include the maximum likelihood linear regression (MLLR) (C. J. Leggetter et al, 1995), the model decomposition (A. P. Varga et al, 1990), parallel model composition (PMC) (M. J. F. Gales et all, 1993), and the structural maximum a posteriori (SMAP) method (K. Shinoda et al, 2001).

2.1 2000 onwards

In 21th century, a variational Bayesian (VB) estimation and clustering techniques were developed (Mohamed Afify et al, 2004). The former is based on a subsequent distribution of parameters. In speech recognition the problem of adaptive learning was solved by Giuseppe Richardi (Mohamed Afify et al, 2005) and also proposed active learning algorithm for ASR. In 2007, the difference in acoustic features between spontaneous and read speech using a large scale speech data base i.e, CSJ have been analyzed (Simon Kinga et al, 2006). The entire cepstral analysis scheme is implemented in hardware by employing intellectual property cores, so as to accelerate recognition process. Hence field programmable gate array (FPGA) based speech recognition systems was being introduced. First softcore implementations on Virtex-4 family FPGA of Lithuanian isolated word recognizer were made by this article authors in (V. Arminas et al, 2010). Use of this soft-core processor Microblaze jointly with intellectual property (IP) cores for signal processing enabled us to boost up word recognition process by 1.55 times (G. Tamulevičius et al, 2010; E. Ivanovas, 2012), but it was still not enough for a real-time operation. The field programmable gate array (FPGA) platform makes use of two techniques i.e. parallelization and pipelining technique in speech recognition. It is a flexible architecture to create systems in contrast to implementations on the ASIC devices. The embedded processor-based solutions on the market have an average 80 % recognition rate and limited size of the dictionary: 32 (EasyVR (A. Chakravarty, 2013) or 75 (NLP [40]) commands. The major concern of such recognizer is to make sure real-time requirements for the feature extraction (especially in comparison stages). Feature extraction and matching techniques in FPGA (J. Choi et al, 2010; R. Veitch et al, 2010; G. Zhang et al, 2011; S. T. Pan et al , 2011;) or GPU (D. Sart et al, 2010; Y. Zhang et al, 2012) implementation can run independently in contrast with this approach.

Work performed by Davel and colleagues (M. Davel et al, 2004; M. Davel et al, 2009) has shown that it is possible for 'developers' with limited linguistic experience to create accurate models' through the use of a developmental process with proper tool support. In March 2008, the very first multi-model speech application for Google maps for mobile (GMM) was being introduced. Most of today's speech recognition systems employ hidden Markov models (HMMs) to deal with the temporal variability of speech and Gaussian mixture models (GMMs) to decide how well each state of each HMM fits a frame or a short window of frames of coefficients that represents the acoustic input. An optional way to evaluate the fit is to make use of a feed-forward neural network that takes several frames of coefficients as input and generates posterior probabilities over HMM states as output. Deep neural networks (DNNs) that have multiple hidden layers and are trained using new methods have been shown to smash GMMs on a variety of speech recognition benchmarks, sometimes by a large margin (Heiga Zen et al, 2013). In order to achieve the best recognition accuracy, high-tech deep neural networks are adopted as acoustic models. With significantly smaller footprints, DNN can provide large accuracy improvements over GMM acoustic models (Xin Lei et al, 2013). Previously, speech recognition on handheld computers and smartphones was investigated in DARPA sponsored Transtac Program, where speech-to-speech translation systems were developed on the phone (J. Zheng et al., 2010; J. Xue et al; 2012; R. Prasad et al, 2013) Accessing information through our voice has been a part of science fiction. Currently, with the help of power-full smartphones and cloud based computing, science fiction is becoming reality. Today's state-of-the –art in speech processing is upcoming field of Google search by voice. The ability to carry out web search simply by speaking (voice search) them to your mobile phones, first appeared on iphone and android phones (Johan schalkwyk, 2013). At Google Japanese voice search led to a major simplification and

speed up the process to deal with new languages in general. Japanese and Koreans are in terms of usage currently (2009, 2011) our one of the largest international languages. Voice search for Mandarin Chinese and Cantonese that was launched by Google in 2009 and 2010 are described in (Jiulong Shan et al, 2010) and (Yun-Hsuan et al, 2011). In September 2013 some of the graduate students and researchers at UT Dallas have developed novel systems that can recognize speaking voices despite conditions that can make it harder to make out a voice, such as whispering, speaking through various emotions, or talking with a stuffy nose. In a recent National Institute of Standards and Technology Speaker Recognition Evaluation challenge, officials sent more than 80 million voice verification trials with added noise – natural background sounds or artificial computer-generated sounds – to more than 50 universities, research labs and companies throughout the world. Teams had to determine whether speech recordings were from certain speakers or not. Recently in 2014, an application namely JUSTSPEAK a universal voice control solution for non-visual access to the android operating system came into existence. As compared to already existing systems JUSTSPEAK offers two contributions. First, it allows system wide voice control on Android that contains any application. Second, it provides more resourceful and natural interface with support of multiple voice commands in the same utterance (Yu Zhong et al, 2014). Google ASR services are being used in the framework of JUSTSPEAK in order to recognize the spoken speech of users (Lei Xin et al; 2013).

3. Speech Recognition Categories

Speech recognition systems can be classified into a number of different classes by describing what types of utterances they have the capability to recognize

3.1 Isolated word speech recognition

Each utterance in an Isolated word recognizer usually needs to be quiet on both sides of the sample window. At a time a single word or a single utterance is accepted. These systems have Listen or Not-Listen states, where a speaker has to wait between the utterances. Isolated utterance may be the best name for this class of words (Zahi N.Karam et al). Isolated word recognition is based on the principle that the signal in a given recording interval consists of an isolated word, preceded and followed by silence or other background noise. Hence, it is supposed that the speech segments can be reliably detached from the silence segments (lori f et al, 1981) An isolated word speech recognition system is considered as a system, where a single utterance is treated in accordance with two implicit assumptions. First assumption involves that the speech to be recognized should consists a single word/phrase and the said is to be recognised as a complete entity without any explicit knowledge for the phonetic content of the word/phrase. The second assumption is that each and every spoken utterance should have accurately defined endpoints i.e. beginning and ending point. One of the applications where this type of speech recognition can be used is Command and Control systems.

3.2 Connected word recognition

Connected words (utterances) are somehow similar to isolated utterances, but allow separate utterances to be 'run-together' with a certain gap of several milliseconds in between them called speech code (Anand Singh et al, 2012). The words that are separated with a small gap (speech code) in between them are known as connected words and system used for these kinds of words is known as connected word recognition system. Resembling isolated word speech recognition, connected word recognition has a property that the basic speech-recognition unit is the word/phrase to much extent. (Rabiner et al. 2010) investigated three algorithms that were designed for connected word recognition: Two level DP approach, Level Building approach and One Pass approach are three algorithms and found them to be providing the l best identical matching string with the identical matching score for connected word recognition. Computational efficiency, storage requirement and ease of realisation in real time hardware were the factors by which they differ. Later (Garg et al., 2011) developed speaker dependent connected digits recognition system by relating unrestricted Dynamic time warping technique in which each digit was recognised by estimating distance with respect to matching of input spoken digit with stored template.

3.2 Continuous speech recognition

In continuous speech recognizers computers determines the content while as humans are allowed to speak naturally. These types of recognizers are very difficult to develop as they make use of special methods to determine the endpoints of the utterances. In Continuous speech recognition words are connected together instead of being separated by pauses (speech codes). As a result the performance of continuous speech recognition systems is affected in terms of unknown boundary information (endpoints) about words, co-articulation, production of surrounding phonemes and rate of speech.

3.3 Spontaneous speech recognition

Spontaneous speech can be considered as speech that is natural sounding but not rehearsed. Spontaneous speech recognizer system should be able to handle a variety of natural speech features such as words being run together, ums and ahs, and even slight stutters.

4. Automatic Speech Recognition Process

The main aim for speech recognition is that a machine should be able to hear (speech recognition), to understand (Natural language processing) and response back (speech synthesis) to spoken information (Taabish Gulzar et al; 2014) and (Rajouriya D. K et al, 2010). In ASR systems usually the aim is to analyse, extract and classify or

recognize the content of information spoken by humans. Speaker recognition system works in four stages.

1: Analysis of speech.
2: Feature extraction.
3: Modeling.
4: Testing.

4.1 Speech Analysis Techniques

When an utterance is being spoken speaker characteristics (e.g. vocal tract length, shape and gender), linguistic content, acoustic surroundings and speaking rate simultaneously influences the acoustics of the whole spoken productivity (Hisashi Wakita, 1977). The analysis of speech signal is done with the below mentioned techniques.

4.1.1 Segmentation analysis

Here the speech signal is being analyzed using the size of frame and a shift in the range of 10-30 ms to pull out the speaker information.

4.1.2 Sub Segmental analysis

This analysis is mainly used to extract the characteristics of excitation state (Nicolás Morales1 et al.).

4.1.3 Supra Segmental analysis

In this case, behavior character of the speaker is analyzed through the analysis of frame size.

4.2 Feature extraction techniques

The primary aim of feature extraction is to find a set of properties of an utterance that have acoustic correlations to the speech-signal, that is parameters that in some way can be computed or estimated through processing of the signal waveform. Such parameters are termed as features. In order to obtain accuracy in speech recognition, the main issue is selection of the appropriate features from a speech signal. Two types of analysis i.e. Temporal and Spectral analysis techniques are used for feature extraction. Speech signal itself is used for analysis informer case, while as in later mainly the spectral representation of a speech signal is carried out for analysis. Feature extraction means tracking out the needful information content out of a speech signal and discarding the unwanted information. During feature extraction the dimensions of the input vector is reduced while the discriminating power of the speech signal is maintained.

Feature extraction can be divided into three steps. Spectral analysis: Here raw features are generated that explains the envelope of the power spectrum of speech segments. The second step is called parametric transformation, where the feature vector that comprises of some static and dynamic features are compiled. Finally, the last stage known as statistical modelling is used to transform the feature vectors into more robust and dense vectors that are given to Back-end processor. Some of the Extraction methods along with their properties and respective method for implementations are given in the table below.

Table 1 shows different Feature Extraction methods along with their properties and implementation procedure.

S.No	Methods	Properties	Procedure for implementation
1	Principal Component analysis (PCA)	Non linear feature extraction method, Linear map, fast, eigenvector-based	Traditional, eigenvector base method, also known as karhuneu-Loeve expansion; good for Gaussian data
2	Linear Discriminate Analysis(LDA)	Non linear feature extraction method, Supervised linear map; fast, eigenvector-based	Better than PCA for classification (M.A.Anusuya et al, 2009).
3	Independent Component Analysis (ICA)	Non linear feature extraction method, Linear map, iterative non-Gaussian	Blind course separation, used for de-mixing non- Gaussian distributed sources(features)
4	Linear Predictive coding	Static feature extraction method,10 to 16 lower order coefficient	It is used for feature Extraction at lower Order
5	Cepstral Analysis	Static feature extraction method, Power spectrum	Used to represent spectral envelope (M.A.Anusuya et al, 2009).
6	Mel-frequency scale analysis	Static feature extraction method, Spectral analysis	Spectral analysis is done with a fixed resolution along a Subjective frequency scale i.e. Mel-frequency Scale
7	Filter bank analysis	Filters tuned required frequencies	
8	Mel-frequency cepstrum (MFFCs)	Power spectrum is computed by performing Fourier Analysis	This method is used for find our Features
9	Kernel based feature extraction method	Non linear transformations	Dimensionality reduction leads to better classification and it is used to redundant features, and improvement in classification error (Kenneth et al, 2003).
10	Wavelet	Better time resolution than Fourier Transform	It replaces the fixed bandwidth of Fourier transform with one proportional to frequency which allow better time resolution at high

			frequencies than Fourier Transform
11	Dynamic feature extractions i)LPC ii)MFCCs	Acceleration and delta coefficients i.e. II and III order derivatives of normal LPC and MFCCs coefficients	It is used by dynamic or runtime Feature
12	Spectral subtraction	Robust Feature extraction method	It is used basis on Spectrogram (W. M. Campbell et al, 2006).
13	Cepstral mean subtraction	Robust Feature extraction	It is same as MFCC but working on Mean statically parameter
14	RASTA filtering	For Noisy speech	It is find out Feature in Noisy data
15	Integrated Phoneme subspace method (Compound Method)	Integrated Phoneme subspace method (Compound Method)	Higher Accuracy than the existing Methods (Sannella).

4.3 Modeling Techniques

Generation of the speaker models making use of specific feature vector is the main aim of modelling technique. Speaker modelling techniques are categorizes into two types speaker recognition and speaker identification. The later automatically decides who is speaking on the basis of individual information integrated in an utterance, while the speaker recognition is in turn divided into two types speaker dependent and speaker independent. In case of speaker dependent mode the recognizer extracts all the speaker characteristics in the acoustic signal (Samudravijay). In speaker independent case no such features are taken into consideration and only the needful message is extracted. A few modeling techniques that are being used in speech recognition systems are mentioned below.

4.3.1 Acoustic phonetic approach

The establishment of acoustic phonetic approach is based on finding speech sounds and providing appropriate labels to these sounds, that presume that there exists a limited phonemes in spoken language and a set of acoustic properties are used for the classification of these phonemes. The very first step of the above mentioned approach is spectrally analyze the speech signal together with feature detection that transforms the spectral measurements to a string of features explaining the broad acoustic properties of various phonemes. In the second step known as segmentation and labelling the utterance is firstly segmented stable acoustic regions and then one more phonetic label are connected to each and every segmented part. In the last and the final step a valid word or a set of words are determined from the phonetic sequence that is produced by the segmentation to labelling. Problem phone recognition, Gaussian Mixture modelling and Support vector machine classification (E. Singer et al; Viet Bac Le at al.) are the three main techniques that have been widely used for language identification, still many of these applications refuses the use of acoustic phonetic approach (D.R.reddy, 1996).

4.3.2 Pattern Recognition

In the recent years, pattern recognition gained immense popularity and has been widely used in a variety of applications. The idea of using pattern recognition strongly emerged from the desire to imitate human recognition behaviour. The key idea of pattern recognition is to optimally extract patterns based on certain conditions and thereby separate the data into different classes. Moreover, pattern recognition approach performs classification without having any idea about the distribution of the measurements in different groups.The pattern-matching method has become the predominant method for speech recognition during the last six decades (Dat Tat Tran, 2000). The vital feature of pattern recognition approach (Itakura 1975; Rabiner 1989; Rabiner and Juang 1993) is that it makes use of a well formulated mathematical framework and generates consistent speech pattern representations, for consistent pattern comparison, from a set of labelled training samples via a formal training algorithm. The pattern-comparison stage of this method involves a direct comparison between the unknown utterances i.e. the speech to be recognized with each possible blueprint learned in the training stage for determining the individuality of the unknown according to the best of match of the patterns.

4.3.3 Template based approaches

The main idea behind this approach is that, a set of ideal speech samples are stored as reference patterns representing the dictionary of candidate s words. Recognition is then done by matching the feature vectors of an unknown utterance (speech to be recognized) with all the data in the reference template in order to achieve the best match. Errors due to segmentation or classification of smaller acoustically more variable units such as phonemes can be avoided by creating templates for entire words. One key idea in template method is to derive typical sequences of speech frames for a pattern (a word) via some procedure, and to rely on the use of local spectral distance measures to compare patterns. One more idea is to use some form of dynamic programming for the rime being alignment of patterns to account for variations in speaking rates across talkers as well as across repetitions of the word by the same talker.

4.3.4 Stochastic approach

Stochastic modelling entails the use of probabilistic models to deal with uncertain or incomplete information. The major paradigm swing for speech-recognition progress has been the introduction of statistical methods, especially stochastic processing with hidden Markov models (HMMs) (J. K. Baker, 1975;] F. Jelinek, 1976) in the early 1970s (H. Poor, 1988). More than 30 years later,

this methodology still predominates. Hidden Markov modelling (HMM) is the most in use stochastic approach now a days. HMMs, which have long dominated the world of acoustic modelling, are used to characterize the speech signal as a parametric stochastic process and statistically characterize the variation of a speech unit (a word/phoneme/syllable). HMMs have the same advantages for speaker recognition as they do for speech recognition. Extraordinarily robust models of speech events can be developed with only small amounts of requirement or information accompanying training utterances. Speaker recognition systems based on an HMM architecture used speaker models derived from a multi-word sentence, a single word, or a phoneme. Usually, multi-word phrases (a string of seven to ten digits, for example) were used, and models for each individual word and for silence were combined at a sentence level according to a predefined sentence-level grammar (J. M. Naik, et. Al, 1989). A negative side effect of this is that HMMs do not provide much accuracy towards the recognition process. Hence, problems are often encountered while analysing the errors of an HMM system in an attempt to improve its performance. On the other hand, vigilant incorporation of knowledge has considerably enhanced HMM based systems.

4.3.5 Dynamic Time Warping (DTW)

DTW is a method which measures the distance between each input frame and each reference frame using the dynamic programming algorithm to find the finest warping of the pattern, and decides the best match by minimizing the distance between the input frame and the reference frame. Automatic speech recognition is a well known application of DTW, in order to deal with different speaking speeds. The time alignment of different speech signal is the centre problem for distance measurement in speech recognition. A slight shift leads to incorrect recognition. Dynamic Time Warping is an efficient method to solve the time alignment problem. DTW algorithm focuses at aligning two sequences of feature vectors by warping the time axis cyclically until an optimal match between the two sequences is found. Continuity plays a less role in DTW than in other pattern matching methods. DTW is an algorithm that is predominantly suitable to matching sequences with missing information, provided there are long enough segments for matching to occur.

4.3.6 Vector Quantization (VQ)

Vector quantization (VQ) is a classical quantization technique from signal processing which allows the modeling of probability density functions by the distribution of prototype vectors. It was basically used for data compression. It works by dividing a large set of points (vectors) into relatively small groups having approximately the same number of points closest to them. Each group is represented by its centroid point, as in k-means and some other clustering algorithms. The primary aim of data compression is to reduce the bit rate for transmission or data storage while maintaining the necessary reliability of the data. The feature vector may represents all possible speech coding parameters including linear predictive coding (LPC) coefficients, cepstrum coefficients (Balwant A. Sonkamble et al, 2012). ASR systems often make use of VQ (R. K. Moore, 1994). In order to use efficient codebooks for reference models and codebook searcher Vector Quantization is used. The test utterance is evaluated by all possible codebooks and ASR selects the word whose codebook gives the lowest distance measure (L.R.Bahl et.al, 1993). Since codebook entries are not arranged in particular order and can come from any part of the training words, thus the codebooks do not posses any specific time information. On the other hand, some indirect durational cues are preserved because the codebook entries are used for minimize average distance across all training frames, and frames, corresponding to longer acoustic segments (e.g., vowels) are more frequent in the training data. Frequently a few code words are sufficient to represent large number of frames during relatively steady sections of vowels, thus allowing even more codeword's to represent short, dynamic portions of the words. Hence as far as vocabularies of similar words are considered, vector quantization can be an advantage over other ASR comparison methods.

4.3.7 Knowledge Based Approach

A hybrid of the acoustic phonetic approach and pattern recognition approach results in Knowledge based approach (Artificial Intelligence approach) (R. K. Moore, 1994). All the information regarding linguistic, phonetic and Spectrogram is used by knowledge based approach. Knowledge engineering design makes use of the direct and explicit incorporation of experts speech knowledge into a recognition system. However, pure knowledge engineering was also encouraged by the significance and research in expert systems. Due to the difficulty in quantifying expert knowledge this approach has only limited success. Introduction of many levels of human knowledge such as phonetics, phonotactics, lexical access, syntax, semantics and pragmatics is one more serious issue.

4.3.8 Connectionist Approaches (Artificial Neural Networks)

Another technology that was (re)introduced in the late 1980's was the concept of artificial neural networks (ANN). Neural networks were first introduced in the 1950's , but was unsuccessful in producing notable results initially (W. S. McCullough et al, 1943). Compared to HMMs, neural networks make no suppositions about feature statistical properties and have numerous qualities making them eye-catching recognition models for speech recognition. Neural networks allow discriminative training in a natural and efficient manner when they are used to estimate the probabilities of a speech feature segment. Recent development in speech recognition is the connectionist modeling and is still the subject of

controversy. Knowledge or constraints are not programmed in individual units, rules, or procedures, but distributed across many simple computing units as far as connectionist models are concerned. Connectionist models mainly depend upon the availability of excellent training or learning strategies. Hardware implementation of the connectionist models because of their simplicity and uniformity in the processing elements becomes more attractive. However, a large number of iterations are required to train the data, and in some cases is more expensive.

4.3.9 Support Vector Machine (SVM)

One of the most promising and in use learning algorithms for classification as well as regression is Support vector machine (SVM). SVM has been effectively applied to many real-world pattern recognition applications. SVM finds a separating surface with a large margin between training samples of two classes in a high dimensional feature space implicitly introduced by a computationally efficient kernel mapping, and the large margin implies a better generalization ability according to the statistical learning theory (V. N. Vapnik, 1995). Support vector machine (SVM) is a linear machine that works in the highly dimensional feature space formed by the nonlinear mapping of the N-dimensional input vector 'x' into a K-dimensional feature space (K > N) through the use of a mapping u(x) (Sadoaki Furui, 1991). The algorithm was originally developed by Shikano (1985), for the classification problem with separable data, and this method was later improved to handle non-separable data as well. The typical fact about SVM is that the learning task is reduced to quadratic programming by using the so-called Lagrange multipliers. In an SVM, the kernel functions that satisfy the mercer conditions are used to do all the operations in learning and testing modes. In the recent years, SVMs have proven to be a dominant technique for pattern classification in many areas like bioinformatics, handwritten character recognition, machine vision, etc. It is a generalized linear classifier with maximum-margin fitting functions. This fitting function offers regularization which helps the classifier generalized better. By controlling the VC dimensions of SVM model, it controls the model complexity.

4.4 Matching techniques

An unknown word is matched to a known word using one of the following techniques (Svendsen et al., 1989).

4.4.1 Whole-word matching

Here the recognizer 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 useful only if the recognition vocabulary is known when the application is developed (S.katagiri).

4.4.2 Sub-word matching

The classifier looks for sub-words - usually phonemes - and then carry out further pattern recognition on those. This technique takes more processing than whole-word matching, but it involves less storage (between 5 and 20 bytes per word). In addition, the pronunciation of the word can be guessed from English text without requiring the user to speak the word beforehand.

(Svendsen et al., 1989 ; Rabiner et al., 1981; Wilpon et al., 1988) discuss that research in the field of automatic speech recognition had been followed for the last three decades, only whole-word based speech recognition systems have found practical use and have obtained dramatic successes.

5. Applications of Speech Recognition

The crucial impact of speech recognition lies whether one can fully integrate the enabling technologies with applications. How to successfully integrate speech into applications often depends on the nature of the user interface and application. An elegant user interface requires carefully considering the particular user group of the application and delivering an application that works effectively and economically. One exclusive challenge in speech-recognition applications is that speech recognition (as well as understanding) is unsatisfactory. It is therefore essential that applications make use of necessary interactive error handling techniques to minimize the impact of these errors. Application developers thus should have adequate knowledge regarding the strengths and weaknesses of the underlying speech technologies and make out the appropriate place to use speech recognition and understanding technology effectively.

There are three broad classes of applications: (1) Cloud-based call center/IVR (Interactive Voice Response): This includes the widely used applications from Tellme's information access over the phone to Microsoft Exchange Unified Messaging; (2) PC-based dictation/command and control: There are a number of dictation applications on the PC. It is a constructive tool for accessibility benefits, but not yet ready for the mainstream. (3) Device-based embedded command control: There is a large variety of devices that do not have a typical PC keyboard or mouse, and the traditional Graphical User Interface (GUI) application cannot be directly extended (Xuedong Huang et al). Mobile phones and automobile scenarios are also very suitable for speech applications. Because of the physical size and hands-busy and eyes-busy constraints, the traditional GUI application interaction model needs a considerable modification. Besides these some more applications of speech recognition domain have been discussed in the following table.

Table 2 shows some practical application of speech recognition.

S.No	Domain	Applications
1.	Education Sector	Teaching students of foreign languages to pronounce vocabulary correctly. Teaching overseas students to pronounce English correctly. Enabling students who are physically handicapped and unable to use a keyboard to enter text verbally.
2.	Domestic sector	Oven, cars, Lock systems, refrigerators, dishwashers and washing machines.
3.	Military sector	High performance fighter aircraft, Battle management, Training air traffic controllers, Telephony and other domains, people with disabilities.
4.	Artificial Intelligence sector	Robotics
5.	Medical sector	Health care, Medical Transcriptions (digital speech to text).
6.	General	Automated transcription, Telematics, Air traffic control, Multimodal interacting, court reporting, Grocery shops, Banking system.
7.	Translation	It is an advanced application which translates from one language to another.

6. Tools for automatic Speech Recognition

6.1 PRAAT

It is free software that can run on wide range of OS platforms and meant for recording and analysis of human speech in mono or stereo. Besides this it can be used cal determining various parameters of speech such as number of samples, time duration, RMS values, Mean power intensity etc.

6.2 AUDACITY

It is free, open source software that can run on wide range of OS platforms and meant for recording and editing sounds. The interface is translated into many languages. Silent features of this software are as:

- Record live audio.
- Convert tapes and records into digital recordings or CDs.
- Edit WAV, AIFF, FLAC, MP2, MP3 or Ogg Vorbis sound files.
- Cut, copy, splice or mix sounds together.
- Change the speed or pitch of a recording

6.3 Hidden Markov Model Toolkit (HTK)

The Hidden Markov Model Toolkit (HTK) is a convenient toolkit for building and manipulating Hidden Markov Models. HTK consists of a set of library modules and tools that are available in C source form. HKT provides sophisticated services for speech analysis, HMM training, testing and result analysis. This software supports HMMs making use of both continuous density mixture Gaussians and discrete distributions and can be used to develop complicated HMM systems. Initially this tool kit has been designed to recognize English, so the characters are stored as 8-bit ASCII standard code. But there are languages which don't make use of this format. For example, Vietnamese language is stored in UTF-8 format. In this case Uni-key software was used to translate Vietnamese character format from UTF-8 to VIQR code which is a convention for writing Vietnamese using ASCII 7 bit format. Taking this into consideration (Nguyen et al., 2004) developed an automatic speech recognition system using HTK.

6.4 SPHINX

CMU Sphinx toolkit has various packages for different tasks and applications. It is sometimes difficult to understand what to choose. To make this thing clear some of them are as:

- Pocketsphinx — recognizer library written in C.
- Sphinxtrain — acoustic model training tools
- Sphinxbase — support library required by Pocketsphinx and Sphinxtrain
- Sphinx4 — adjustable, modifiable recognizer written in java
- CMUclmtk — language model tools

Sphinx 4 is a latest version of Sphinx series of speech recognizer tools that is written entirely in Java programming language. It provides a clear outline for research in speech recognition.

6.5 KALDI

Kaldi is identical in aims and scope to HTK. The primary aim is to have sophisticated and flexible code, written in C++, that is simple to adapt and extend. Important features include:

- Code-level integration with Finite State Transducers (FSTs)
- Extensive linear algebra support
- Extensible design
- Open license

6.6 VOXFORGE

VoxForge is a free speech corpus and acoustic model storehouse for open source speech recognition engines. VoxForge was developed to accumulate transcribed speech to create a free GPL speech corpus for use with open source speech recognition engines. The speech audio files will be 'compiled' into acoustic models for use with open source speech recognition engines such as Julius, ISIP, and Sphinx and HTK.

6.7 JULIUS

It is an open source high performance two-pass large, vocabulary continuous speech recognition decoder software that perform well on Linux OS. During its last revision, a grammar based recognition parser Julian has been integrated it. (Mathur et al. 2010) used Julius to built a domain specific speaker independent continuous speech recognizer. Despite of creating just a base line recognizer, the results were encouraging.

6.8 SCARF

Segmental Conditional Random Field Toolkit for speech recognition (SCARF) is a software toolkit designed for doing speech recognition with the help of segmental conditional random fields. It is designed to permit for the integration of numerous, possibly redundant segment level acoustic features, together with a complete language model, in a coherent speech recognition framework. SCARF is designed to make it particularly suitable to use acoustic detection events as input, such as the recognition of energy bursts, phonemes, or other events.

6.9 Dragon NaturallySpeaking

Dragon NaturallySpeaking is widely considered to be the market leader for Speech-recognition software. The latest version of this software is 12.0 which support both 32-bit and 64-bit editions of Windows XP Vista, 7 and 8. The software has three main areas of functionality: dictation, text-to-speech and command input. The user is able to dictate and have speech transcribed as written text, have a document synthesized as an audio stream, or issue commands that are recognized as such by the program.

7. Performance of Speech Recognition System

Speed and accuracy are the two factors that specify performance of speech recognition systems. Accuracy may be measured in terms of performance accuracy which is usually rated with word error rate (WER), whereas speed is measured with the real time factor. To calculate speed of an ASR system real time factor (RTF) is the parameter used. If an input of duration 'I' takes time 'P' to process, then real time factor is mathematically defined as,

$$RTF = P / I \tag{4}$$

Word Error Rate (WER)

The word error rate calculates misrecognitions at the word level: it compares the words outputted by the recognizer to those the user really spoke. Every error (substitution, insertion or deletion) is counted against the recognizer.

WER = Number of Substitution + Insertions + Deletions/ Total number of words.

$$WER = (S + D + I / N) \tag{5}$$

When considering the performance of a speech recognition system, at that time word recognition rate (WRR) is used instead,

$$WRR = 1 - WER = (N - S - D - I / N) = (H - I / N)$$

Where, H is N-(S+D), the number of correctly recognized words.

8. Technical Challenges and Future Research

Despite successful applications of speech recognition in the marketplace and people's lives, the technology is still far from being faultless and technical challenges abound. The two main technical challenges in adopting speech recognition are: (1) making speech-recognition systems robust in noisy acoustic environments, and (2) creating workable speech-recognition systems for natural, free-style speech. Since then, huge stride has been made in overcoming these challenges, and yet the problems remain unsolved. The single simplest, best way for current high-tech recognition systems to improve performance on a given task is to enhance the amount of task-relevant training data from which its models are created. System capabilities have progressed directly along with the amount of speech corpora available to capture the huge inconsistency inherent in speech. Despite all the speech databases that have been exploited so far, system performance consistently improves when more relevant data are available. This circumstance clearly specifies that more data is required to capture crucial information in the speech signal. This is a key aspect in increasing the facility with which we can learn, understand, and subsequently automatically recognize a wide variety of languages. This capability will be a crucial component in enhancing performance not only for transcription within any given language but also for spoken language machine translation, cross-language information retrieval, and so on. The rich areas for future research of speech may be considered as part of Infrastructure, Knowledge Representation, or Models, Algorithms, and Search. One of the finest areas regarding the future of speech recognition systems is the Creation of high-quality annotated corpora. To improve performance on a given task, the only best way for the present high-tech recognition system is to simply increase the amount of task-relevant training data from which its models are generated. To develop machine representations of meaning that has the capability to capture the communicative intent of a spoken utterance, is one more future research of automatic speech recognition systems. This should enhance the current word error rate measure for speech recognition performance.

Conclusion

One of the most significant ways of communication among humans is language and prime medium used for the said is speech. A number of practical limitations have been encountered that often sets obstacles in the widespread deployment of application and services

provided by ASR systems. The articles briefly described the state-of-the-art of ASR systems. An attempt has been made through this article to provide a review that how much this technology has been emerged in the last seven decades through the never ending efforts of researcher all over the world.

References

L. Rabiner , B.H. Juang (1993), Fundamentals of speech processing, *New Jersey: Prentice Hall.*

F. Jelinek (1997), Statistical methods for speech recognition. *Cambridge, MA: MIT Press.*

V. Zue, R. Cole, W. Ward (1996), Speech Recognition.Survey of the State of the Art in Human Language Technology. *Kauii, Hawaii, USA*

Z. Mengjie (2001), Overview of speech recognition and related machine learning techniques, *http://www.mcs.vuw.ac.nz/comp/Publications/archive/CS-TR-01/CS-TR-01-15.pdf*

Aravind Ganapathiraju , Jonathan Hamaker, Joseph Picone. (2001), Syllable-Based Large Vocabulary Continuous Speech Recognition , *IEEE Transactions On Speech And Audio Processing*, Vol. 9, No. 4.

Sadaoki Furui (2005), 50 years of Progress in speech and Speaker Recognition Research , *ECTI Transactions on Computer and Information Technology*,Vol.1. No.2

K. H. Davis, R. Biddulph, and S. Balashek (1952), Automatic Recognition of spoken Digits, *J.Acoust.Soc.Am*., 24(6):637-642.

J. W. Forgie and C. D. Forgie (1959), Results obtained from a vowel recognition computer program , *J.A.S.A.*, 31(11),pp.1480-1489

T. B. Martin, A.L. Nelson, H.J. Zadell (1964). Speech recognition by feature abstraction techniques, Tech. Report AL-TDR-64-176, Air Force Avionics Lab,

T. K. Vintsyuk (1968), Speech discrimination by dynamic programming, *Kibernetika*, 4 (2), pp. 81-88.

J. Suzuki and K. Nakata (1961), Recognition of Japanese Vowels Preliminary to the Recognition of Speech , *J.Radio Res.Lab* 37(8):193-212

T. Sakai and S. Doshita (1962), The phonetic typewriter, information processing 1962 , *Proc.IFIP Congress* , pp. 445-45

D. R. Reddy (1966), An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave, *Tech.Report No.C549, Computer Science Dept., Stanford Univ*

V. M. Velichko and N .G .Zagoruyko (1970), Automatic Recognition of 200 words, *Int.J.Man-Machine Studies,2:223*

H. Sakoe and S. Chiba (1978), Dynamic Programming Algorithm Optimization for Spoken Word Recognition ,*IEEE Trans.Acoustics, Speech, Signal Proc*.,ASSP-26(1):43- 49

F. Itakura (1975), Minimum Prediction Residula Applied to Speech Recognition ,*IEEE Trans.Acoustics, Speech,Signal Proc*., ASSP-23(1):67-72

C. C. Tappert, W. D. Chapman, X. R. Dixon, and A. S. Rabino (1971)., Automatic recognition of continuous speech utilizing dynamic segmentation, dual classification, sequential decoding and error recovery, *Rome Air Dev. Cen, Rome, NY, Tech. Report TR-71-146*

F. Jelinek, L. R. Bahl, and R. L. Mercer (1975)., Design of a linguistic statistical decoder for the recognition of continuous speech, *IEEE Trans. Information Theory*, IT-21, pp. 250-256.

R. K. Moore (1994), Twenty things we still don t know about speech , *Proc.CRIM/ FORWISS Workshop on Progress and Prospects of speech Research an Technology* .

H. Sakoe (1979), Two Level DP Matching A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition , *IEEE Trans. Acoustics, Speech, Signal Proc*., ASSP-27:588-595.

J. S. Bridle and M. D. Brown (1979), Connected Word Recognition Using whole word templates , *Proc. Inst.Acoust.Autumn Conf.,25-28.*

C. S. Myers and L. R. Rabiner (1981), A Level Building Dynaimc Time Warping Algorithm for Connected Word Recognition , IEEE *Trans. Acoustics, Speech Signal Proc*., ASSP-29:284-297.

C. H. Lee and L. R. Rabiner (1989), A Frame Synchronous Network Search Algorithm for Connected Word Recognition , *IEEE Trans. Acoustics, Speech, Signal Proc*., 37(11):1649-1658.

S. Furui (1986), Speaker independent isolated word recognition using dynamic features of speech spectrum, *IEEE Trans.Acoustics, Speech, Signal Processing*, ASSP-34, pp. 52- 59.

S. Katagiri (2003), Speech pattern recognition using neural networks, *Pattern Recognition in Speech and Language Processing*, CRC Press, pp. 115-147.

R. P. Lippmann (1987), An introduction to computing with neural nets, *IEEE ASSP Mag*., 4 (2), pp. 4-22.

A. Weibel, et. al. (1989), Phoneme recognition using time-delay neural networks, *IEEE Trans. Acoustics, Speech, Signal Proc.*, 37, pp. 393-404.

B. H. Juang and S. Furui (2000), Automatic speech recognition and understanding: A first step toward natural human machine communication , *Proc.IEEE*,88,8,pp.1142- 1165

C. J. Leggetter and P. C. Woodland (1995), Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, *Computer Speech and Language*, 9, pp. 171-185

A. P. Varga and R. K. Moore (1990), Hidden Markov model decomposition of speech and noise, *Proc. ICASSP*, pp. 845-848

M. J. F. Gales and S. J. Young (1993), Parallel model combination for speech recognition in noise, *Technical Report, CUED/F-INFENG/TR135*,

K. Shinoda and C. H. Lee (2001), A structural Bayes approach to speaker adaptation, IEEE *Trans. Speech and Audio Proc.*, 9, 3, pp. 276-287

Mohamed Afify and Olivier Siohan (2004), Sequential Estimation With Optimal Forgetting for Robust SpeechRecognition , IEEE *Transactions On Speech And Audio Processing*, Vol. 12, No. 1, pp. 19-26.

Mohamed Afify, Feng Liu, Hui Jiang (2005), A New Verification-Based Fast-Match for Large Vocabulary Continuous Speech Recognition , *IEEE Transactions On Speech And Audio Processing*, Vol. 13, No. 4, pp 546-553.

Simon Kinga and Joe Frankel (2006), Recognition ,Speech production knowledge in automatic speech recognition , *Journal of Acoustic Society of America.*

V. Arminas, G. Tamulevicius, D. Navakauskas, E. Ivanovas (2010), Acceleration of feature extraction for FPGA based speech recognition, in *Proc. SPIE 7745, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments*, pp. 774511–774511-6

G. Tamulevičius, V. Arminas, E. Ivanovas, D. Navakauskas (2010), Hardware accelerated FPGA implementation of Lithuanian isolated word recognition system, *Elektronika ir elektrotechnika* (*Electronics and Electrical Engineering*), vol. 99, no. 3 , pp. 57–62.

E. Ivanovas (2012), Development and implementation of means for word duration signal processing, *Ph.D. dissertation, Dept. of Electronic Systems., Vilnius Gediminas Technical Univ., Vilnius.*

A. Chakravarty (2013), *Speech recognition toolkit for the Arduino*,. [Online]. Available: http: //arjo129.github.com/uSpeech/

Natural Language Processor, Sensory, 2010. [Online]. Available: http: //www.sensoryinc.com/products/NLP-5x.html

J. Choi, K. You, W. Sung (2010), An FPGA implementation of speech recognition with weighted finite state transducers, *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2010, pp. 1602– 1605.

R. Veitch, L. M. Aubert, R. Woods, S. Fischaber (2010), Acceleration of HMM-based speech recognition system by parallel FPGA Gaussian calculation, *VI Southern Programmable Logic Conf.* , pp. 197–200

G. Zhang, J. Yin, Q. Liu, Ch. Yang (2011), A real-time speech recognition system based on the implementation of FPGA, *in Proc. Cross Strait Quad-Regional Radio Science and Wireless Technology Conf.,* pp. 1375–1378.

S. T. Pan, C. C. Lai, B. Y. Tsai (2011), The implementation of speech recognition systems on FPGA-based embedded systems with SoC architecture, *Int. Journal of Innovative Computing, Information and Control*, vol. 7, no. 11, pp. 6161–6175.

D. Sart, A. Mueen, W. Najjar, V. Niennattrakul, E. Keogh (2010), Accelerating dynamic time warping subsequence search with GPUs and FPGAs, *IEEE 10th Int. Conf. on Data Mining*, pp. 1375–1378.

Y. Zhang, K. Adl, J. Glass (2012), Fast spoken query detection using lowerbound dynamic time warping on graphical processing units, *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, , pp. 5173–5176. [Online].

M. Davel and E. Barnard (2004), The efficient generation of pronunciation dictionaries: Human factors during bootstrapping, *Interspeech/ICSLP*

M. Davel and O. Martirosian (2009), Pronunciation dictionary development in resource-scarce environments, *Interspeech.* Pp 1-8.

Heiga Zen, Andrew Senior, Mike Schuster (2013) , Statistical parametric speech synthesis using deep neural networks. *Proceedings of the IEEE*

International Conference on Acoustics, Speech, and Signal Processing (ICASSP), IEEE , pp. 7962-7966

Xin Lei, Andrew Senior, Alexander Gruenstein and Jeffrey Sorensen (2013), Accurate and Compact Large Vocabulary Speech Recognition on Mobile Devices, *INTERSPEECH*

J. Zheng et al.,(2010) Implementing SRI's Pashto speech-to-speech translation system on a smart phone, in *SLT*.

J. Xue, X. Cui, G. Daggett, E. Marcheret, and B. Zhou,(2012), Towards high performance LVCSR in speech-to-speech translation system on smart phones, *in Proc. Interspeech*,.

R. Prasad et al.,(2013), BBN Transtalk: Robust multilingual two-way speech-to-speech translation for mobile platforms, *Computer Speech and Language*, vol. 27, pp. 475–49.

Johan schalkwyk, Doug Beeferman, Francoise Beaufags Bill Byrne, Ciprian chelba, Mike Cohen, Maryam Garrette and Brain Strope,(2010) , Google search by voice: A case study, *in advances in speech recognition,Mobile Envirronments call centres, and clinics, Amy Neustein, Ed. Springer- Verlag*.

Jiulong Shan, Genqing Wu, Zhihong Hu, Xiliu Tang, Martin Jansche and Pedro(2010) Search by voice in Mandarin Chinese, *in Proceedings of interspeech*. Pp 354-357

Yun-Hsuan, Martin Jansche and Pedro Moreno,(2011) , Deploying search by voice in Cantonese, *InProceedings of interspeech*,.

Yu Zhong, T.V. Raman, Casey Burkhardt, Fadi Biadsy and Jeffrey P. Bigham (2014), JustSpeak: Enabling Universal Voice Control on Android.

Lei Xin, Andrew Senior, Alexander Gruenstein, and Jeffrey Sorensen(2013). Accurate and Compact Large Vocabulary Speech Recognition on Mobile Devices, *interspeech*.

Zahi N.Karam,William M.Campbell ,A new Kernel for SVM MIIR based Speaker recognition, MIT Lincoln Laboratory, Lexington, MA, USA.

Lori f. lamel, lawrence r. rabiner, aaron e. rosenberg and jay g. wilpon,(1981) An Improved Endpoint Detector for Isolated Word Recognition, *IEEE transactions on acoustics, speech, and signal processing*, vol. assp-29, no. 4.

Anand Singh, Dr. Dinesh Kumar Rajoriya and Vikash Singh, (2012), Database Development and Analysis of Spoken Hindi Hybrid Words Using Endpoint Detection, *International Journal of Electronics and Computer Science Engineering*, Vol.1.

Rabiner, L. Juang, B. H., Yegnanarayana, B.(2010), Fundamentals of Speech Recognition, *Pearson Publishers*.

Garg, A., Nikita, Poonam,(2011), Connected digits recognition using Distance calculation at each digit, *IJCEM International Journal of Computational Engineering & Management*, Vol. 14, ISSN (Online): 2230-7893.

Taabish Gulzar, Anand Singh and Sandip Vijay(2014). An Improved Endpoint Detection Algorithm using Bit Wise Approach for Isolated, Spoken Paired and Hindi Hybrid Paired Words. *International journal of computer applications*, 0975-8887, Volume 92 – No.15

Rajoriya, D.K., Anand, R.S. and Maheshwari R.P. (2010), Hindi paired word recognition using probabilistic neural network, *International Journal Computational Intelligence studies, Vol.1*.

Hisashi Wakita(1977), Normalization of Vowels by Vocal Tract Length and Its Applications to Vowel Identification, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol.25.

Nicolás Morales1, John H. L. Hansen2 and Doorstep T. Toledano1, MFCC Compensation for improved recognition filtered and band limited speech, *Center for Spoken Language Research, University of Colorado at Boulder, Boulder (CO), USA.*

M.A.Anusuya ,S.K.Katti(2009) ,Speech Recognition by Machine: A Review, *International journal of computer science and Information Security*. Vol. 6, No. 3.

Kenneth Thomas Schutte(2003) ,Parts-based Models and Local Features for Automatic Speech Recognition, *B.S.,University of Illinois at Urbana-Champaign S.M., Massachusetts Institute of Technology*.

W. M. Campbell_, D. E. Sturim W. Shen D. A. Reynolds, J. Navr´atily(2006), The MIT- LL/IBM Speaker recognition System using High performance reduced Complexity recognition, *MIT Lincoln Laboratory IBM*

Sannella,,M Speaker recognition Project Report report, *From http://cs.joensuu.fi/pages/tkinnu/research/index.*

Samudravijay K Speech and Speaker recognition report source: http://cs.jounsuu.fi/pages/tkinnu/reaserch/index.

E. Singer, P.A. Torres-Carrasquillo, T.P. Gleason, W.M. Campbell, and D.A. Reynolds ,Accoustic ,phonetic and discriminative approach to automic Language Idantification.

Viet Bac Le, Laurent Besacier, and Tanja Schultz, Acousticphonetic unit similarities for context dependant acoustic model portability Carnegie Mellon University, Pittsburgh, PA, USA

D.R.reddy (1996), An Approach to Computer speech Recognition by direct analysis of the speechwave ,*Tech.Report No.C549,Computer Science Department ,Stanford University,*

Dat Tat Tran (2000), Fuzzy Approaches to Speech and Speaker Recognition, A thesis submitted for the degree of Doctor of Philosophy of the university of Canberra.

J. K. Baker (1975), Stochastic modeling for automatic speech recognition, *Speech Recognition, D. R. Reddy, Ed. New York: Academic*,

F. Jelinek (1976), Continuous speech recognition by statistical methods, *Proc. IEEE*, vol. 64, no. 4, pp. 532–557.

H. Poor (1988), An Introduction to Signal Detection and Estimation *(Springer Texts in ElectricalEngineering)*, J. Thomas, Ed. New York: Springer-Verlag

J. M. Naik, et. al .(1989) Speaker verification over long distance telephone lines, *Proc. ICASSP*, pp. 524-527.

Balwant A. Sonkamble , D. D. Doye (2012), Speech Recognition Using Vector Quantization through Modified K-meansLBG Algorithm. Computer Engineering and Intelligent Systems ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)Vol 3, No.7.

L.R.Bahl et.al (1993), A method of Construction of acoustic Markov Model for words, *IEEE Transaction on Audio ,speech and Language Processing* ,Vol.1

W. S. McCullough and W. H. Pitts,(1943), A Logical Calculus of Ideas Immanent in Nervous Activity, *Bull. Math Biophysics*, Vol. 5, pp. 115-133

V. N. Vapnik (1995), The Nature of Statistical Learning Theory, *Springer, New York*.

Sadoaki Furui (1991), Speaker dependent feature extraction, recognition and processing techniques. *Speech communication* vol 10, 505- 520.

Svendsen T., Paliwal K. K., Harborg E., Husy P. O. (1989). Proc. ICASSP'89, Glasgow.

S.katagiri,, Speech Pattern recognition using Neural Networks.

Rabiner, L., R., and Wilpon, J. G., (1979). Considerations in applying clustering techniques to speaker-independent word recognition, *.Journal of Acoustic Society ofAmerica*vol .66(3):pp663-673.

Wilpon J.G., D.M.DeMarco,R.P.Mikkilineni (1988) Isolated word recognition over the DD telephone network -Results of two extensive field studies, *Proc. ICASSP*,pp. 55-58

Xuedong Huang and Li Deng, An Overview of Modern Speech Recognition. *Microsoft Corporation.*

Nguyen Hong Quang, Trinh Van Loan, LE The Dat(2004), Automatic Speech Recognition for Vietnamese using HTK.

Mathur, R., Babita, Kansal, A.(2010), Domain specific speaker independent continuous speech recognizer using Julius, *Proceedings of ASCNT , CDAC, Noida, India*, pp. 55 – 60.