UGC MAJOR RESEARCH PROJECT

FINAL REPORT

Speech-based Content Retrieval for Visually Impaired

Principal Investigator : **Dr. P. Dhanalakshmi**
Associate Professor
Department of Computer Science and Engineering
Annamalai University
Annamalainagar
Tamil Nadu - 608 002

Co-Investigator : **Dr. M. Balasubramanian**
Assistant Professor
Department of Computer Science and Engineering
Annamalai University
Annamalainagar
Tamil Nadu - 608 002

Table of Contents

1. Introduction to Speech based Content Retrieval

Speech based document retrieval system provides the facility to retrieve the spoken document on the basis of speech query. This system is designed for users who wish to access the Internet in a non-visual way. This includes blind or partially sighted users, people with dyslexia or learning difficulties, and users who are learning new languages. It provides a speech interface that allows visually impaired users to interact independently and efficiently with the computer. It is designed to interact directly with the information on a web page, and to translate it into speech (Mark S. Hawley, 2013). Such system can be implemented by combining three techniques: Speech Recognition, Document Retrieval and Speech Synthesis. Fig. 1 depicts the methodologies implicated in the proposed system. Speech Recognition is the process of translating the spoken words into text. Keyword query is converted into text using speech recognition. Section 3 describes the working procedure of speech to text conversion to isolate words. Text retrieval or document retrieval system extracts the relevant documents for the given search query. This query can be of single search term or multi-search term. The steps involved in document retrieval are discussed in Section 4. Finally, speech synthesis system is implemented, which converts the written text into speech utterances and is illustrated in Section 5. Finally, performance of speech based content retrieval system is depicted in Section 6.

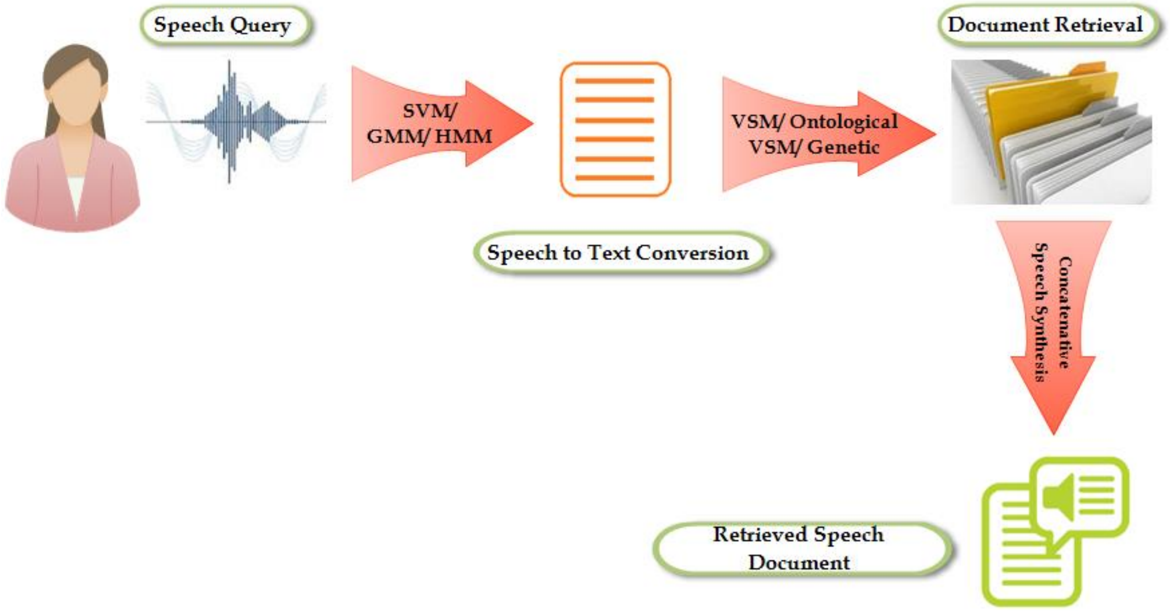

Speech Query
SVM/ GMM/ HMM
Speech to Text Conversion
VSM/ Ontological VSM/ Genetic
Document Retrieval
Concatenative Speech Synthesis
Retrieved Speech Document

Fig. 1: Techniques involved in speech based spoken document retrieval

Speech Recognition (SR) is an operating system which enables to convert spoken words into written text. In general, it involves extraction of pattern from digitized speech samples and representing them using an appropriate data model. These patterns are subsequently compared to each other using mathematical operations to determine their contents. SR is used to translate the words spoken by human so as to make them system recognizable (Anusuya, M.A. and Katti, S.K., 2011). For the past few decades plenty of researches have been investigated in speech processing. Generating speaker independent speech recognition system is a challenging task in the field of speech processing and it has an increasing attention in many real world applications like computer aided system, hearing impaired, visually impaired etc. The first module aims to generate speaker independent isolated word recognition system. Acoustic features namely Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC) are extracted from the speech signal. Support Vector Machine (SVM), Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) modeling techniques are used to model these acoustic features extracted from the spoken query word. The performance of keyword extraction from the speech is analyzed for SVM, GMM and HMM. Each isolated word segment from the test sentence is matched against the models generated by the SR system for finding the semantic representation of the test input speech. After extracting the keywords from the speech utterances, the resultant keywords are given to the document retrieval system.

Plenty of researches have been investigated in the field of Document/ Information Retrieval. Document retrieval helps users to find relevant information from the huge repository of data that matches the user's requirements. In general, search engines uses keyword/query based approach. If a document contains the query in the vocabulary then it is returned as match and their relevant documents are retrieved. Ranking algorithms are used to retrieve the most relevant results at the top of the list and the least relevant at the bottom (Geir Solskinnsbakk and Jon Atle Gulla, 2010). Typically search engines return the results as a text; the current work focuses on retrieving the results as a speech.

In this work, document retrieval is achieved by Vector Space Model (VSM), ontology assisted VSM and Genetic Approach (GA). VSM and ontological VSM is a simple and effective method for retrieving the document from the collection of text database. Document retrieval system is achieved by three stages: keyword extraction, query re-formulation and similarity measures (based on cosine angle) between the documents and the query respectively.

Genetic Algorithm (GA) is a probabilistic algorithm which simulates the process of natural selection of living organisms to find an approximate solution to a problem (Goldberg, 2003). Chromosome representation gives the collection of information regarding individuals present in the chosen population. Initially, genotype information for the entire population is mapped into the decision variable space which is termed as phenotype. Query term is mapped against phenotype information of each individual using fitness functions. In this work, fitness function is computed using three similarity measures: i) Dice co-efficient ii) Jaccard co-efficient and iii) Cosine Similarity respectively. Fittest individual is identified and its appropriate text documents are retrieved using Roulette Wheel Selection procedure. Experiments have been conducted for different fitness functions and results are evaluated. Retrieval system returns an ordering of document over the collection of document for the required query. Documents are ranked and the resultant documents are given into the synthesis system.

Speech synthesis or Text to Speech (TTS) system converts symbolic linguistic representations/regular languages like syllabic and phonetic transcriptions into speech. A computer system used for this purpose is termed as Speech Synthesis (SS). Synthesized speech can be extracted by concatenating pieces of recorded speech that are stored in database. Every system differs in the size of the stored database. Recent inclination is to collect huge collection of fluent speech database and to select an optimal sequence of acoustic units at run time to synthesize a particular utterance. The effectual method for TTS conversion is concatenation method.

In this work, text normalization is initially performed. In general, text normalization is the conversion of text that includes non-standard words such as statistics, abbreviations, misspelling into standard word representation (Mahwash Ahmed and Shibli Nisar, 2014). Phonemes/syllables for each isolated word in the retrieved documents are extracted using dictionary based phonemic transcription. Synthesizer generates the actual utterance of speech by concatenating the pre-recorded pieces of phonemes.

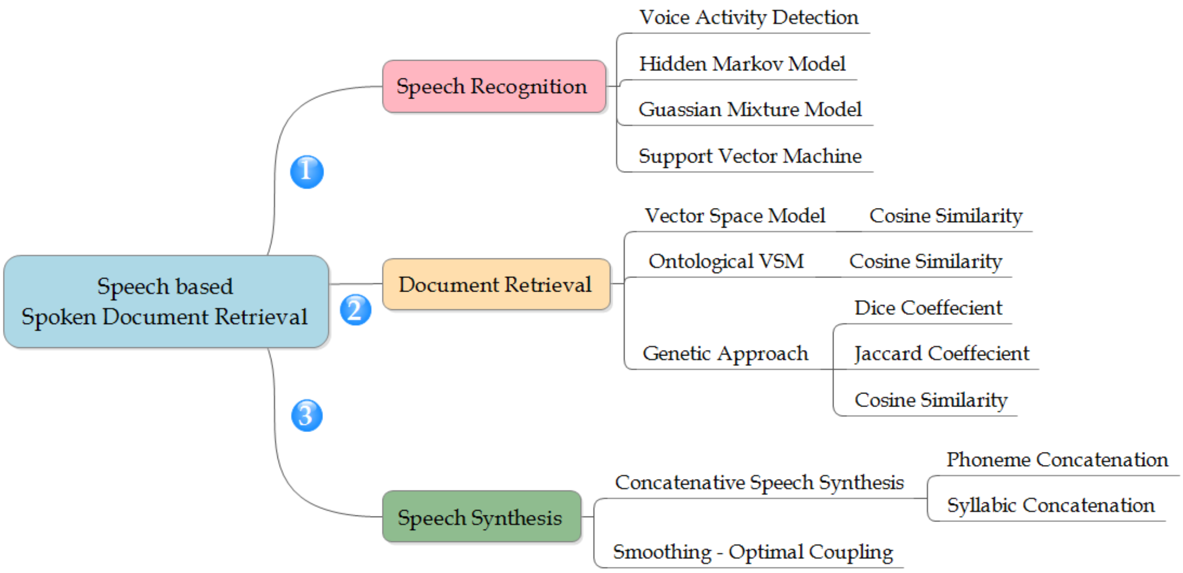


Fig. 2: Framework of the proposed system

The TTS engine identifies the beginning and ending of sentences. The pitch likely falls near the end of a statement and rises, for a question. Similarly, the machines starts speaking the phoneme/syllable and it falls on to the last word, and then pauses are placed in between the sentences or phonemes/syllables for clear reading. This process is achieved by prosody generation (Jong Kuk Kim et al. 2009). After concatenating the phoneme/syllable sound clip, it is found that a discontinuity exists in the speech. To avoid such problems and to increase the efficiency of the system, smoothing of speech utterances can be done. Optimal coupling is an effective method for smoothing and is also implemented to improve the pleasantness and naturalness of the speech.

1.1 Speech Recognition

Speech Recognition is an operating system which enables to convert spoken words into written text. An effective measure of intelligibility of a speech recognition system is the probability of correct recognition of the transmitted message. SR is used to translate the words spoken by human so as to make them system recognizable (José A. González, 2013). In general, it involves extraction of pattern from digitized speech samples and representing them using an appropriate data model. These patterns are subsequently compared to each other using mathematical operations to determine their contents.

Speech recognition system performs two fundamental operations: Acoustic feature extraction and pattern matching. Acoustic feature extraction is the process of converting speech signal into a set of parameters (Sabato Marco Siniscalchi et al.,2013). Pattern matching is the task of finding parameter set from memory which closely matches the parameter set obtained from the input speech signal. Spectral shaping is the process of converting the speech signal from sound pressure wave to a digital signal; and emphasizing important frequency components in the signal (Al-HaddadS.A.R. et. al., 2009). Feature extraction is the process of obtaining different features such as pitch, power, and vocal tract configuration from the speech signal.

During recent years, a number of researchers have investigated in the field of speech processing because of its increasing attention in many real world applications. Many solutions have emerged in the past few decades to ease the day to day life of differently abled people. The proposed work focuses on Speech Recognition (SR), information retrieval and speech synthesis system, which is applicable for visually disabled persons. Speech disorders or hearing disorders are a type of communication disorder where 'normal' speech or hearing is disrupted.

1.1.1 Basics of speech and speech production mechanism

Speech is the natural mode of communication for people and is the vocalized form of human communication. Speech is researched in terms of speech production and speech perception of sound used in vocal languages. Isolated word recognition is a manner of reading based on immediate perception of what word a familiar grouping of letters represents. Due to the pressure of the glottis and the air pushed from lungs, the vocal cords can open and close very quickly, which generates vibrations in the air. Speech sounds can be analysed from several point of view: Articulatory, acoustics, phonetic and perceptual. A smallest unit of meaningful sound is defined as phoneme. Vowels are voiced sound and consonants are voiceless sound. Vowels and consonant phonemes are classified in terms of place and manner of articulation and voicing. Place of articulation refers to the location in the vocal tract. Manner of articulation refers how the vocal tract restricts airflow.

Generally the place of articulation is named according to the passive articulator, the one which moves less in forming the sound. Still, it is more precise to name both articulators, active and passive. For example, in English *t* is alveolar, and more precisely, apico-alveolar. The most frequently mentioned places of articulation (in order from the lips towards the throat) are bilabial, dental, palatal, alveolar, palato-alveolar, labiodentals, retroflex, velar, and glottal. Place of articulation and manner of articulation is shown in Fig.3 and Fig. 4. These bilabial, dental, palatal, alveolar, palato-alveolar, labiodentals, retroflex, velar, and glottal are the speech organs to generate different signals for pronunciation.

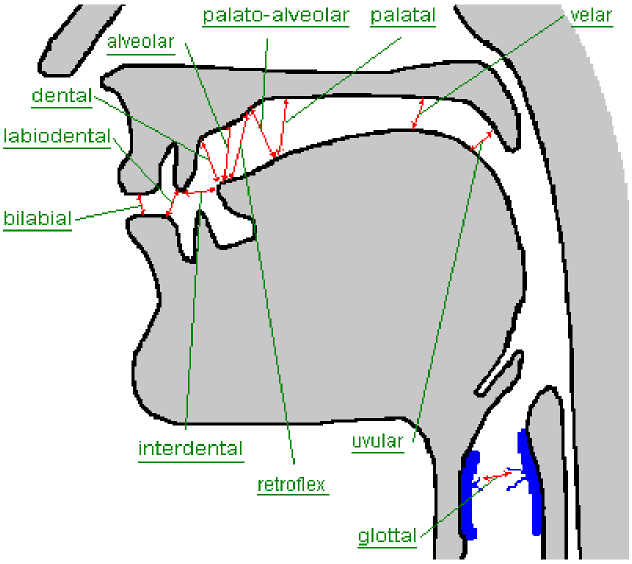


Fig. 3: Place of articulation

Human speech production to speech recognition involves the following steps:

- Rapid open and close of vocal cords (glottis) to generate the vibration in air flow.
- Resonance of the nasal cavity, oral cavity and pharyngeal cavity.
  - The vibrations of air
  - The reception of the inner ear
  - The recognition by the brain

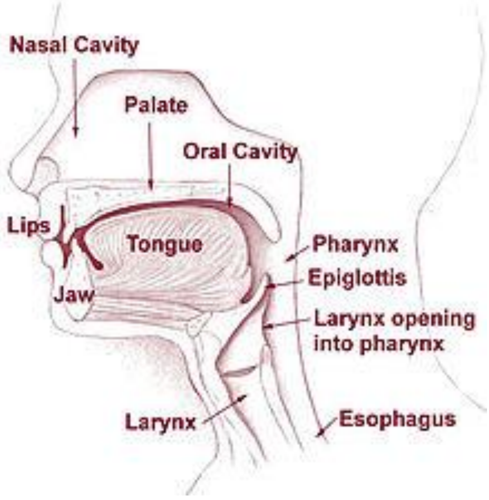


Fig. 4: Manner of articulation

Speech Recognition is a multileveled pattern recognition task in which acoustic signal is examined and structured into hierarchy of sub word units (Phones) words, phrases and sentences.

1.1.2 Phonetics

Phonetics is a branch of linguistics that comprises the study of the sounds of human speech, sign languages and the equivalent aspects of sign. It is concerned with the physical properties of speech sounds or signs (phones): their physiological production, acoustic properties, auditory perception, and neuro-physiological status. Phonology is concerned with the abstract, grammatical characterization of system of sound or signs.

The field of phonetics is a multiple layered subject of linguistics that focuses on speech. Phonetics as a research discipline which has three main branches:

- Articulatory phonetics
- Acoustic phonetics
- Auditory phonetics

Articulatory phonetics is concerned with the articulation of speech: the position, shape, and movement of articulators or speech organs, such as the lips, tongue, and vocal folds. Acoustic phonetics is concerned with acoustics of speech: the spectro-temporal properties of the sound waves produced by speech, such as their frequency, amplitude, and harmonic structure. Auditory phonetics is concerned with speech perception: the perception, categorization, and recognition of speech sounds and the role of the auditory system and the brain in the same.

Acoustic phonetics is concerned with acoustics of speech: The spectro-temporal properties of the sound waves produced by speech, such as their frequency, amplitude, and harmonic structure. Auditory phonetics is concerned with speech perception: the perception, categorization, and recognition of speech sounds and the role of the auditory system and the brain in the same. Vowel qualities are classified according to tongue position (high/mid/low, closed/ open and front/central/back) and rounding of the lips. Consonants are classified primarily according to place of articulation, manner of articulation, and voicing. Consonants are classified primarily according to place of articulation, manner of articulation, and voicing. Phonetic symbols for vowels and consonants are shown in Fig. 5 and Fig.6.

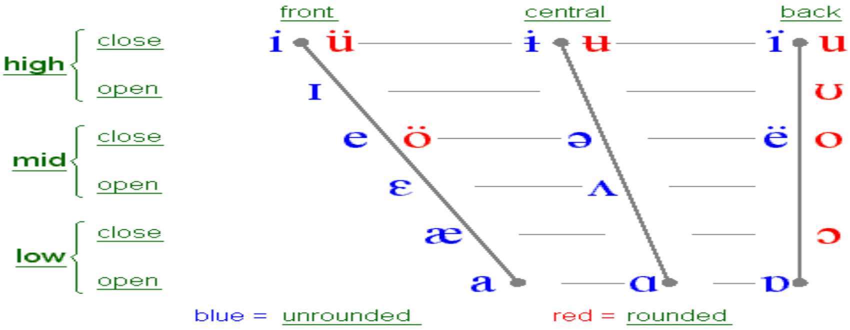
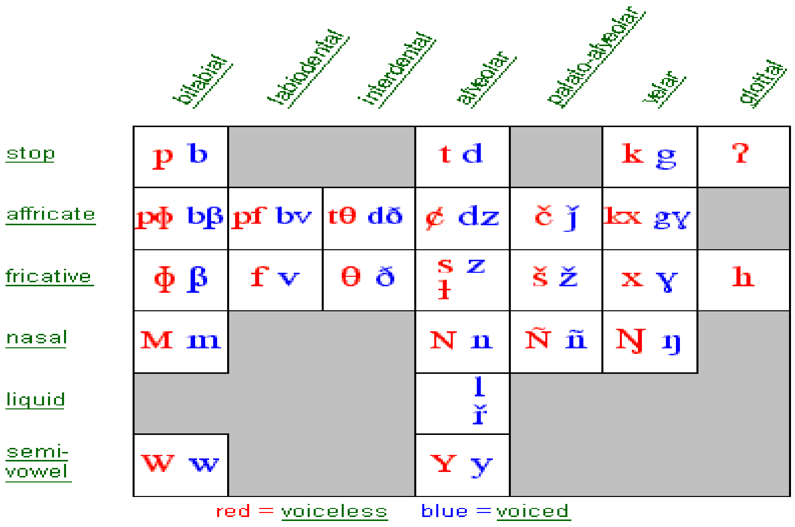


Fig. 5: Phonetic symbol for vowel



	bilabial	labiodental	interdental	alveolar	palato-alveolar	velar	glottal
stop	p b			t d		k g	ʔ
affricate	pɸ bβ	pf bv	tθ dð	ȼ dz	č ǰ	kx gɣ	
fricative	ɸ β	f v	θ ð	s z ɬ	š ž	x ɣ	h
nasal	M m			N n	Ñ ñ	Ŋ ŋ	
liquid				l ř			
semi-vowel	W w			Y y			

red = voiceless blue = voiced

Fig. 6: Phonetic symbol for consonants

1.1.3 Speech Communication

Speech communication refers to the process associated with the production and perception of sounds used in spoken language. Fig. 7 shows the applications of speech communication. Speech communications are highly applicable in the following fields: Digital transmission and storage, Speech Recognition (SR), Speech Synthesis, Speaker Identification (SI) and Speaker Verification (SV). The proposed work focuses on Speech Recognition (SR) system. It is exceedingly applicable for disabled persons. Speech disorders or hearing disorders are a type of communication disorder where ’normal’ speech or hearing is disrupted.

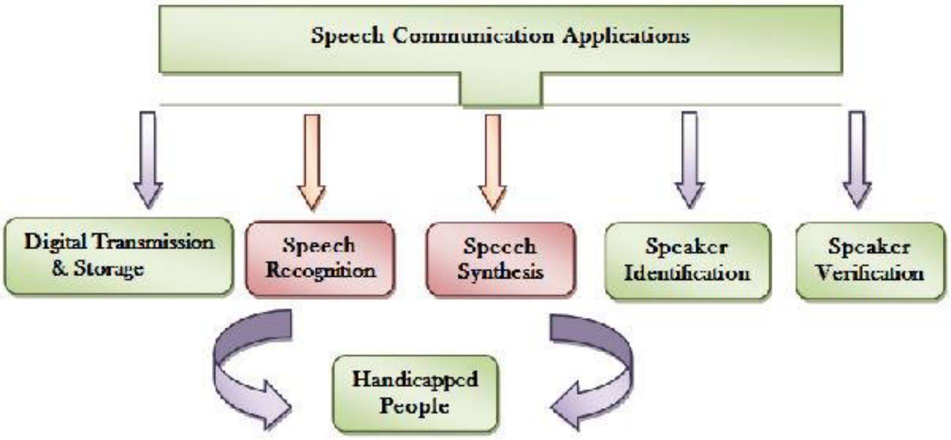


Fig. 7: Real time applications of speech communication system

1.1.4 Types of Speech Recognition

Speech Recognition is classified into two main classes namely Speaker-Dependent and Speaker-Independent. Speaker dependent systems generally involve training a system to recognize each of the vocabulary words uttered single or multiple times by a specific set of speakers while for speaker independent systems usually such training methods are not applicable and words are recognized by analyzing their inherent acoustical properties.

Speech Recognition system can be separated in different classes by describing the type of utterances they can recognize. They are

1. Isolated word recognition
2. Connected word recognition
3. Continuous speech recognition
4. Spontaneous speech recognition

In this work an isolated word recognition system is designed for speech recognition. Science of Isolated word recognition indicates that the human use the letter within a word to recognize a word. It accepts single word or single utterances at a time.

1.1.5 Issues in Speech Recognition

The main challenge of SR involves modeling the variation of the same word as spoken by different speakers depending on speaking styles, regional, pitch, gender, loudness, voice, accents, etc., In addition, changing of signal properties over time and background noises also may create major problem in Speech Recognition (Bishnu Prasad Das and Ranjan Parekh, 2012). To overcome this concern, it involves extraction of patterns from digitized speech samples and representing them using an appropriate data model.

1.2 Document Retrieval

Document retrieval or text retrieval is a branch of information retrieval where the information is stored primarily in the form of text. Due to increasing interest in text retrieval it becomes a critical area of research, since it is fundamental to all search engines. Document retrieval system can be considered as a basic and important tool for text mining that is capable of taking a user's information need into an account (GeirSolskinnsbakk and Jon AtleGulla, 2010). Document retrieval is a hard task if multi topic lengthy documents have to be restricted with a very short description (a few keywords) of the information need. In general Information Retrieval (IR) is defined as the problem of selection of document information from storage in response to search questions provided by a user. It deals with document database containing textual, pictorial or vocal information and it process the user's queries and tries to allow the user to access relevant information in an appropriate time interval (John Lafferty and Chengxiang Zhai, 2001). Fig. 8 illustrates the framework for information retrieval system.

1.2.1 Keyword Search

An isolated/individual word in the user requirement is considered as keywords to retrieve documents in an information system, for instance, a catalogue or a search engine. A popular form of keywords on the web is tags or terms which are directly visible and can be assigned by non-experts also. Index terms can consist of a word, phrase, or alphanumerical term. They are created by analyzing the document either manually with subject indexing or automatically with automatic indexing or more sophisticated methods of keyword extraction. Index terms can either come from a controlled vocabulary or be freely assigned.

Keywords are stored in a search index. Common words like articles (a, an, the) and conjunctions (and, or, but) are not treated as keywords because it is inefficient to do so.

Almost every English-language site on the Internet has the article "the", and so it makes no sense to search for it.

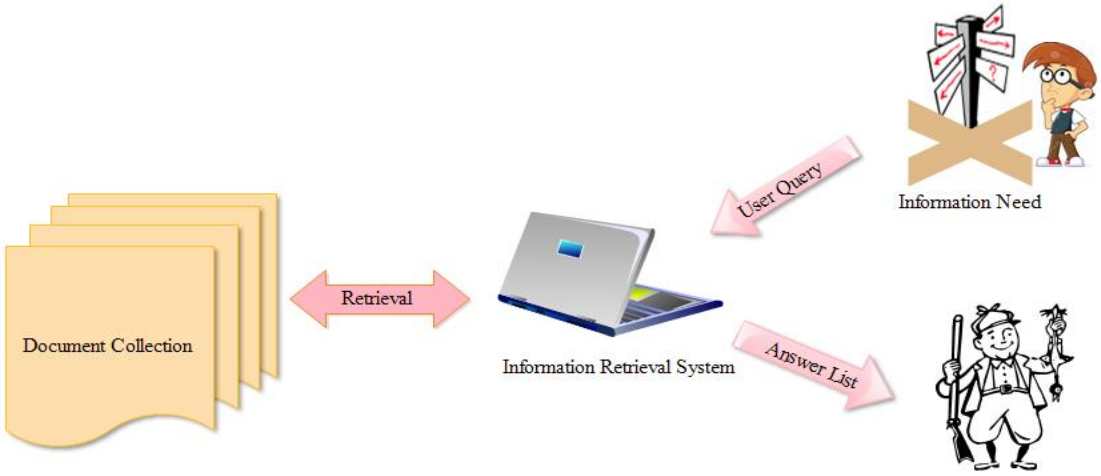


Fig. 8: Information/document retrieval system

In massive collection of text retrieval, all the words in each document are considered to be keywords. The word '*term*' is used to refer the words in a document. Ranking of documents on the basis of estimated relevance to a query is critical. Document retrieval is a hard task if multi topic lengthy documents have to be restricted with a very short description (a few keywords) of the information need.

1.2.2 Similarity based Ranking

A similarity measure is a function that computes the degree of similarity between two vectors. Similarity can be used to refine answer set for the keyword query. User selects few relevant documents from those retrieved document and system finds other documents similar to these selected documents. Using a similarity measure between the query and each document, it is possible to rank the retrieved documents in the order of presumed relevance. It is possible to enforce a certain threshold so that the size of the retrieved set can be controlled.

1.3 Speech Synthesis

Speech Synthesis (SS) is the imitation production of human speech. Speech Recognition and Speech Synthesis plays an imperative role in Human-Machine Interaction. Synthesized speeches are extracted from concatenating pieces of pre-recorded speech utterances from the database. The proposed work converts the written text into a syllabification (syllable text representation) and subsequently it converts syllable representation into modified syllable waveform clips. Such clips can be combined together to generate a sound. Syllabic transcription attempts to describe the individual variations that occur between speakers of a dialect or language. Concatenative Speech Synthesis method generates highly understandable speech utterance. Speech Synthesis refers to the capability of computer to reproduce the normalized text into machine generated speech.

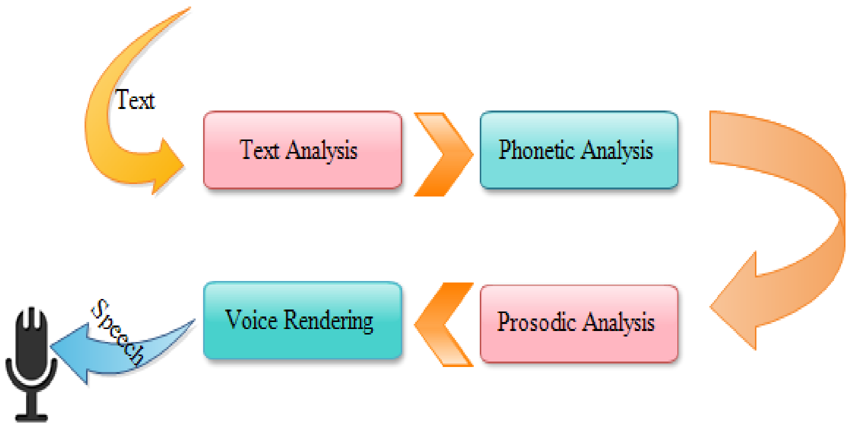


Fig. 9: Speech synthesis system

In general, Speech Synthesis functions as a medium which converts text into speech and is shown in Fig.9. Systems may differ in their size of the stored database depend on its requirements or applications. Recent inclination is to collect massive database of fluent speech. The effectual method for text to speech conversion is concatenation method. In this method, the system intends to select an optimal sequence of acoustic units at run time to synthesize a particular utterance.

1.3.1 Speech Synthesis Methods

The most important qualities of a speech synthesis system are naturalness and intelligibility. Naturalness describes how closely the resultant sounds like human speech, while intelligibility is the ease with which the output is understood. The ideal speech synthesizer usually tries to maximize both natural and intelligible characteristics.

- Formant Synthesis
- Concatenative Synthesis
- Articulatory Synthesis

Current work focuses on concatenative synthesis method which concatenates pieces of recorded speech that are pre-recorded into the database. Naturalness of speech in concatenative SS is high when compared to other synthesis methods.

2. Literature Review

The popularity and ubiquity of multimedia associated with spoken documents has spurred a lot of research interest in spoken document retrieval (SDR) in the recent past (Berlin Chen et al., 2014). (Kuan-Yu Chen and Berlin Chen, 2010) focuses on comparison of two common categories of topic modeling techniques for spoken document retrieval (SDR), namely document topic model (DTM) and word topic model (WTM). In (Berlin Chen et al., 2011) many efforts have been devoted to develop elaborate indexing and modeling techniques for representing spoken documents, but only few approaches are developed to improve query formulations for better representing the users' information needs. The query and documents with different granularities of index features to work in conjunction with the various relevance and/or non-relevance cues are investigated in (Berlin Chen et al., 2011). Language Modeling (LM) framework is used to combine several kinds of information cues into the process of feedback document selection for enhanced query formulation in SDR Yi-Wen Chen et al., (2013).

Human beings impose some constraints on the sequence of sound units while producing speech (Rao K. S. and Yegnanarayanan B., 2007). Speech Recognition usually involves extraction of patterns from digitized speech samples and representing them using an appropriate data model. These patterns are subsequently compared to each other using mathematical operations to determine contents (Bishnu Prasad Das and Ranjan Parekh, 2012). In (Linga Murthy M. K. and Murthy G. L. N., 2012), LPC analysis extracts the features of given words and vector quantization is used for feature matching. Several works have been reported in (Benesty J et al., 2007) for speech processing: from the view of literature, different types of spectral features such as LPCC, PLP and MFCC are repeatedly and widely used feature extraction techniques in speech recognition.

A serious problem with the LPC is that they are highly correlated but it is desirable to obtain less correlated features for acoustic modeling (Anusuya M. A. and Katti S. K., 2011). From (Utpal Bhattacharjee, 2013), it has been observed that the performance of LPCC based system degrades more rapidly compare to MFCC. In (Ramu Reddy V, Sudhamay Maity and Sreenivasa Rao, 2013), prosodic characteristics are acquired over a period of time and prosodic events appear to be time-aligned with syllables or group of syllables. An iterative algorithm based on conventional DTW algorithm and on an averaging technique is used for determining the best prototype during the training phase in order to increase model discrimination (Hocine Bourouba, Mouldi Bedda and Rafik Djemili, 2006). It is found that MFCC is used widely for feature extraction of speech and GMM and HMM is best among all modeling technique (Anusuya M. A. and Katti S. K., 2009).

In (Bishnu Prasad Das and Ranjan Parekh, 2012) English words spoken by a set of 28 speakers and its corresponding digits zero to nine is recognized. Gaussian MixtureModels are used in (Shashidhar G. Koolagudi and Rao Sreenivasa Krothapalli, 2011), for developing the emotion models. Selecting the right amount of negative examples to build a proper hyperplane is important for training an effective recognition model using an SVM (John H. L. Hansen et al., 2014).

(Vapnik V. and Chapelle O., 2000) derived the expectation of the error boundary for the hyperplane using an SVM, and these error estimation methods were used to tune the kernel parameters (Duan K, Keerthi S, and Poo A, 2003) (Chapelle O et al., 2002). The results in (Bilginer Gulmezoglu et al., 1999), shows that good recognition rates can still be obtained when the sampling frequency of the time-domain samples is reduced from 9.6 kHz to 0.6kHz. In (Hocine Bourouba, Mouldi Beddaand Rafik Djemili, 2006) the authors show that the DTW/GHMM system increases the average recognition rate by 2-10% more than the HMM-based recognition method. At 20dB SNR level the recognition accuracy for the MFCC based system gives 97.03% whereas under same conditions, the LPCC based system gives 73.76% i.e., there is nearly 24% differences in recognition accuracy (Utpal Bhattacharjee, 2013).

In (Sivaprakasam T. and Dhanalakshmi P., 2013), BPNN is very effective recognition method and can accomplish event recognition in a short time and achieve a recognition rate of 91.7%. Though the method proposed in (Hocine Bourouba, Mouldi Bedda and Rafik Djemili, 2006) acheived performance, there are still some issues to be further investigated. If explicit effective features can be extracted, the recognition may have a better performance. The system described in (Bishnu Prasad Das and Ranjan Parekh, 2012) achieves 94% accuracy with isolated digit recognition. Error rates of 23.1 are detected if only MFCC and PLP features are considered separately in (Zolnay A et al., 2005). An accuracy of 91.4% is reported in (Thiang and Wijoyo S., 2011) and 79.5% in (Abushariah A. A. M. et al., 2010).

(Wafa Maitah et. al., 2013) investigated the use of adaptive genetic algorithm (AGA) under vector space model, Extended Boolean model, and Language model in information retrieval (IR). The algorithm comprised of crossover and mutation operators with variable probability and the system resulted in faster attainment of better solutions. A genetic algorithm based feature selection is used by (Ajimi Ameer et. al., 2014) for CBIR system and the authors proved that by fusing multi- feature similarity score the system's retrieval performance improved. (Philomina Simon and Siva Sathya, 2009) described a general framework of information retrieval system and the applicability of genetic algorithms in the field of information retrieval. (Suphakit Niwattanakul et. al., 2013) proposed the similarity measurement method between words by deploying Jaccard Coefficient. Technically, the authors developed a measure of similarity Jaccard with Prolog programming language to compare similarity between sets of data. A question answering system is presented by (Davide Buscaldiba et. al., 2009) based on redundancy and a passage retrieval method that is specifically oriented to question answering. The passage retrieval engine is almost language-independent since it is based on n-gram structures.

(Vikas Thada et. al., 2014) presented a method for finding out the most relevant document among a set of documents for the given set of keyword. Relevance checking is done with the help of Rogers-Tanimoto, MountFord and Baroni-Urbani/Buser similarity coefficients. An effective keyword search method for data-centric extensive markup language (XML) documents is discussed by (LI Guoliang et. al., 2009).

The method divides an XML document into compact connected integral subtrees, called self-integral trees (SI-Trees), to capture the structural information in the XML document. The authors showed that this method costs 10-100 ms to answer a keyword query, and outperforms existing approaches by 1-2 orders of magnitude. A general architecture for ontology-driven data integration based on XML technology is introduced by (Stephan Philippi et. al., 2004) as a proof of concept, a prototypical implementation of this architecture based on a native XML database and an expert system shell is described for the realization of a real world integration scenario.

Among various approaches for expressive speech synthesis (ESS), (D. Govind·S.R. and Mahadeva Prasanna, 2012) presented a system which focused on the development of ESS systems by explicit control. In this approach, the ESS is achieved by modifying the parameters of the neutral speech which is synthesized from the text. Performance evaluation, in a complete speech analysis-synthesis system, has been carried out by (Sandeep Kumar et. al, 2013) for a wavelet-based pitch detection scheme and showed significant results.

(Björn Schuller et. al, 2012) showed that the usage of synthesized emotional speech in acoustic model training can significantly improve recognition of arousal from human speech in the challenging cross-corpus setting. An Auto Associative Neural Network (AANN) based unrestricted prosodic information synthesizer is proposed by (Sudhakar Sangeetha et. al, 2013). The proposed system is applicable to all the languages if the syllabification rule has been changed.

3. Isolated Word Recognition

Speech Recognition system can be separated in different classes by describing the type of utterances they can recognize. In this work an isolated word recognition system is designed for speech recognition.

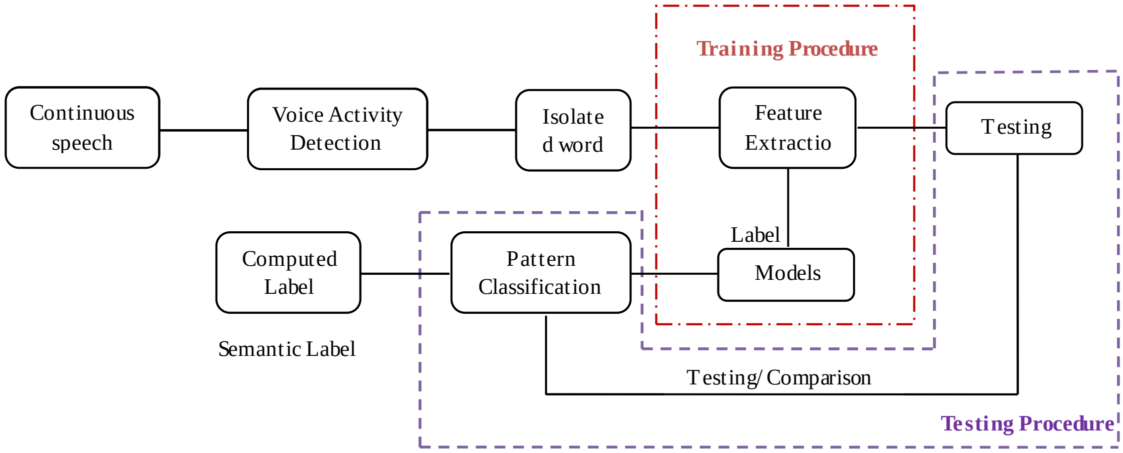


Fig. 10: Framework of proposed system

3.1 Voice Activity Detection

Voice Activity Detection Algorithms are language independent and specifically afford the concerned information regarding where the speech signal is present i.e., it identifies where the speech is voiced, unvoiced or sustained. These details help to deactivate the process during non-speech segment in a speech. It makes the smooth progress of speech processing. Isolated words in an audio speech were exploited using the long pauses in a speech which is shown in Fig. 11. The spectral and temporal envelop of a signal provide maximum information about the signal content. In this work the temporal envelop through RMS energy of the signal is derived for separating/segregating individual words out of the long speeches. Based on the threshold value the temporal environment is analyzed to find the region of words. Isolated words in an audio speech were exploited using the long pauses in a dialog. The spectral and temporal envelop of a signal provide maximum information about the signal content. In this work the temporal envelop through RMS energy of the signal is derived for separating/segregating individual words out of the long speeches. Based on the threshold value the temporal environment is analyzed to find the region of words.

RMS over the window of size which is shown in Eqn. 1. *wlen* is the length of the window. The optimum threshold is chosen through trial and error. In this work 0.5 is chosen as the threshold over the RMS energy window of 20ms. When the energy envelope exceeds the pre-defined threshold value then that sample is marked as the beginning of the segment. Likewise the adjacent sample which falls below the threshold is termed as the end of the segment.

$$RMS = \sqrt{x^2 \otimes wlen} \qquad (1)$$

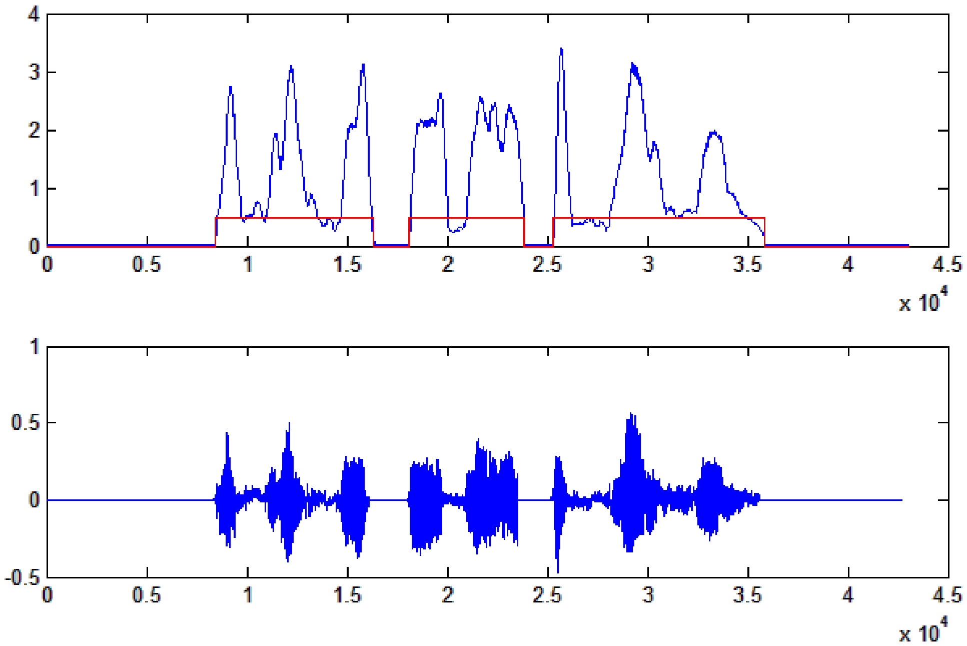


Fig. 11: Sample isolated word separation for the word acoustic feature extraction

Similarly other segment in the energy environment is found by differentiating the threshold signal. Finally, that value is used to extract the isolated words samples from that original speech.

3.2 Acoustic Feature Extraction

In this work LPC, LPCC and MFCC features are extracted.

3.2.1 Linear Prediction Coefficients (LPC)

Linear Prediction (LP) is a mathematical operation which provides an estimation of the current sample of a discrete signal as a linear combination of several previous samples. The theory of Linear Prediction is closely linked to modeling of the vocal tract system, and relies upon the fact that a particular speech sample may be predicted by a linear weighted sum of the previous samples. The number of previous samples used for prediction is known as the order of prediction. The weights applied to each of the previous speech samples are known as linear prediction coefficients (LPC). They are calculated so as to minimize the prediction error. Fig. 12 shows the Block diagram of LPC computation. Any speech sample can be predicted from the past p speech samples which is depicted in Eqn. (2).

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \cdots + a_p s(n-p) \qquad (2)$$

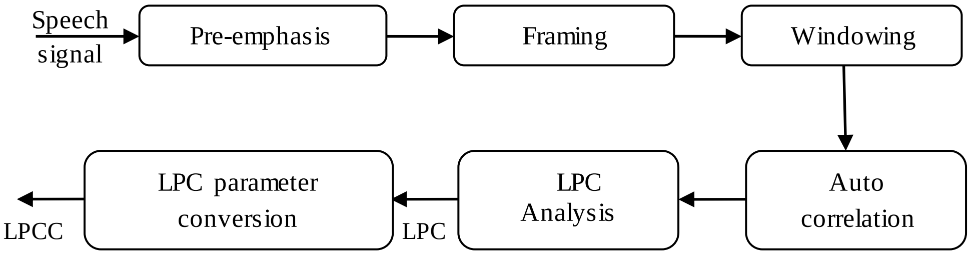


Fig. 12: Block diagram of LPC

3.2.2 Linear Prediction Cepstral Coefficients (LPCC)

The process of extracting LPCC features from an audio signal is summarized as follows: The recursive relation (2) between the predictor coefficients and cepstral coefficients is used to convert the LP coefficients (LPC) into LP cepstral coefficients $\{c_k\}$.

$$\mathrm{c}_0 = \ln \sigma^2$$

$$\begin{aligned} c_m &= a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k\, a_{m-k} \qquad 1 \le m \le p \\ &= \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k\, a_{m-k} \qquad p < m \le d \qquad (3) \end{aligned}$$

Where σ^2 the gain term in the LP analysis and d is the number of LP Cepstral coefficients. A 19 dimensional weighted linear prediction cepstral coefficient (LPCC) for each frame is used as a feature vector.

3.2.3 Mel Frequency Cepstral Coefficients (MFCC)

MFCCs are short-term spectral features and are widely used in the area of audio and speech processing. The mel frequency cepstrum has proven to be highly effective in recognizing the structure of speech signals and in modeling the subjective pitch and frequency content of speech signals. Fig. 13 describes the procedure for extracting the MFCC features.

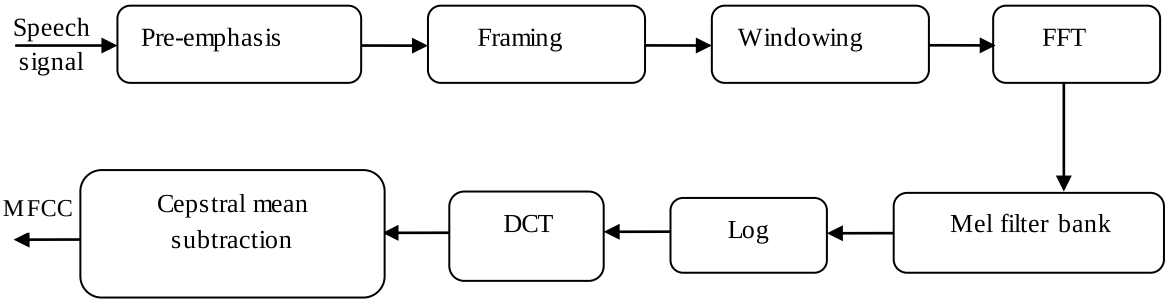


Fig. 13: Block diagram of MFCC

Mel frequency wrapping:

Magnitude spectrum is computed for each of these frames using fast Fourier transform (FFT) and converted into a set of mel scale filter bank outputs. The human ear resolves frequencies non-linearly across the speech spectrum and empirical evidence suggests that designing a front-end to operate in a similar non-linear manner improves performance. A popular solution is therefore filterbank analysis since this provides a much more straightforward route to obtain the desired non-linear frequency resolution. However, filterbank amplitudes are highly correlated and hence, the use of a cepstral transformation in this case is virtually mandatory.

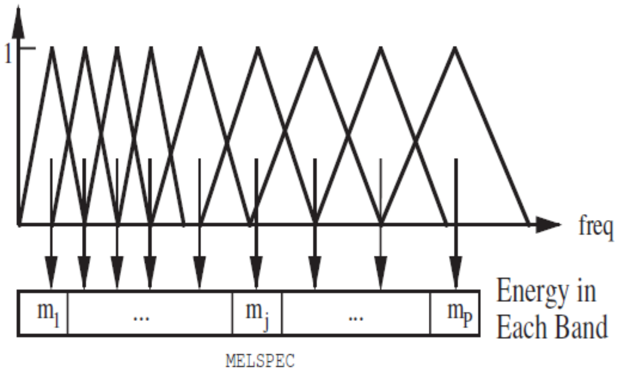


Fig. 14: Mel Scale Filter Bank

A simple Fourier transform based filterbank is designed to give approximately equal resolution on a mel-scale. Fig. 14 illustrates the general form of this filterbank. As can be seen, the filters used are triangular and they are equally spaced along the mel-scale which is defined by

$$Mel(f) = 2595\, log_{10}\,(1 + \frac{f}{100}) \qquad (4)$$

To implement this filterbank, the window of speech data is transformed using a Fourier transform and the magnitude is taken. The magnitude coefficients are then binned by correlating them with each triangular filter. Here binning means that each FFT magnitude coefficient is multiplied by the corresponding filter gain and the results are accumulated. Thus, each bin holds a weighted sum representing the spectral magnitude in that filterbank channel. Normally the triangular filters are spread over the whole frequency range from zero upto the Nyquist frequency. However, band-limiting is often useful to reject unwanted frequencies or avoid allocating filters to frequency regions in which there is no useful signal energy. For filterbank analysis , lower and upper frequency cut-offs can be set. When low and high pass cut-offs are set in this way, the specified number of filterbank channels are distributed equally on the mel-scale across the resulting pass-band.

Cepstrum:

Logarithm is then applied to the filter bank outputs followed by discrete cosine transformation to obtain the MFCCs. Because the mel spectrum coefficients are real numbers (and so are their logarithms), they may be converted to the time domain using the Discrete Cosine Transform (DCT). In practice the last step of taking inverse DFT is replaced by taking discrete cosine transform (DCT) for computational efficiency. The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Typically, the first 39 MFCCs are used as features.

In this work, the 39 MFCC coefficients (c_1, c_2, c_3,..., c_{39}) for a frame are extracted from the speech signal at the segmental level. The 'null' MFCC coefficient c_0 is excluded from the DCT, since it represents the mean value of the input signal which carries little information. The dynamic parameters derived from 13th order static Cepstral coefficients (c_0, c_1, c_2, c_3,..., c_{39}) have been suggested and shown to improve the performance in speech recognition systems. These dynamic features include the delta-cepstrum (the first-order difference of the short-time static cepstrum), the delta-delta-cepstrum (the second-order difference of the static cepstrum), delta and delta-delta-energy. Especially, the dynamic features are verified to be more robust than the static features in noisy conditions. A 39th order MFCC is used to capture the static and dynamic features of speech signal spectrum which contains 13th order static coefficients, 13th order delta coefficients and 13th order acceleration (delta-delta) coefficients. This results in a 39 dimensional MFCC feature vector for each frame.

3.3 Modeling Technique

3.3.1 Support Vector Machines

Support vector machines (SVM) are a kernel-based technique which is based on the principle of Structural Risk Minimization (SRM). SVM constructs a linear model to estimate the decision function using non-linear class boundaries based on support vectors. If the data are linearly separated, SVM trains linear machines for an optimal hyperplane that separates the data without error and into the maximum distance between the hyperplane and the closest training points. The training points that are closest to the optimal separating hyperplane are called support vectors (Chapelle, O. et al., 2002). Through some nonlinear mapping SVM maps the input patterns into a higher dimensional feature space. SVM generally applies to linear boundaries. In some cases linear boundary is inappropriate SVM maps the input vector into a high dimensional feature space. Fig. 15 shows the example for SVM kernel function Φ(X). It maps the transformation of 2 dimension input space (x_1, x_2) into higher three dimension feature space (x_1^2, x_2^2, $\sqrt{}2x_1, x_2$).

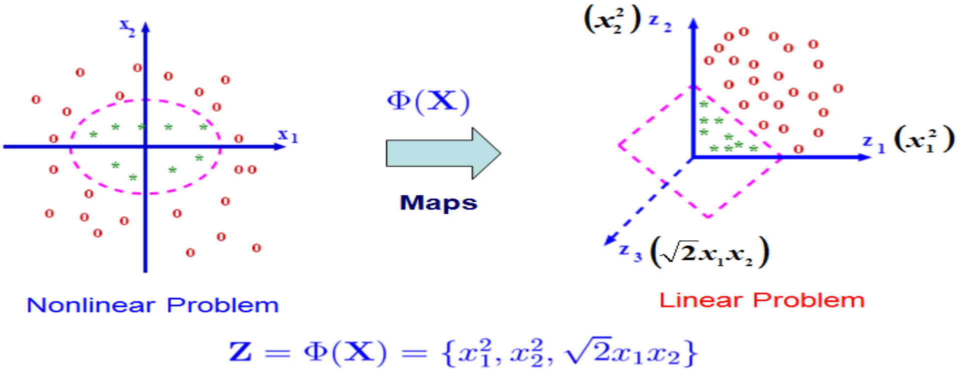


Fig. 15: An Example for SVM kernel function (2D input space into 3D feature space)

The function K is defined as the kernel function for generating the inner products to construct machines with different types of non linear decision surfaces in the input space. The kernel function may be any of the symmetric function that satisfies the Mercer's conditions. Fig. 16 shows the architecture of SVM. There are several SVM kernel functions as given in Table 1.

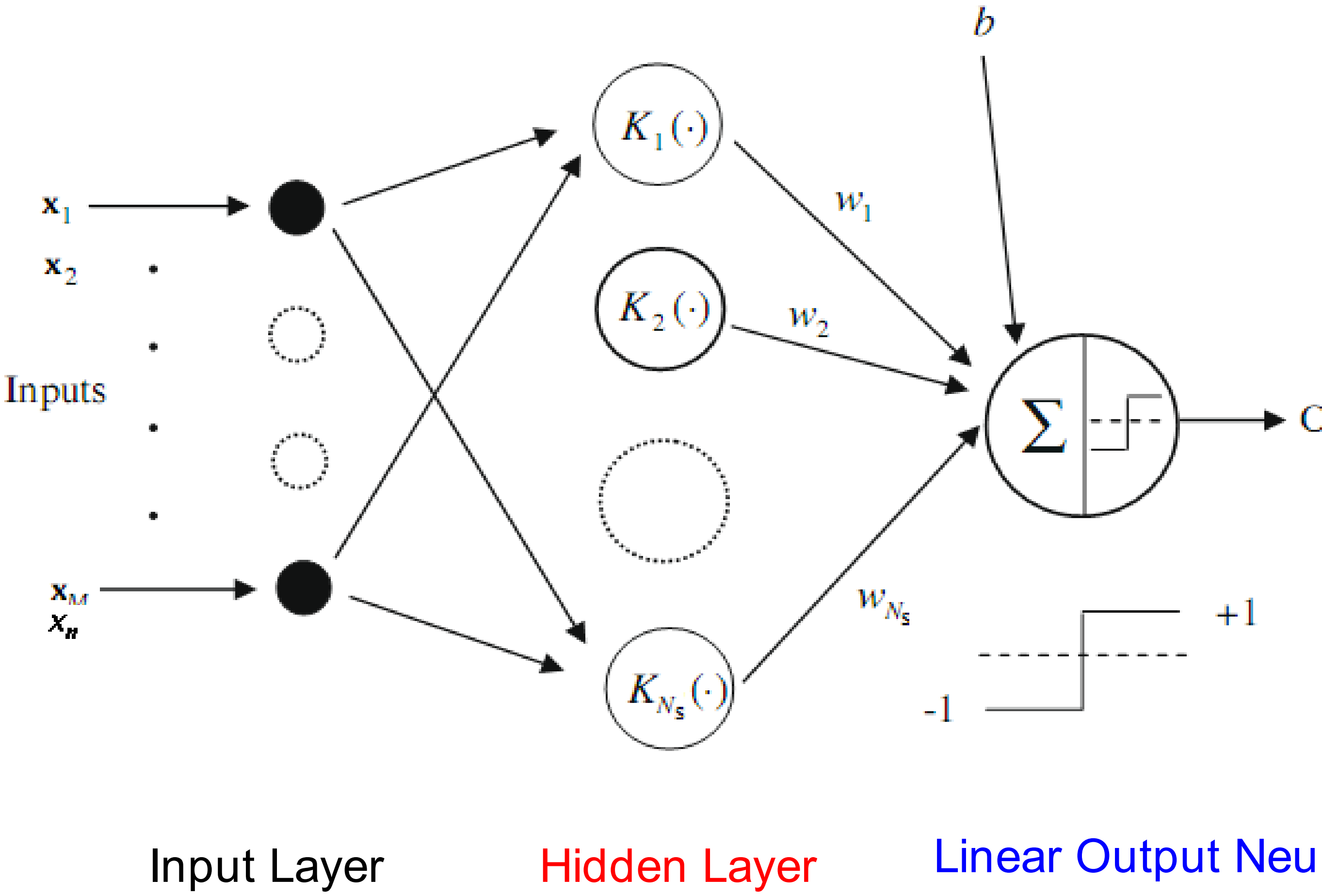


Fig. 16: Architecture of SVM

Where,

$x = (x_1, x_2, \ldots, x_n)$ is input pattern (feature vector),

N_s is the number of support vectors,

$K_1(.)$, $K_2(.)$,…, $K_{Ns}(.)$ are kernel functions,

$w_1, w_2, \ldots\ldots w_{Ns}$ are weights from hidden layer to output, b is a bias.

$y = (+1$ or $-1)$ is the output of SVM model.

A linear SVM is used to classify data sets which are linearly separable.SVM finds a separating hyperplane which separates the data with the largest margin. The discriminant function is given by following Eqn. (5)

$$g(x) = w^t x + b \qquad (5)$$

such that g $(x) \geq 0$ for $y = +1$ and g $(x) < 0$ for $y = -1$. SVM learns an optimal separating hyper plane from a given set of positive and negative examples. It minimizes the structural risk, that is, the probability of misclassifying yet-to-be-seen patterns for a fixed but unknown probability distribution of the data. This is in contrast to traditional pattern recognition techniques of minimizing the empirical risk, which optimizes the performance on the training data.

Table 1: Types of SVM Inner Product Kernels

Types of SVM	Inner Product Kernel K(x, x_i)	Details
Polynomial Kernel	$(x^T x_i+1)^P$	Where x is input patterns, x_i is support vectors, σ^2 is variance, i -1,2,…..,Ns, Ns is number of support vectors,β_0 β_1-constant values
Gaussian Kernel (Radial bias Function)	$\exp\left(-\frac{\lVert x\text{-}x_i\rVert^2}{2\sigma^2}\right)$	
Sigmoidal Kernel	tanhs($\beta_0 x^T x_i+\beta_1$)	

3.3.2 Gaussian Mixture Model

Gaussian mixture model (GMM) is a powerful statistical tool which is widely used in pattern recognition. GMM has been used successfully for speaker identification in recent years. Gaussian mixture model (GMM) is a mixture of several Gaussian distributions and can therefore represent different subclasses inside one class. Fig. 17 shows the mixture of two Gaussians. The probability density function is defined as a weighted sum of Gaussians. GMM can be viewed as one component of HMM under certain circumstances.

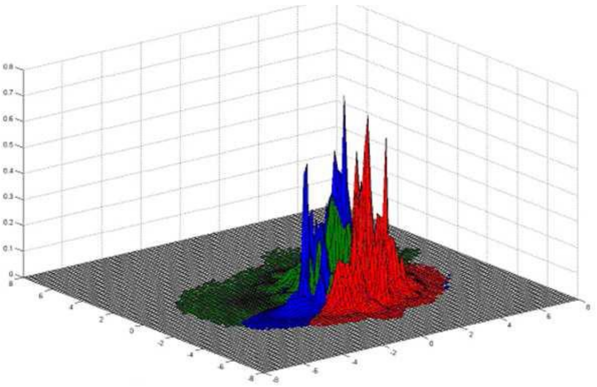


Fig. 17: Gaussian mixture model

The probability distribution of feature vectors is modeled by parametric or nonparametric methods. Models which assume the shape of probability density function are termed parametric. In non-parametric modeling, minimal or no assumptions are made regarding the probability distribution of feature vectors. The potential of Gaussian mixture models to represent an underlying set of acoustic classes by individual Gaussian components, in which the spectral shape of the acoustic class is parameterized by the mean vector and the covariance matrix, is significant. Also, these models have the ability to form a smooth approximation to the arbitrarily shaped observation densities in the absence of other information. The basis for using GMM is that the distribution of feature vectors extracted from a class can be modeled by a mixture of Gaussian densities as shown in Fig. 17. For a D dimensional feature vector ***x***, the mixture density function for category is defined as

$$p\left(\frac{x}{\lambda^s}\right) = \sum_{i=1}^{m} \alpha_i^s \, f_i^s(x) \qquad (6)$$

The mixture density function is a weighted linear combination of m component unimodal Gaussian densities $f_i^s(.)$. Each Gaussian density function $f_i^s(.)$ is parameterized by the mean vector μ_i^s and the covariance matrix using $\sum_i^s$ using

$$f_i^s(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi^d \, |\sum_i^s|}} \; \exp(-\frac{1}{2}(\boldsymbol{x} - \mu_i^s)^T \left(\sum_i^s\right)^{-1}(\boldsymbol{x} - \mu_i^s)) \qquad (7)$$

where $(\sum_i^s)^{-1}$ and $|\sum_i^s|$ denotes the inverse and determinant of the covariance matrix $\sum_i^s$, respectively. The mixture weights $(\alpha_1^s, \alpha_2^s, \alpha_3^s, \ldots, \alpha_m^s)$ satisfy the constrain $\sum_{i=1}^{m} \alpha_i^s = 1$. Collectively, the parameters of the model λ^s are denoted as $\lambda^s = \{\alpha_i^s, \mu_i^s, \sum_i^s\}$, i=1,2,...,m. The number of mixture components is chosen empirically for a given data set. The parameters of GMM are estimated using the iterative expectation maximization algorithm.

Steps involved in implementing GMM:

Training Procedure:

Training procedure involves Expectation step and Maximization procedure for selecting the gaussian.

E-Step:

1. K-means clustering is used for initializing the seed mean vector.
2. The expected mean and co-variance are computed.

M-Step:

1. Based on the probability density function the expected mean and covariance are maximized to generate the new mean 7 covariance.
2. The process is iterated until the gaussian is perfectly fitted or some stopping criteria are reached.

Testing Procedure:

1. Probability density function for each and every frame in the isolated word is computed against every model. The most distributed model is considered as a winner.
2. Based on the freq of winning gauss models the most occurring model is considered as the gaussian representing that word's feat vector.

3.3.3 Hidden Markov Model

Hidden Markov Model (HMM) is powerful statistical tool which is widely used in pattern recognition. Especially, the HMM has been developed extensively in speech recognition system over the last three decades. There are two main reasons for choosing HMM in speech processing. First, the transition and duration parameters in HMM may properly reveal the evolution of features over time, which is very important in modeling speech/audio perception. Second, there are many kinds of variations of the HMM as well as experiences of using them which are developed in speech recognition researches. This makes HMM a mature technique to be applied in this research. A Hidden Markov Model is considered as a generalization of a mixture model where the hidden variables control the mixture components has to be selected for each observation, these observations are related through a Markov process rather than independent of each other (HocineBourouba, MouldiBedda, RafikDjemil, 2006). Hidden Markov Model assumes that successive acoustic features of a spoken word are state independent. The occurrence of one feature is independent of the occurrence of the others.

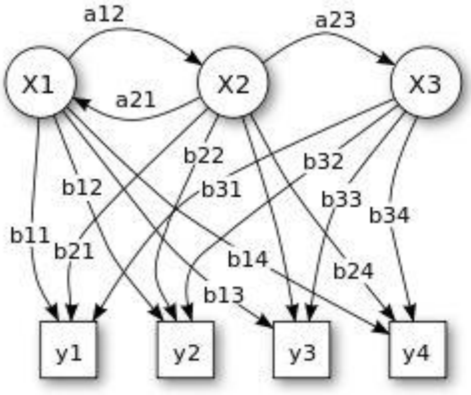


Fig. 18: Probabilistic parameters of a hidden Markov model (example)

Markov Model is a stochastic model with finite state automaton in which the sequence of states is a Markov chain. Each Markov Model corresponds to a deterministic event, whereas each output of HMM corresponds to probabilistic density function; the generating state sequence of HMM is hidden. Probability starts with a particular event. Markov model is defined as, the states represent possible event types (e.g., the different words in this example) and the transitions represent the probability of one event type following another. It is easy to depict a Markov model as a graph. HMM for Speech Recognition are typically an interconnected group of state. According to an emission probability density function, each state is assumed to emit a new feature vector for each individual frame. Each new observation frame can be associated with any state. However, the topology of the HMM and the associated transition probabilities provide temporal constraints.

In Fig. 18, x_1, x_2, x_3 are the states of Hidden MarkovModel. y_1, y_2, y_3 and y_4 are the possible observations a_{12}, a_{21}, a_{23} are state transition probabilities and output probabilities are mentioned as b. This is for a general description for 3 states. A hidden markov model for discrete symbol observations is characterized by the following parameters

1. N, the number of states in the model. Individual states are labeled as 1, 2, ...,N, and denote the state at time t as q_t.
2. M, the number of distinct observations symbols in all states, i.e., the discrete alphabet size. Here individual symbols are denoted as $V = v_1, v_2, ..., v_M$
3. The state-transition probability distribution $A = \alpha_{ij}$ where,
$$\alpha_{i,j} = P[q_{t+1} = j | q_t = i], \qquad 1 \le i,j \le N \qquad (8)$$
4. The observation symbol probability distribution $B = b_j(k)$, in which

$$\alpha_{jj}(k) = P[X_t = \upsilon_k | q_t = j], \qquad 1 \leq k \leq M \tag{9}$$

defines the symbol distribution in state j, $j = 1, 2, \ldots, N$.

5. The initial state distribution π_i,

$$\pi_i = P[q_1 = i], 1 \leq i \leq N \tag{10}$$

Thus, complete specification of an HMM includes two model parameters, N and M, the observation symbols, and the three sets of probability measures A, B, and π.

$$\Lambda = (A, B, \pi) \tag{11}$$

Eqn. 11 shows the compact notation, which is used to indicate the complete parameter set of the model.

It is used to define a probability measure for the observation sequence X, i.e., $P = (X \,|\, \Lambda)$, which can be calculated according to a forward procedure as defined below.

Consider the forward variable α_t(i) defined as

$$\alpha_t(i) = P(x_1, x_2, \ldots, x_t, q_t = i \,|\, \Lambda) \tag{12}$$

which is the probability of the partial observation sequence $x_1, x_2, \ldots, x_t$ and state

Steps involved in implementing HMM:

Training Procedure:

1. Transition probability and emission probability is randomly initialized based on the observation (o=19)
2. Based on the randomized emission and transition probability the model is trained for the given word feature vector.
3. The new transition and emission probability are estimated based on the word vector.
4. The HMM for the individual word is represented by the parameters namely mean, co-variance, mixmat, emission, transition and LL.

Testing Procedure:

1. Log likelihood ratio for each and every frame in the isolated word is computed against every model. The most likely model is considered as a winner.
2. Based on the frequency of winning HMM models the most occurring model is considered as the markov model representing that word's feat vector.

3.4 System Evaluation

3.4.1 Database

Under controlled environment, domain restricted speech sentences are recorded at 16 kHz sampling frequency. Speech database is collected from 50 visually impaired persons to gather the knowledge about their personal search interest as well to improve the efficiency of the system. A database of 2700 different queries and 3600 isolated words are recorded from normal (36 male and 32 female) and visually impaired persons (22 male and 28 female). Each speech clip consists of 1 to 2 minutes duration. Hence, utterances of individual words ranges around 1-2 seconds. Table 2 depicts the database creation for Isolated word recognition/Speech Recognition. Subsequently recorded sentences were segregated into isolated word with the help of VAD. Table 3 shows the recorded isolated word samples relevant to various domains. In Table 3, an isolated word (domain) relevant to speech processing is assumed as D_1, pattern recognition as D_2, image processing as D_3, medical processing as D_4 and preposition in D_1 to D_4 are considered as P respectively.

Table 2: Speech Recognition Database Description

Total Number of Speakers	:	118
No. of normal speaker	:	68
No. of impaired persons	:	50
No. of isolated words recorded	:	3,600
Number of Sentences Recorded	:	2,700
No. of Isolated word in Recorded Sentences	:	7,400
Total no. of isolated word	:	11,000

Table 3: Recorded word samples relevant to various domains

Terms	Speech Proc.	Pattern Recogn.	Image Proc.	Medical Proc.	Prepositions
No. of Isolated Words	1910	1780	1550	1740	400

3.4.2 Feature extraction

In this work the pre-emphasized signal containing the continuous speech is taken for testing. Through VAD the isolated words are extracted from the sentences. From which unvoiced excitations present in the frames are removed by thresholding the segment size (100ms). Features such as LPC, LPCC and MFCC are extracted from each frame of size 320 window with an overlap of 120 samples. Thus it leads to 14 LPCs, 19 LPCCs and 39 MFCCs respectively which are used individually to represent the isolated word segment. During training process each isolated word is separated into 20ms overlapping windows for extracting 14 LPCs, 19 LPCCs and 39 MFCCs features respectively.

3.4.3 Modeling

System may differ based on the stored size of speech units. For specific usage of domains, the storage of entire words or sentences allows the system for high quality outcomes. In this work, words or sentences related to Speech processing, Pattern Recognition, Image Processing, Medical Processing and finally prepositions were stored in the database. By providing various initialization techniques for each modeling technique, the overall recognition accuracy is increased. In GMM, components in Gaussians are varied and results are analyzed. Similarly, Kernels in SVM and Gaussians in HMM are also analyzed.

3.5 Experimental Results

Using VAD isolated words in a speech is separated as discussed in Section 3.1. N SVMs are created for each isolated word. For training, 4,850 isolated words were considered. Hence, this results in 4,850 feature vectors each of 14 dimensional LPC, 19 dimensional LPCC and 39 dimensional MFCC for 4,850 isolated words respectively. The training process analyzes speech training data to find an optimal way to classify speech frames into their respective classes. The derived support vectors are used to classify speech data. For testing, 6,150 isolated words were considered. During testing, 14 dimensional LPC, 19 dimensional LPCC and 39 dimensional MFCC feature vectors (1 sec of speech file) are given as input to SVM model and the distance between each of the feature vectors and the SVM hyperplane is obtained. The average distance is calculated for each model. The text corresponding to the query speech is decided based on the maximum distance. The same process is repeated for different query speech, and the performance is studied. The performance of SR for different kernels: Polynomial, Gaussian and Sigmoidal are compared for LPC, LPCC and MFCC acoustic features. From the exhaustive analysis, Gaussian kernel function in SVM using MFCC features provides better performance when compared to others, which is shown in Table 4. Word Error Rate (WER) and Word Accuracy (W_{acc}) for the speech recognition system using SVM with MFCC features is shown in Fig. 19.

Table 4: Different Kernel functions in SVM using LPC, LPCC and MFCC features

Types of kernels	Polynomial (in %)	Gaussian (in %)	Sigmoidal (in %)
LPC	76.09	79.23	78
LPCC	80.59	82.37	81.23
MFCC	82.93	84.56	83.21

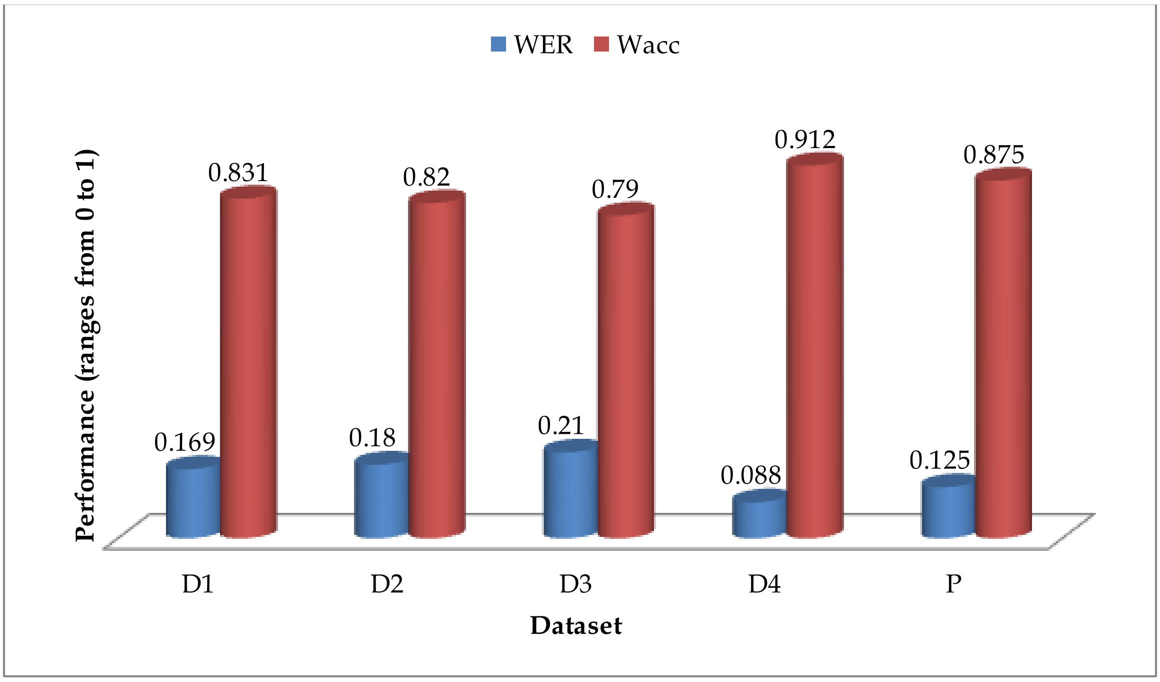


Fig. 19: Performance of gaussian kernel function in SVM using MFCC features

In GMM the database comprises of polysyllabic terms that leads to fitting of each syllable to individual components. Thus component setting of 6 or more provide better accuracy than others and their WER and W_{acc} is shown in Fig. 19. Based on the perceptual characteristics of the polysyllabic words the system efficiently recognize each term. Various components in GMM using LPC, LPCC and MFCC features are analyzed and are shown in Table 4.

Table 4: Various components in GMM using LPC, LPCC and MFCC features

No. of Mixtures	**2**	**4**	**6**	**8**
LPC	74.26	79.68	81.73	81.73
LPCC	77.82	87.16	89.29	89.29
MFCC	79.07	92	93.74	93.74

The number of Gaussian mixtures is increased from 2 to 10. The performance in terms of classification accuracy is studied. When the number of mixtures is 2, the performance is very low. When the mixtures are increased from 2 to 6, the classification performance slightly increases. When the number of mixtures varies from 6 to 10, there is no considerable increase in the performance and the maximum performance is achieved. There is no considerable increase in the performance when the number of mixtures is above 10. With GMM, the best performance is achieved with 6 Gaussian mixtures. Performance of GMM using MFCC features is depicted in Fig. 20.

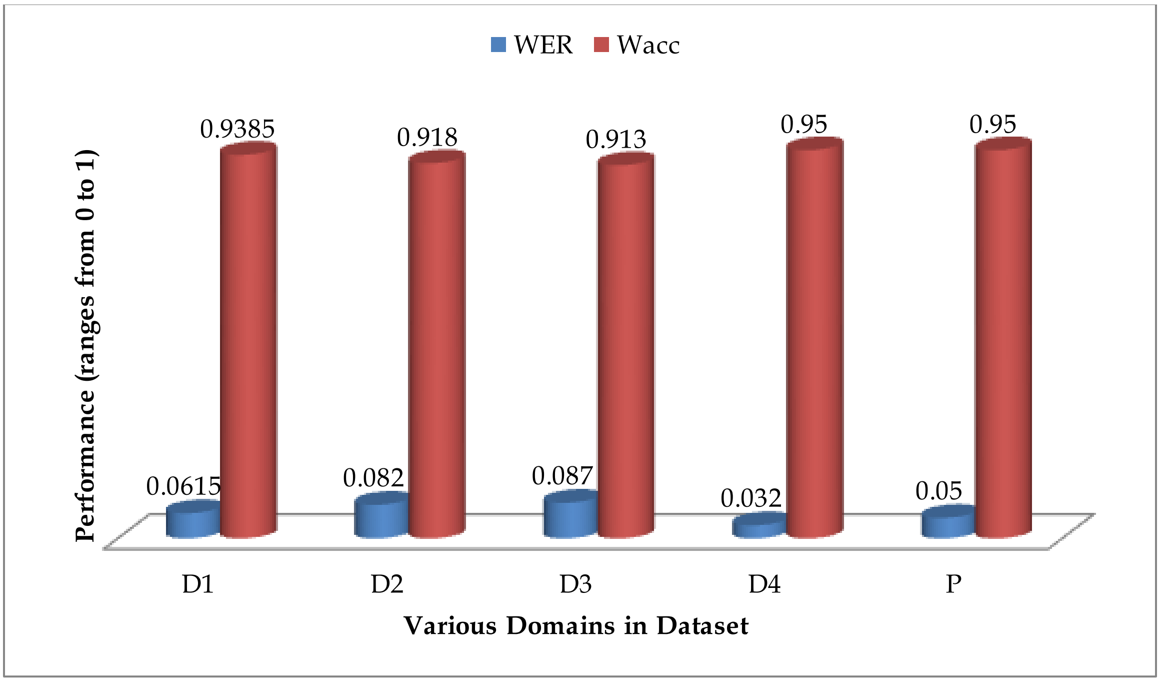


Fig. 20: Performance of GMM using MFCC features

In Hidden Markov Model each unit speech utterance is mapped on to states of HMM. Based on the given speech input the individual models in the HMM is matched to find the path. The path which follows maximum probability is selected as the appropriate model. States in HMM is varied and its performance is analysed. 4 state in HMM provides better results when compared to other and the results are premium which are depicted in Fig 21.

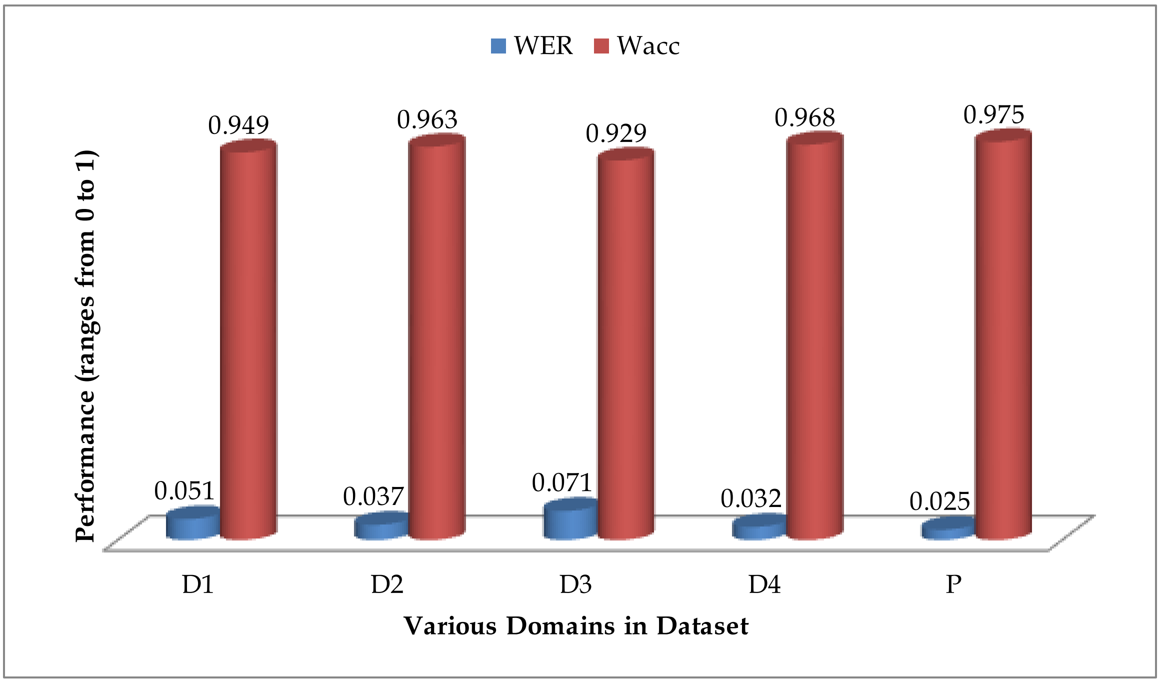


Fig. 21: Performance of HMM with 4 States

Performance of HMM with 4 state markov model is shown in Fig. 21. Word Error Rate (WER) for three modeling techniques is depicted in Fig. 22. Fig. 23 shows overall performance of the speech recognition system using SVM, GMM and HMM with its acoustic feature (MFCC).

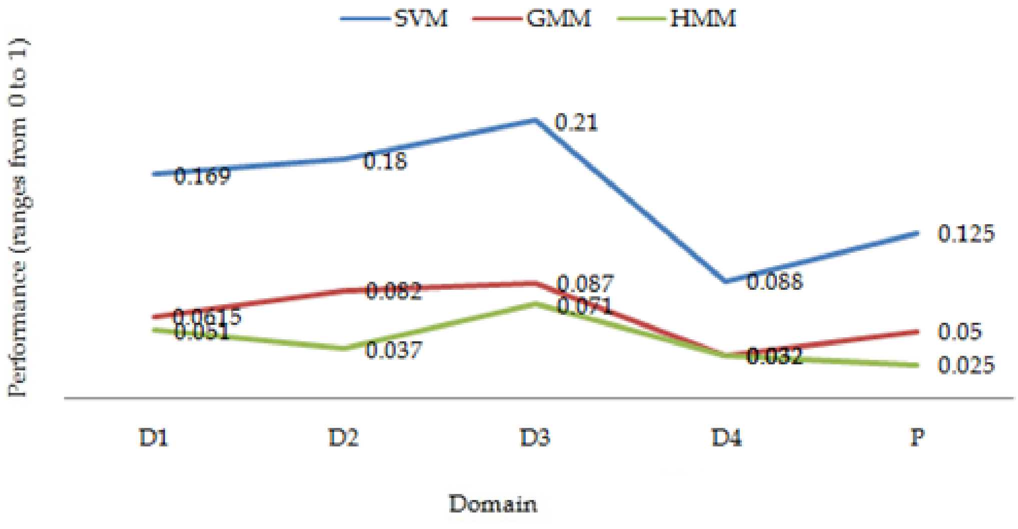


Fig. 22: Comparison of WER for SVM, GMM and HMM Techniques

From the exhaustive analysis of the above modeling technique and acoustic feature, Hidden Markov Model with MFCC performs better results than other modeling technique From Fig. 24, it is proven that the results of Hidden Markov Model provide good results.

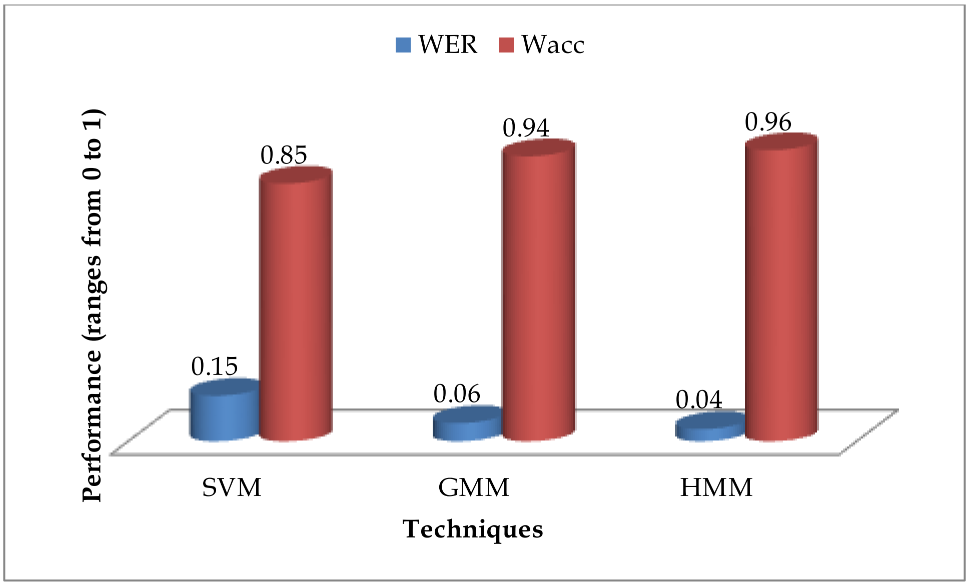
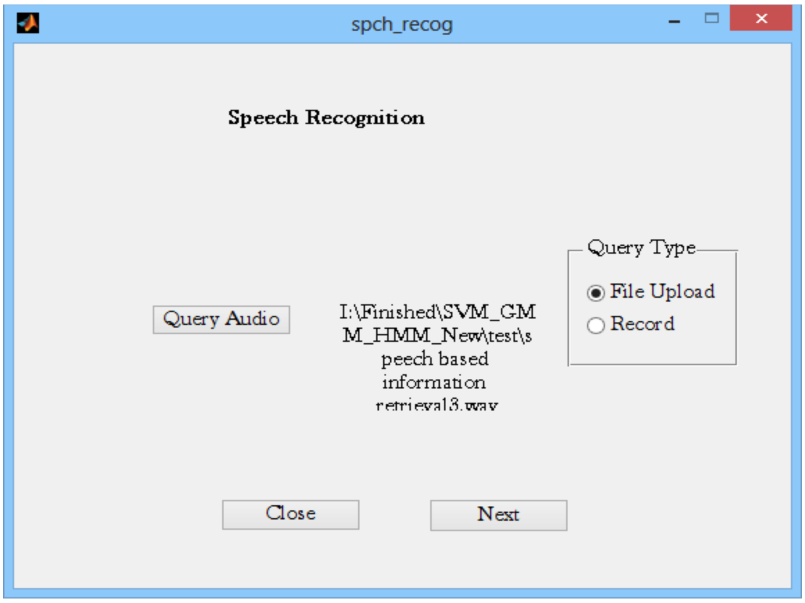


Fig. 23: Overall Performance of the SR system (isolated word recognition)

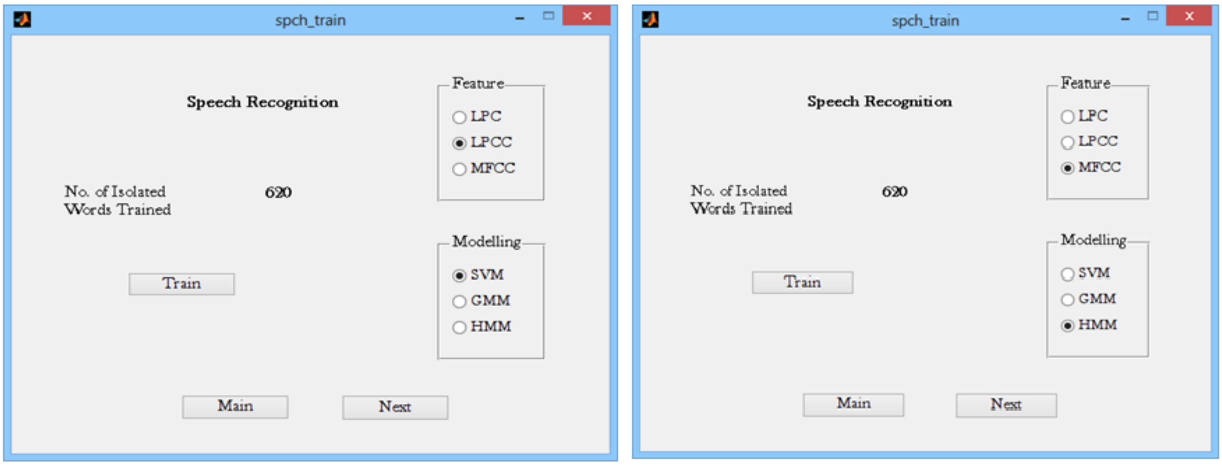
For the chosen domain and database, speech recognition system generates WER as 0.15, 0.06 and 0.04 for SVM, GMM and HMM respectively. An average accuracy of 84.56% for SVM, 93.74% for GMM, and 95.68% for HMM is obtained respectively. It is evident that, HMM shows better results while comparing to other modeling techniques. Snapshots of isolated word recognition system in various stages with various techniques and features are depicted in Fig. 24, 25 and 26 respectively.



I:\Finished\SVM_GMM_HMM_New\test\speech based information retrieval3.wav

● File Upload
○ Record

Fig. 24: Querying speech input by the user



spch_train

Speech Recognition

No. of Isolated Words Trained 620

Feature
○ LPC
● LPCC
○ MFCC

Modelling
● SVM
○ GMM
○ HMM

spch_train

Speech Recognition

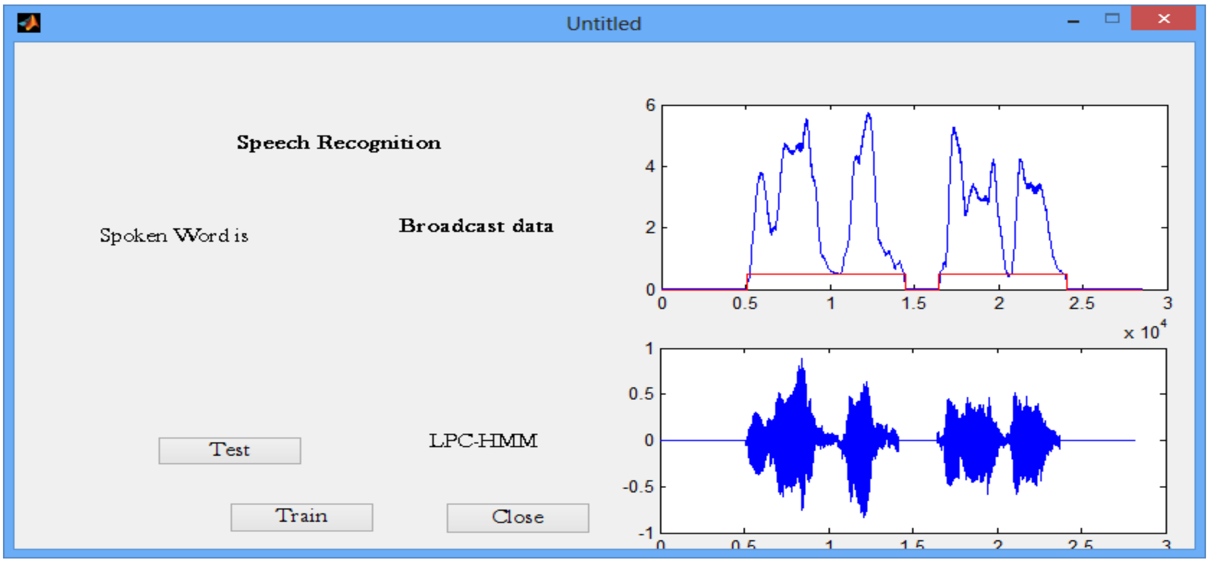
No. of Isolated Words Trained 620

Feature
○ LPC
○ LPCC
● MFCC

Modelling
○ SVM
○ GMM
● HMM

Fig. 25(a) Fig. 25(b)

Fig. 25: Selection of different modeling technique with different acoustic feature for training isolated word. a) SVM Modeling technique with LPCC features. b) HMM modeling technique with MFCC features



Untitled

Speech Recognition

Spoken Word is Broadcast data

LPC-HMM

Fig. 26: Snapshot of the speech Recognition System using LPC features for HMM Technique

3.6 Summary

A system has been developed to convert spoken word into text using SVM, GMM and HMM. Acoustic features namely LPCC and MFCC are extracted to model the words. Voice Activity Detection (VAD) is used for segregating individual words. Features for each isolated word are extracted and those models are trained successfully. In testing phase, each isolated word segment from the test sentence is matched against these models for finding the semantic representation of the test input dialogue. SVM shows an accuracy of 84.56% for MFCC, GMM shows an accuracy of 93.74% for 4th component using MFCC and HMM shows an accuracy of 95.68% for 4 state markov models. As a result HMM provides optimum results while compared with other modeling techniques.

4. Query based Text Document Retrieval

Most of the data/text mining tasks use Information Retrieval (IR) methods to preprocess text documents. These methods are quite different from traditional preprocessing methods used for relational tables. IR helps users to find the needed information from the huge collection of data that matches their requirements. This work aims to implement vector space model to represent text document. In boolean method, search terms are logically combined using AND, OR and NOT (boolean operators). Boolean retrieval method is based on exact match approach, in which the system retrieves each document that makes the query logically true and ignores the others. Ontology based approach retrieves the document on the base of entity/search term relationship as well as interrelationship. In general this approach relates the query term with another term based on the relationship. In boolean retrieval and ontology approach, retrieval results are usually quite poor because these systems do not consider term frequency of the search term. VSM model is considered as best model for information retrieval because of its angle projection between the similar terms, and tf-idf weighting. Even for very low similarities between the term and the document, this method results in a smallest angle which was not efficiently projected by other methods like Euclidian distance.

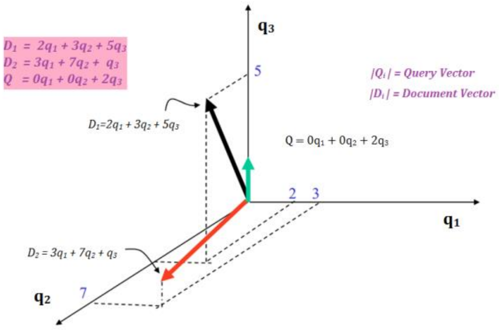
In this work ontology based VSM and Genetic Approach are proposed for increasing the efficiency of the retrieval system. The user queries ranges from multi-sentence description of information need to a few words, and so search engines are in need of mapping the inflectional variants into root and also it identifies the rare term present in the user query. This process is termed as stemming and stopword removal which is done in preliminary stage. These preprocessed queries are further reformulated by the ontological method. Later these reformulated queries are considered for computing the similarity between the document and the query. Section 4.1 and 4.2 describes the implementation of Ontology assisted VSM approach. In Genetic Algorithm, genotype representation of document and query are considered and as discussed in Section 4.3.

4.1 Vector Space Model

Vector Space Model (VSM) is also known to be as term vector model. The major applications of VSM are information filtering, indexing, information retrieval and relevancy rankings. Search term may be of single word, keywords, query or longer phrases (Turney, P. and Pantel, P., 2010). If the chosen terms are words, then the number of words in the vocabulary decides the dimensionality of the vector. In other words, the numbers of distinct words occur in the corpus (Louis S. Wang, 2009). In the classic vector space model, the term-specific weights in the document vectors are products of local and global parameters. This model is known as term frequency-inverse document frequency (tf-idf) model (Xindong Wu et al., 2008). Vector Space Model is the convenient and effective way of ranking documents. It is analyzed by measuring how close their vectors are to a query vector (Belkin and Marchetti, 2001). If q_i is the given input search term and D_i is the collection of document where i ranges from 0,1,...,n, then the vector matrix is as given as follows:

$$\begin{pmatrix} & q_1 & q_2 & q_3 & \cdots & q_n \\ D_1 & d_{11} & d_{12} & d_{13} & \cdots & d_{1n} \\ D_2 & d_{21} & d_{22} & d_{23} & \cdots & d_{2n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ D_n & d_{n1} & d_{n2} & d_{n3} & \cdots & d_{nn} \end{pmatrix}$$

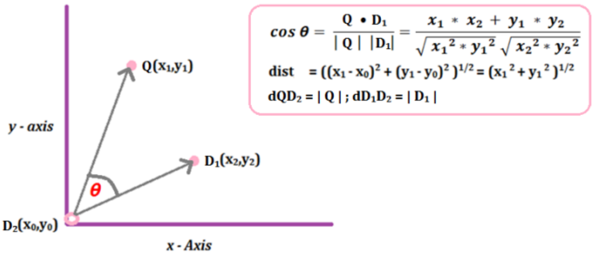
In Fig. 27 the VSM model with three geometric text, $D_1=2q_1+3q_2+5q_3$, $D_2 = 3q_1+7q_2+q_3$ and $Q = 2q_3$ are described. For instance, consider the point $Q(x_1, y_1)$ represents a query and points $D_1(x_2, y_2)$, $D2(x_0, y_0)$ etc., represents documents.



$D_1 = 2q_1 + 3q_2 + 5q_3$
$D_2 = 3q_1 + 7q_2 + q_3$
$Q = 0q_1 + 0q_2 + 2q_3$
q3
5
$|Q_i| =$ Query Vector
$|D_i| =$ Document Vector
$D_1 = 2q_1 + 3q_2 + 5q_3$
$Q = 0q_1 + 0q_2 + 2q_3$
2
3
q1
$D_2 = 3q_1 + 7q_2 + q_3$
q2
7

Fig. 27: VSM approach

The cosine angle between Q (the query), with each document is computed and are sorted in decreasing order based on its cosine angles. The cosine angle between document D_i and query Q is shown in Fig. 28. The same procedure is to be followed for the entire collection of documents with query.



$cos\, \theta = \frac{Q \bullet D_1}{|Q|\ |D_1|} = \frac{x_1 * x_2 + y_1 * y_2}{\sqrt{x_1{}^2 * y_1{}^2}\sqrt{x_2{}^2 * y_2{}^2}}$
dist $= ((x_1 - x_0)^2 + (y_1 - y_0)^2)^{1/2} = (x_1{}^2 + y_1{}^2)^{1/2}$
$dQD_2 = |Q|$; $dD_1D_2 = |D_1|$
$Q(x_1,y_1)$
y - axis
$D_1(x_2,y_2)$
θ
$D_2(x_0,y_0)$
x - Axis

Fig. 28: An example for computing the cosine angle

4.2 Ontology assisted Vector Space Model

Ontology based vector space model is an effective method for retrieving the ranked document based on the required query given by the user. This is achieved by computing the cosine similarity between the reformulated queries based on ontological approach with documents in the database. Ontological approaches are best known technique for query reformulation. Ontology based query reformulation technique retrieves the document on the base of entity/search term relationship as well as interrelationship between the huge collections of document (Nadia, L. and 2014). In this work document retrieval system is achieved by three stages: keyword extraction, query re-formulation and the similarity measures (based on cosine angle) between the documents and the query respectively. Fig. 29 depicts the framework of the retrieval system based on VSM.

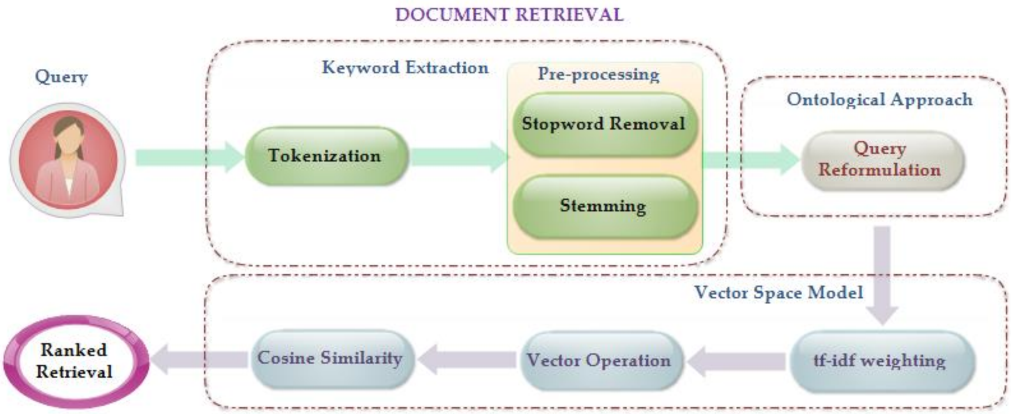


Fig 29: Framework for the Document Retrieval System

4.2.1 Keyword Extraction

Keyword extraction is done in two steps: document linearization, text pre-processing (stopword removal and stemming).

Document Linearization

Document Linearization or Tokenization is the process of normalizing the document, in which all the text documents are parsed, lower cased or upper cased (depends on user preference) and punctuations and special symbols are removed (Ting KM., 2002). In this work the documents are lineared by parsing the text and these parsed texts are converted into lower case sentences for normalizing the collection of document as well as the given query.

Text Pre-processing

Stopword removal: -

Rare terms are more informative than frequent terms (BjörnSchuller, 2012). Terms like is, a, the, an, and, of, for, with, etc., will occur most frequently in the collection of documents. It is very essential to discard these frequent terms from the search query for computing the score especially in large collection of databases. Stop words may confuse the retrieval system; (Carlson, R., & Granström, B., 2007) these are not useful for searching or text mining which results in 20 - 30% of total word count. If these terms are also considered as search terms it accounts in many numbers of occurrences than other term. Even if the document does not contain the required query, the document may be misclassified as relevant with these frequent terms. The computing time, indexing file size can be reduced and efficiency can be increased by removing these stopwords. For developing an application, additional domain specific stopwords list can also be constructed.

Stemming:-

Stemming is the technique used to find out the root/stem of a word used to improve the effectiveness of the system (Belkin and Marchetti, 2001). Stemming is the process of mapping inflectional variants to root (e.g. see, sees, seen, saw -> see). In such system, combining roots with same roots may reduce indexing size as much as possible which is nearly equivalent to 40 – 50 %. There are certain procedures like remove ending, transform words etc., to be followed for mapping inflectional variants into root. It improves the recall measures of the information retrieval system. These extracted keywords are reformulated using ontological approach.

4.2.2 Weighting Scheme in Information Retrieval

Weighting scheme plays a vital role in vector based information retrieval. These can be categorized by three common schemes local weighting (*tf*), global weighting (*idf*) and *tf-idf* weighting (Björn Schuller, 2012). Local weighting is the process of counting the occurrence of query *q* within each single document. Global weighting is the method of computing the number of documents containing the term over *n* number of document collection in the database. The system is in need of defining the number of document containing the term as well as the number of occurrence in each document for computing the most similar documents. Hence most of the retrieval system uses *tf-idf* weighting.

The weighting based on *tf-idf* is the best way to convert the textual representation of information into a VSM (sparse features) (Björn Schuller, 2012). *Tf-idf* weight of a search term (or query) is the product of its local weight (*tf*) and its global weight (*idf*). *Tf-idf* weighting is a best known weighting scheme in information retrieval and machine learning.

$$w = tf * idf \tag{13}$$

Weighting factor *w* is the fundamental factor for computing the query vector $|Q|$ and document vector $|D_i|$ which is depicted in Section 4.2.3

4.2.3 Computing Cosine Similarity based Ranking

The similarity of the document vector $|D_i|$ to a query vector $|Q|$ is equivalent to the cosine of the angle between $|D_i|$ and $|Q$. Cosine similarity with *tf-idf* weighting is the most common term weighting approach for vector space model retrieval approach (Xindong Wu et al., 2008). Eqn. 14 shows the cosine angle between the query *q* with Documents (1,2,...,n).

$$\text{Cosine}\ \theta_{\mathrm{D_i}} = \frac{\mathrm{Q} \bullet \mathrm{D_i}}{|\mathrm{Q}| * |\mathrm{D_i}|} \tag{14}$$

Where,

Q – Query term

D_i = Document (i=1,2,...,n)

$|Q|$ = Query Vector (similarity between query term)

$|D_i|$ = Document Vector (similarity between document)

The similarity analysis between the query term $|Q|$ and document $|D_i|$ for *n* number of documents are shown in Eqn. 15 and 16:

$$|\mathrm{Q}| = \sqrt{\sum \mathrm{w}_{\mathrm{Q}}^{2}} \tag{15}$$

$$|\mathrm{D_i}| = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{w}_{\mathrm{i}}^{2}} \tag{16}$$

Where,

w_Q – weighting factor of query term

w_i – weighting factor of Document term (i range from 1 to n number of documents)

4.3 Genetic Algorithm

Genetic algorithms are a powerful searching mechanism known for its robustness and quick search capabilities. There are three major components essentially taken care while designing Genetic Algorithm. The first component is coding the problem solutions, next is to find a fitness function which has to optimize the performance and finally, the set of parameters including the population size, genetic operators and their percentages.

Genetic Approach (GA) is a relevance feedback approach and is a search heuristic that mimics the process of natural selection (Goldberg, D.E.. 2003). The evolutionary process begins from a population of randomly generated individuals with the population in each iteration called as generation. Each individual has a set of properties like chromosome information or genotype which can be altered or muted traditionally from which solutions are represented in binary as strings of 0s and 1s. The major objective behind the application of genetic algorithm in this work is to find important feature elements that contribute more to classifier for distinguishing one word from others. At the same time, the number of feature elements is also reduced. This reflects in to the reduced length or size of the feature vector.

The length of chromosome of GA is to be decided based on the length of feature vector. This chromosome has real value between 0 and 1, randomly generated at each position, in its first form. The position values will further be modified by comparing with some randomly generated value between 0 and 1. If the position value is less than this instantaneous random value then the position in chromosomes will be made to zero. This modification transforms the wide range variations in usable percentage variation of total feature elements. It also enables evaluation of the chromosome's performance of recognition with even small percentage of elements. This chromosome may be then multiplied (element wise) with feature vector to be optimized, before using it for recognition.

GAs are suitable for the information retrieval for the following reasons.

- The document search space represents a high dimensional space.
- In comparison with the classical information retrieval models, GA manipulates a population of queries rather than a single query.
- Each query may retrieve a subset of relevant documents that can be merged.

GA contributes to maintain useful information links representing a set of terms indexing the relevant documents. Fig. 30 describes the genetic based document retrieval architecture.

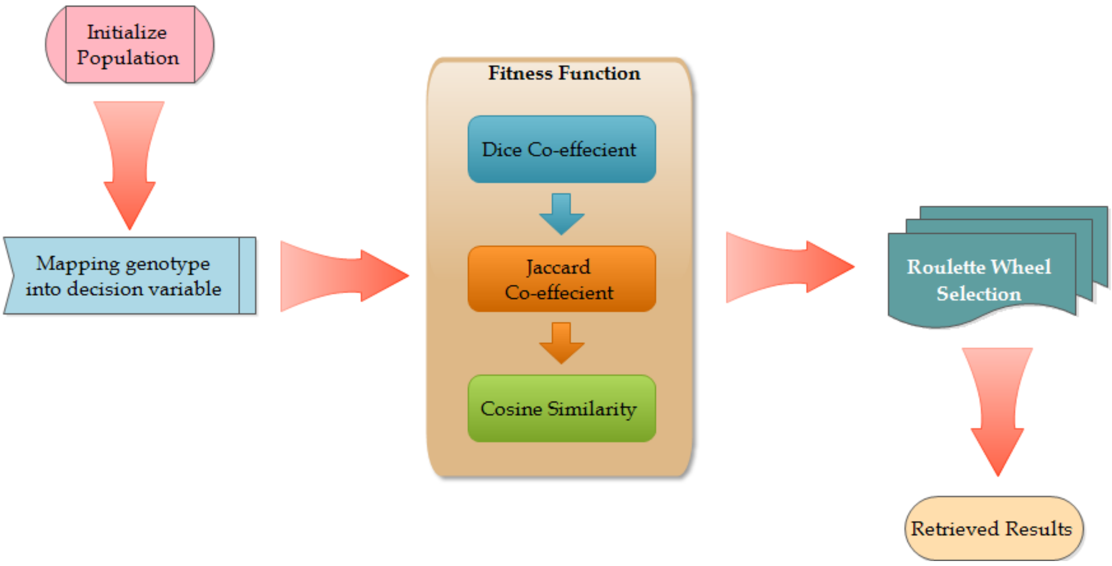


Fig. 30: Document Retrieval - Genetic Approach

Chromosome representation gives the collection of information regarding individuals present in the chosen population. Chromosome is the data structure which denotes the entire population into a single matrix (Anubha Jain et. al., 2014).

$$Chrom = \begin{pmatrix} N_{ind1}\,L_{ind1} & N_{ind1}L_{ind2} & \cdots & \cdots & N_{ind1}L_{ind2804} & N_{ind1}L_{ind2805} \\ N_{ind2}\,L_{ind1} & N_{ind1}L_{ind2} & \cdots & \cdots & N_{ind2}L_{ind2804} & N_{ind2}L_{ind2805} \\ N_{ind3}\,L_{ind1} & N_{ind3}L_{ind2} & \cdots & \cdots & N_{ind3}L_{ind2804} & N_{ind3}L_{ind2805} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ N_{ind50}\,L_{ind1} & N_{ind50}L_{ind2} & \cdots & \cdots & N_{ind50}L_{ind2804} & N_{ind50}L_{ind2805} \end{pmatrix}$$

Where

N_{ind} - Number of individuals in the initial population

L_{ind} - Length of chromosome

N_{var} is the number of strings in each individual and it varies for every individual depends on the chosen population. In initial stage N_{var} for each individual population is mapped to L_{ind} for representing the genotype and further genotype is mapped into decision variable space to obtain phenotype representation. The length of chromosome depends on the number of individuals and the count of keywords present in it. Mapping of genotype into the decision variable space is known to be a Phenotype. Fig. 31 shows the phenotype representation for 50 individual population with 2805 as length of individual chromosome. In each generation, fitness of each individual from the population is evaluated based on the similarity measures. Query term is mapped against genotype information of each individual using fitness functions. In this work, fitness function is computed using three similarity measures: 1) Dice coefficient 2) Jaccard coefficient and 3) Cosine Similarity respectively. Similarity measure for computing Dice coefficient, Jaccard coefficient and Cosine similarity measures are shown in eqn. 17, 18 and 19 respectively.

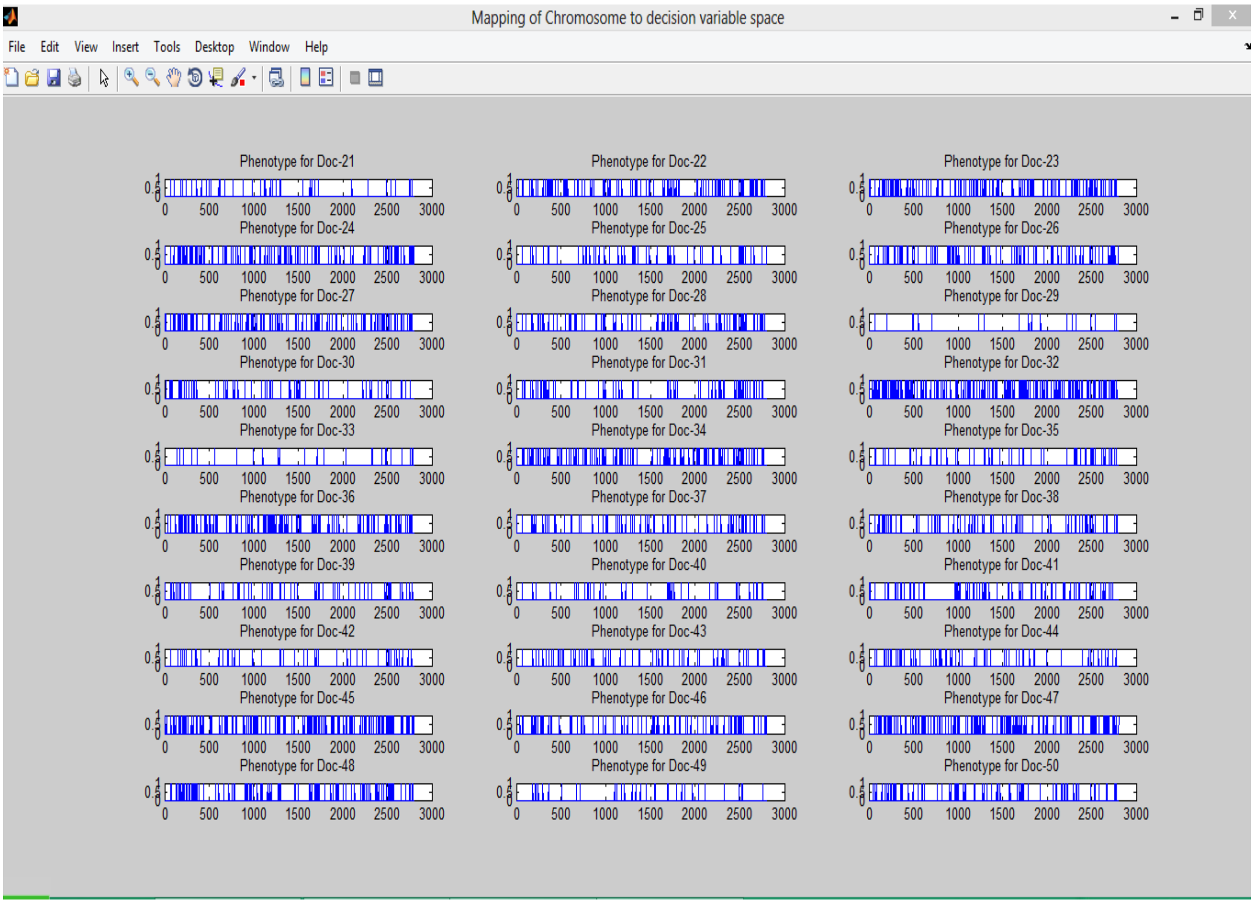


Fig. 31: An example for Mapping of Genotype into Decision variable space (Phenotype) for N_{ind} = 50 and L_{ind} = 2805

4.3.1 Dice Co-efficient

Dice's coefficient measures how similar query genotype and individual genotype are in terms of the number of common bigrams (a bigram is a pair of adjacent letters in the string) (Vikas Thada et. al 2013). Eqn. 17 gives the dice co-efficient between the query genotype and the individual genotype.

$$\mathrm{D_i} = 2\,\frac{|\,\mathrm{q} \cap \mathrm{ind_i}|}{|\mathrm{q}| + |\mathrm{ind_i}|} \qquad (17)$$

4.3.2 Jaccard Co-efficient

Jaccard similarity coefficient is used for comparing the similarity and diversity of sample sets (Vikas Thada, 2013). Jaccard Index and distance is shown in eqn. 18 and in 19.

$$J_i = \frac{|\,\mathrm{q} \cap \mathrm{ind_i}|}{|\mathrm{q}| + |\mathrm{ind_i}| - |\,\mathrm{q} \cap \mathrm{ind_i}|} \qquad (18)$$

$$J_d = 1 - J_i \qquad (19)$$

4.3.3 Cosine Similarity Measure

Degree of Similarities between document genotype and query genotype are computed using cosine angle (Jinn-Tsong Tsai, 2014). In other words it is a measure between genotype of an inner product space that measures the cosine angle between them.

$$C_i = \frac{|\,q \cap \text{ind}_i\,|}{|q|^{1/2} \,.\, |\text{ind}_i|^{1/2}} \qquad (20)$$

Where,

D_i - Dice co-efficient for ith individual

J_i - Jaccard co-efficient/Index for ith individual

J_d - Jaccard distance for ith individual

C_i - Cosine similarity for ith individual

Q - Query Term / Keyword

q - Genotype of Query 'Q'

ind_i - Genotype of individual population 'i'

Experiments have been conducted for various fitness functions and results are evaluated. During each successive generation, a proportion of the existing population is selected to breed a new generation. Individual solutions are selected through a fitness-based process, where fitter solutions are typically more likely to be selected. In this work, Roulette wheel selection procedure is used to identify best fitter solutions.

4.3.4 Roulette Wheel

Roulette Wheel (RW) selection algorithm or Fitness proportionate selection is a randomized algorithm. Fitness functions for 6 relevant documents are shown in Fig. 32 from which best fit is identified by the roulette wheel selection algorithm.

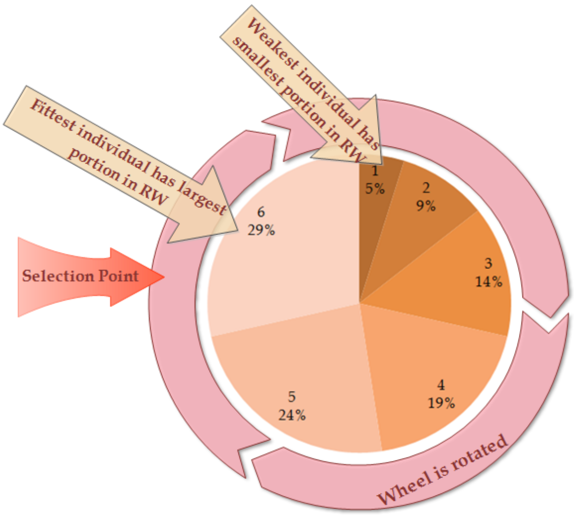

Weakest individual has smallest portion in RW
Fittest individual has largest portion in RW
Selection Point
1
5%
2
9%
3
14%
4
19%
5
24%
6
29%
Wheel is rotated

Fig. 32: Roulette Wheel Approach: based on fitness for 6 relevant documents

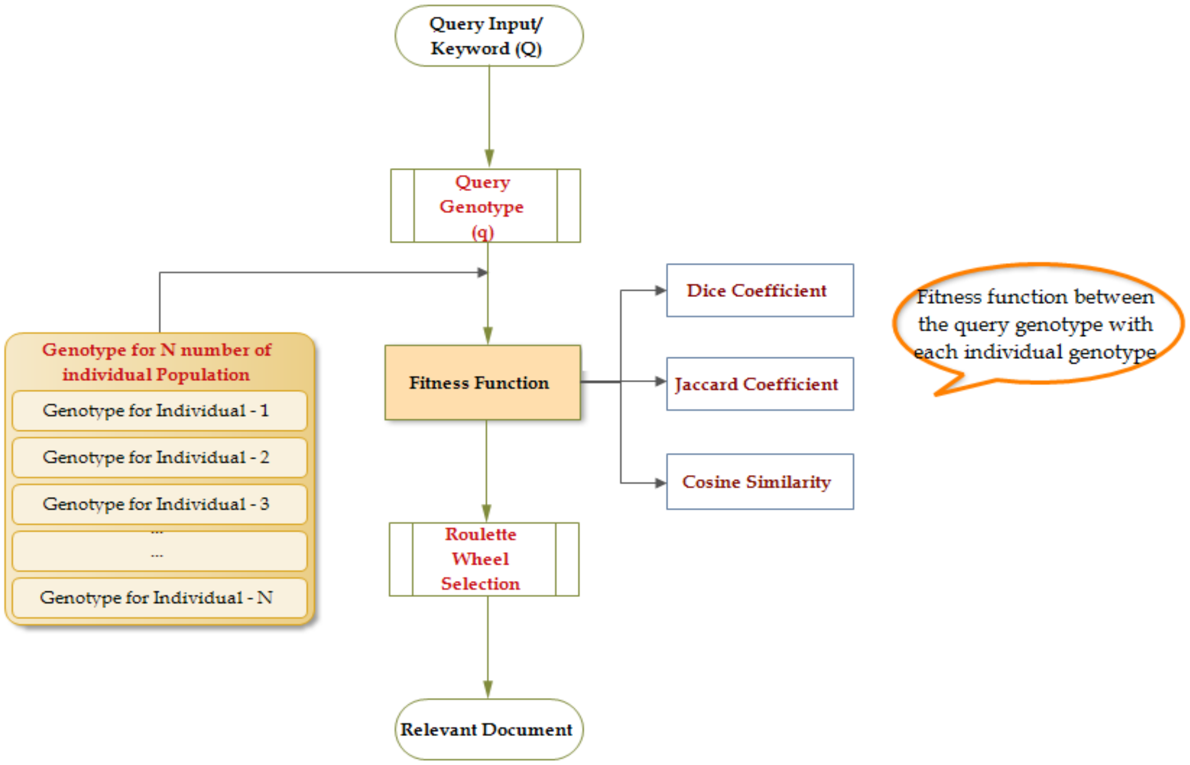

Query Input/ Keyword (Q)
Query Genotype (q)
Genotype for N number of individual Population
Genotype for Individual - 1
Genotype for Individual - 2
Genotype for Individual - 3
...
...
Genotype for Individual - N
Fitness Function
Dice Coefficient
Jaccard Coefficient
Cosine Similarity
Fitness function between the query genotype with each individual genotype
Roulette Wheel Selection
Relevant Document

Fig. 33: Framework of the Proposed Retrieval Algorithm using GA

The idea behind the RW Selection is that the fittest individuals have a greater chance of endurance than weaker ones from the chosen population. The probability of each individual's fitness function is selected randomly for evaluating the selection procedure and their results are normalized (ranges from 0.0 to 1.0) for identifying the best fit.Finally, retrieval system returns an ordering of document over the collection of document for the required query. Fig. 33 depicts the overall steps involved in GA.

4.4 Dataset

4.4.1 Dataset for VSM

The proposed system is evaluated using two different databases A and B (DS - A and DS - B) which consist of 243 documents. DS - A consists of 23 distinct documents (D_1, D_2...D_{23}). However, it is essential to compare large documents and hence the second database has nearly 220 similar documents (D_1, D_2 ...D_{220}) for studying the performance of the system and to check the misclassification between the query and the documents. Database description for DS - A is shown in Table 5 and DS - B in Table 6.

Table 5: Database description for DS – A

Collection Name	No. of Documents	No. of Queries
Acoustics	2	6
Analog and Digital Signals	4	10
Feature Extraction	3	8
Perception	2	4
Phonetics	3	7
Speaker Identification	2	5
Speech Recognition	4	8
Speech Synthesis	3	7

Table 6: Database description for DS – B

Collection Name	No. of Documents	No. of Queries
Linguistics	19	42
Signal Processing	52	76
Speech Processing	58	75
Image Processing	42	69
Video Processing	49	70

To improve the retrieval efficiency it is necessary to collect huge amount of data, most comparable collection of documents may help the user to find the required search results as well as to improve the systems precision. The performance of the proposed system is evaluated by two different types of datasets which is described in Section 4.6.

4.4.2 Dataset for GA

GA is evaluated using 962 queries over 734 documents which are preferred by the user based on their requirement.

Table 7: Database description for 20 Newsgroup dataset

Dataset Name	No. of Individuals	Length of Individuals
Atheism	1000	38813
Computer Graphics	1000	41076
Microsoft Windows	1000	39257
PC Hardware	1000	33174
MAC Hardware	1000	30637
Windows	1000	47626
Forsale	1000	33436
Autos	1002	34131
Motorcycles	1000	31324
Baseball	1000	34959
Hockey	1000	39611
Cryptogrpahy	1000	43622
Electronics	1000	34222
Medical	1000	44703
Space	1000	45254
Christian Religion	997	45492
Politics Guns	1000	43068
Politics Mideast	1000	53883
Politics Misc	1000	50365
Religion Misc	1000	44442
Total No. of Population	**19,999**	**8,09,095**

In addition to this, 20 newsgroup datasets were used for measuring the performance and accuracy of the system. News group dataset, discuss about particular subject/topic consisting of notes written to a central internet site. The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. The 20 newsgroups collection has become a popular data set for experiments in text applications of machine learning techniques and document retrieval.

In this work, 20 news group datasets were considered for evaluating the performance of the system which is described as follows. Dataset description of the proposed system is shown in Table 7 and distinct domain separation is shown in Table 9. In Table 9, the length of the chromosome depends on the number of chromosome extracted from the entire document collection.

4.5 Performance Analysis

Performance of the system is analyzed using precision and recall. Precision is the fraction of the documents retrieved that are relevant to the user's information need. Precision and recall are single-value metrics based on the whole list of document returned by the system. Precision is analogous to positive predictive value, it is computed by considering all retrieved documents into account and recall takes all the relevant documents into an account.

$$P = \frac{|\{Relevant\ Documents\} \cap \{Retrieved\ Documents\}|}{|\{Retrieved\ Documents\}|} \qquad (21)$$

$$R = \frac{|\{Relevant\ Documents\} \cap \{Retrieved\ Documents\}|}{|\{Relevant\ Documents\}|} \qquad (22)$$

F-Measure is the harmonic mean of precision and recall which is used to measure the accuracy of the system and G-Measure is the geometric mean of precision and recall.

$$F1\ Score = 2\frac{(Precision) * (Recall)}{(Precision) + (Recall)} \qquad (23)$$

$$G\ Measure = \sqrt{Precision} * Recall \qquad (24)$$

4.6 Experimental Results

4.6.1 Experimental Evaluation of Vector Space Mode

Experiments are conducted for both databases. Table 8 shows the local weighting (*tf*) for the dataset DS - A. Here 23 documents are compared for each query q_i.

Table 8: Sample Term frequency calculation for single term query

Query	*tf values for the documents (D_1 to D_{23})*																					
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
Acoustics	0	**5**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Auditory	0	1	0	1	**13**	0	0	0	0	0	0	0	0	4	0	2	0	0	0	0	0	0	0
Digital	0	0	9	11	0	**13**	0	0	0	0	2	6	0	0	0	3	0	0	0	0	0	0	0
Phoneme	0	0	0	0	0	0	0	0	**10**	3	0	0	0	0	0	0	0	0	0	0	0	0	0
Recognition	0	0	0	0	0	0	1	1	0	0	0	0	0	0	**7**	0	2	0	0	1	0	0	0
Signal	1	0	**26**	9	0	11	0	0	0	0	0	15	0	0	0	3	0	0	0	0	0	4	0
Sound	**6**	2	2	0	4	0	0	0	0	0	2	0	0	4	0	0	0	1	2	0	0	0	1
Speech	0	0	0	0	1	0	0	0	4	5	5	0	7	0	3	**24**	3	17	4	0	2	4	13

In Table 8, Eight single term sample search queries are given as input to the retrieval system and number of occurrences of the q_i is counted over 23 distinct documents from DS - A. For instance, sample isolated search terms are shown in Table 8 and its inverse document frequency for query term is computed for DS - A. Weighting factor *(w)* for query vector $|Q|$ is measured using the dot product of query term with *idf* and document vector $|D|$ is computed using the dot product of document term with idf. A similarity measure for most relevant (or winning) document for DS-A is illustrated in Fig. 34.

In Table 8 the chosen query is isolated word hence it is not necessary to parse or pre-process the data. By analyzing the similarity measure it is clear that non-zero vectors are the documents containing the search term from which higher measure is considered as relevant document. Highest cosine similarity measure among *n* number of documents for the query q_i is shown in Fig. 34.

Fig. 35 shows the highest cosine similarity measure for 10 different sentences from DS-B which consists of 220 documents. By computing the cosine similarity between the query and the document the highest value among the similarity is considered as more relevant document.

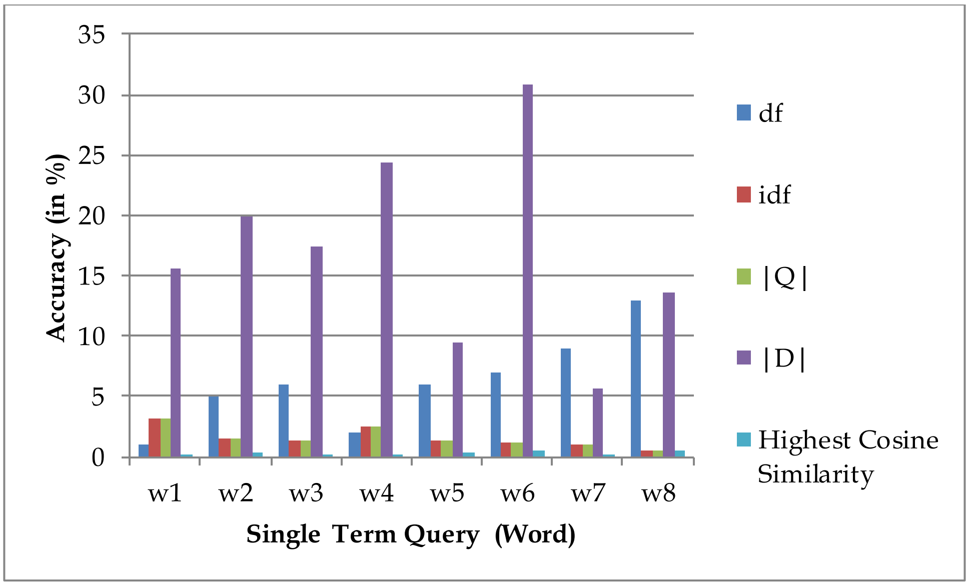


Fig. 34: Sample Query and Document Vector calculation for single term query

Documents are re-arranged based on cosine similarity in descending order and its higher cosine similarity is assigned as Rank 1, next higher cosine similarity as Rank 2 respectively. Cosine similarity for 10 sample sentences which is retrieved from DS-B is shown in Fig. 35.

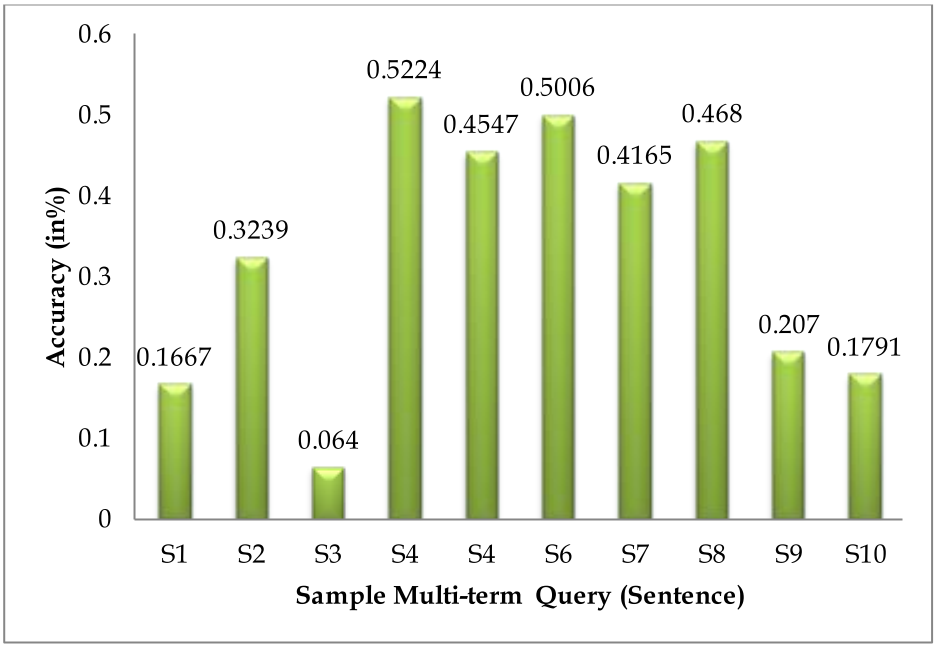


Fig. 35: Highest cosine similarity measure for 10 samples in DS - B

The performances of DS - A and performance of DS - B are analyzed in Fig. 36 and Fig. 37. Fig. 36 depicts the performance analysis for Dataset - A for 8 different groups (G_1,G_2,...G_8 respectively) of domain with different queries. For DS - A, the system provides an average precision as 94.38%, recall as 96.88%, F1 score and G measure as 95.03% and 95.33% respectively.

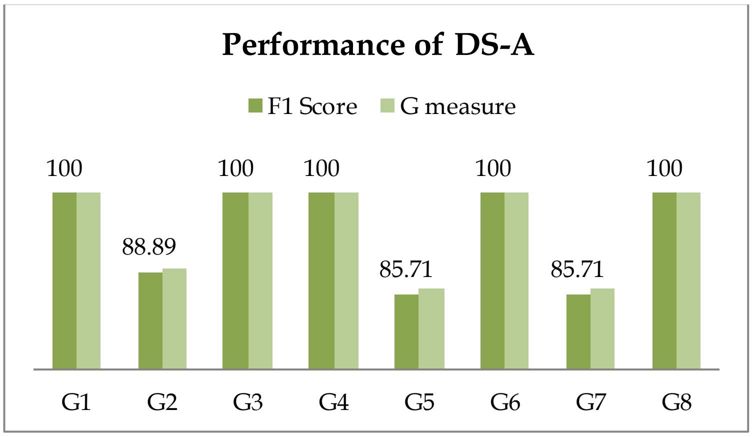


Fig. 36: Performance of 8 different clusters in Dataset - A

Fig. 37 depicts the F1score and G - measures for the Dataset-B for 5 different groups (G_1,G_2,...G_5 respectively) of domain with different queries. From the exhaustive analysis it is proven that the retrieval system shows the performance of 95.68% as average precision and Recall as 98.08% from which average F1 score and G measure are computed as 96.7% and 96.79% respectively.

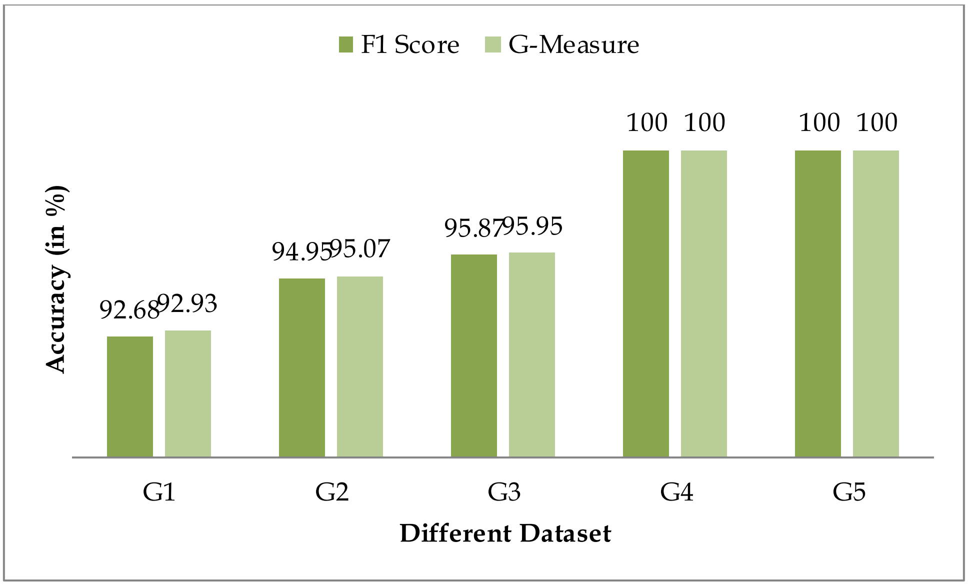


Fig. 37: Performance of 5 different clusters in Dataset - B

Fig. 38 shows the overall performance of the retrieval system. From the analysis it is proven that the system shows an accuracy of 95.86% and 96.06%.

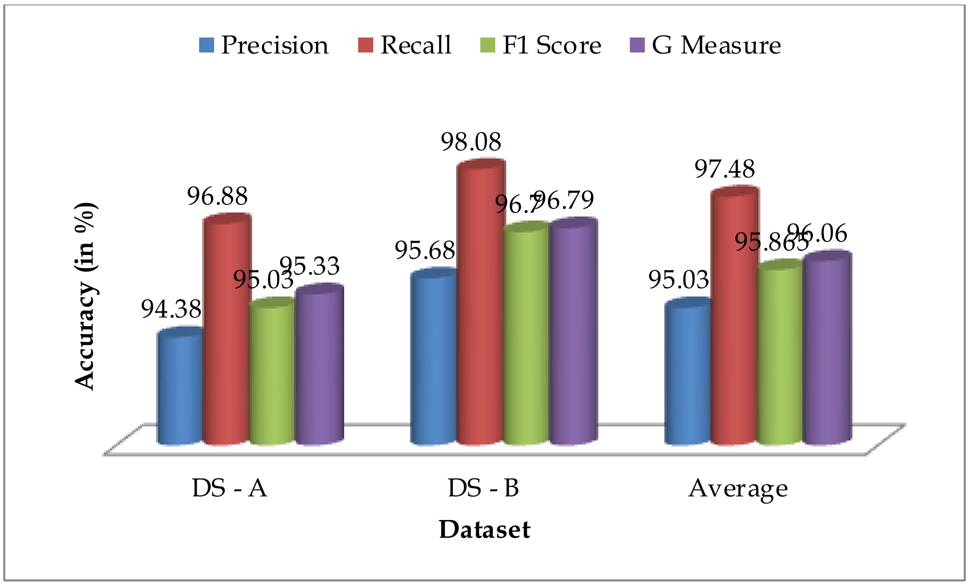


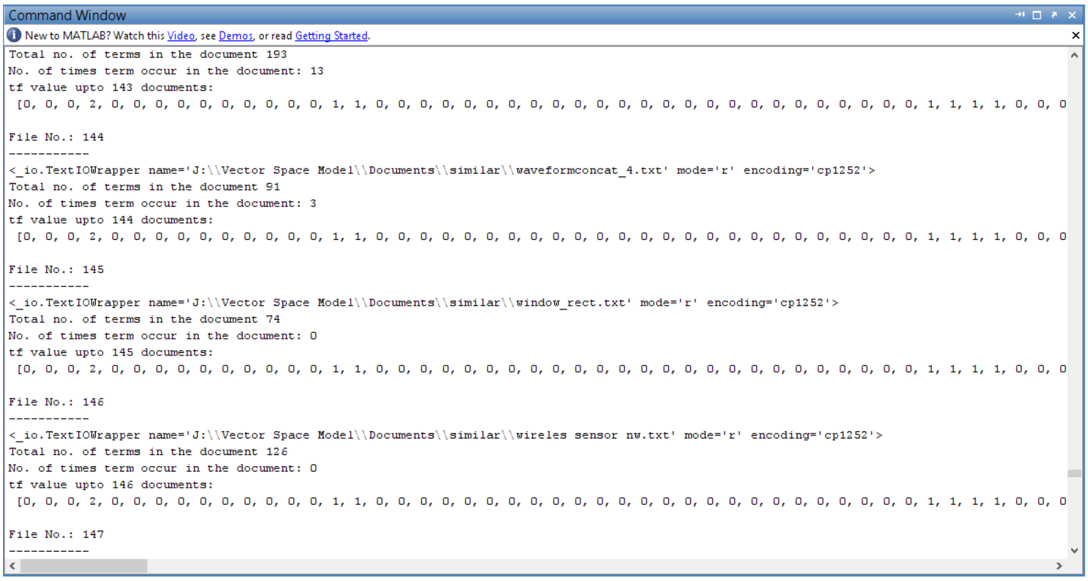
Fig. 38: Overall Performance of the VSM Model

4.6.2 Experimental Evaluation of Genetic Algorithm

Fitness function between the query chromosome and the document chromosomes are evaluated using various fitness functions like dice coefficient, Jaccard coefficient and cosine similarity. Table 10, shows the performance of the fitness function using various similarity measures namely dice co-efficient, Jaccard co-efficient and cosine similarity, for the given query and the genotype. Fig. 39, 40, 41 and 42 shows the Snapshots of VSM model.

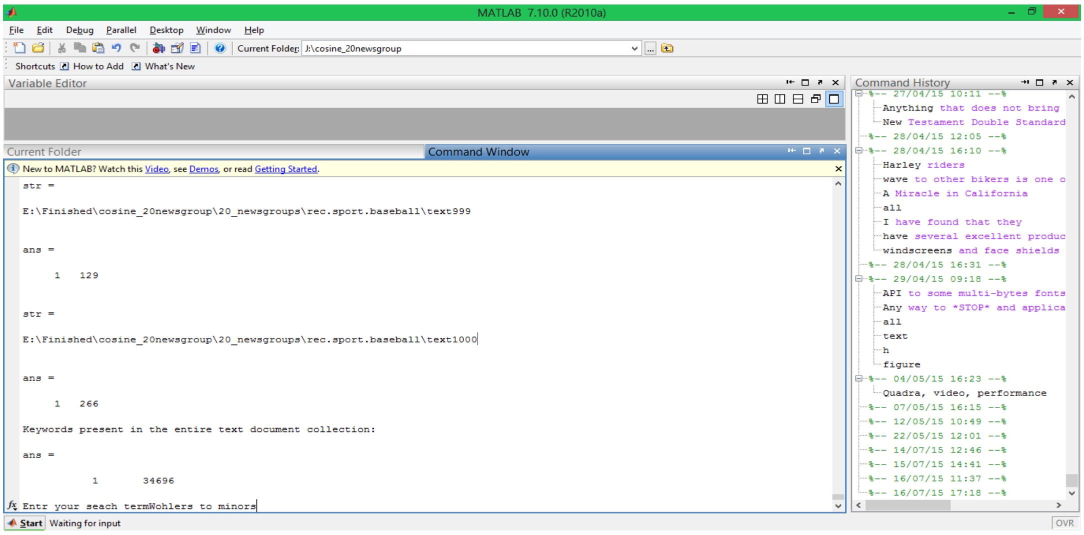
Table 9: Domain based 20 Newsgroup Dataset

Domain of 20NG Dataset	No. of Dataset	N_{ind}	L_{ind}	No. of Queries
Computer	7	7000	268828	5495
Medical	1	1000	44703	893
Electronics	1	1000	34222	936
Space/ Earth	1	1000	45254	758
Games	2	2002	74570	1683
Vehicles	2	2000	65455	1632
Religion	3	2997	128747	2000
Politics	3	3000	147316	2530
Total	**20**	**19999**	**809095**	**15927**


Command Window
New to MATLAB? Watch this Video, see Demos, or read Getting Started.

```
Total no. of terms in the document 193
No. of times term occur in the document: 13
tf value upto 143 documents:
 [0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0

File No.: 144
-----------
<_io.TextIOWrapper name='J:\\Vector Space Model\\Documents\\similar\\waveformconcat_4.txt' mode='r' encoding='cp1252'>
Total no. of terms in the document 91
No. of times term occur in the document: 3
tf value upto 144 documents:
 [0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0

File No.: 145
-----------
<_io.TextIOWrapper name='J:\\Vector Space Model\\Documents\\similar\\window_rect.txt' mode='r' encoding='cp1252'>
Total no. of terms in the document 74
No. of times term occur in the document: 0
tf value upto 145 documents:
 [0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0

File No.: 146
-----------
<_io.TextIOWrapper name='J:\\Vector Space Model\\Documents\\similar\\wireles sensor nw.txt' mode='r' encoding='cp1252'>
Total no. of terms in the document 126
No. of times term occur in the document: 0
tf value upto 146 documents:
 [0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0

File No.: 147
-----------
```

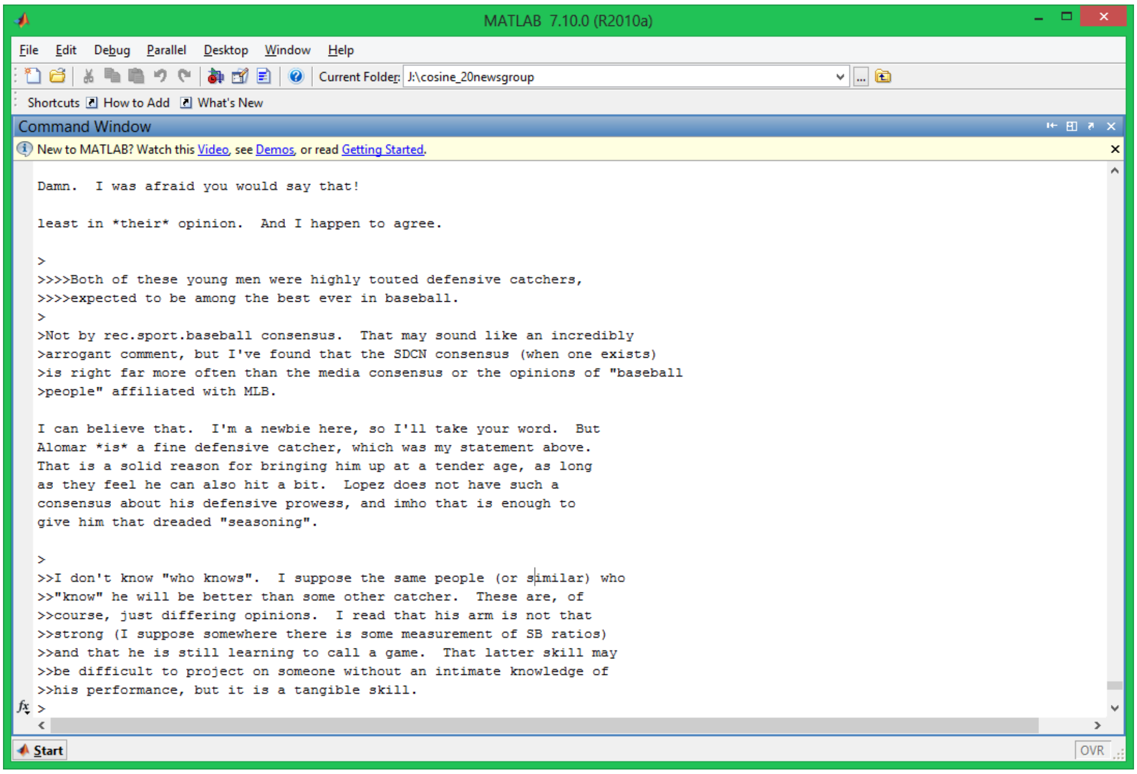
Fig. 39: Snapshot of Vector Space Model in Matlab



```
str =

E:\Finished\cosine_20newsgroup\20_newsgroups\rec.sport.baseball\text999


ans =

     1   129


str =

E:\Finished\cosine_20newsgroup\20_newsgroups\rec.sport.baseball\text1000


ans =

     1   266

Keywords present in the entire text document collection:

ans =

            1        34696

fx Entr your seach termWohlers to minors
```

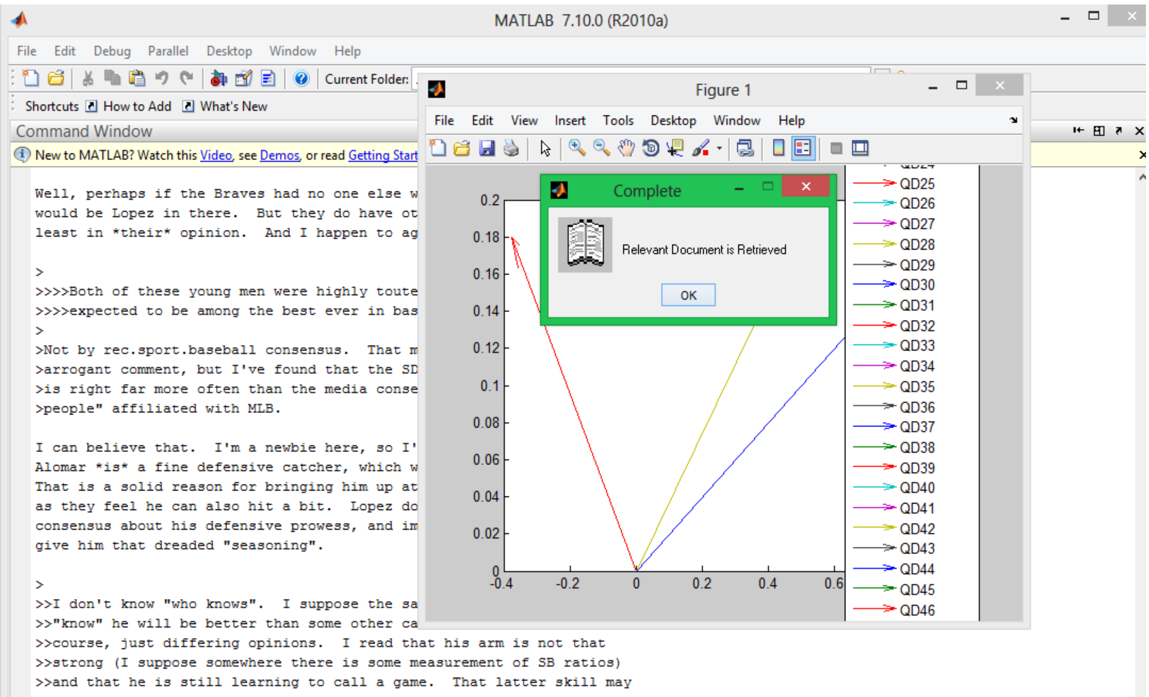
```
%-- 27/04/15 10:11 --%
   Anything that does not bring
   New Testament Double Standard
%-- 28/04/15 12:05 --%
%-- 28/04/15 16:10 --%
   Harley riders
   wave to other bikers is one o
   A Miracle in California
   all
   I have found that they
   have several excellent produc
   windscreens and face shields
%-- 28/04/15 16:31 --%
%-- 29/04/15 09:18 --%
   API to some multi-bytes fonts
   Any way to *STOP* and applica
   all
   text
   h
   figure
%-- 04/05/15 16:23 --%
   Quadra, video, performance
%-- 07/05/15 16:15 --%
%-- 12/05/15 10:49 --%
%-- 22/05/15 12:01 --%
%-- 14/07/15 12:46 --%
%-- 15/07/15 14:41 --%
%-- 16/07/15 11:37 --%
%-- 16/07/15 17:18 --%
```

Fig. 40: Snapshot of Genetic Algorithm (Training process of GA for NG-10)


MATLAB 7.10.0 (R2010a)
File Edit Debug Parallel Desktop Window Help
Current Folder: J:\cosine_20newsgroup
Shortcuts How to Add What's New
Command Window
New to MATLAB? Watch this Video, see Demos, or read Getting Started.
Start
OVR

```
Damn.  I was afraid you would say that!

least in *their* opinion.  And I happen to agree.

>
>>>>Both of these young men were highly touted defensive catchers,
>>>>expected to be among the best ever in baseball.
>
>Not by rec.sport.baseball consensus.  That may sound like an incredibly
>arrogant comment, but I've found that the SDCN consensus (when one exists)
>is right far more often than the media consensus or the opinions of "baseball
>people" affiliated with MLB.

I can believe that.  I'm a newbie here, so I'll take your word.  But
Alomar *is* a fine defensive catcher, which was my statement above.
That is a solid reason for bringing him up at a tender age, as long
as they feel he can also hit a bit.  Lopez does not have such a
consensus about his defensive prowess, and imho that is enough to
give him that dreaded "seasoning".

>
>>I don't know "who knows".  I suppose the same people (or similar) who
>>"know" he will be better than some other catcher.  These are, of
>>course, just differing opinions.  I read that his arm is not that
>>strong (I suppose somewhere there is some measurement of SB ratios)
>>and that he is still learning to call a game.  That latter skill may
>>be difficult to project on someone without an intimate knowledge of
>>his performance, but it is a tangible skill.
fx >
```

Fig. 41: Snapshot of Document Retrieval using Genetic Algorithm (Testing process of GA for NG-10)



Relevant Document is Retrieved

```
Well, perhaps if the Braves had no one else w
would be Lopez in there.  But they do have ot
least in *their* opinion.  And I happen to ag

>
>>>>Both of these young men were highly toute
>>>>expected to be among the best ever in bas
>
>Not by rec.sport.baseball consensus.  That m
>arrogant comment, but I've found that the SD
>is right far more often than the media conse
>people" affiliated with MLB.

I can believe that.  I'm a newbie here, so I'
Alomar *is* a fine defensive catcher, which w
That is a solid reason for bringing him up at
as they feel he can also hit a bit.  Lopez do
consensus about his defensive prowess, and im
give him that dreaded "seasoning".

>
>>I don't know "who knows".  I suppose the sa
>>"know" he will be better than some other ca
>>course, just differing opinions.  I read that his arm is not that
>>strong (I suppose somewhere there is some measurement of SB ratios)
>>and that he is still learning to call a game.  That latter skill may
```

Fig. 42: Retrieved relevant documents with its cosine angle

Table 10: Performance analysis of various Fitness functions using Genetic approach

Similarity Measure	**Dice Co-efficient**	**Jaccard Co-efficient**	**Cosine Similarity**
Performance (in %)	81.43	96.79	**98.84**

In Fig. 43, performance of different similarity measures using Genetic Approach is depicted.

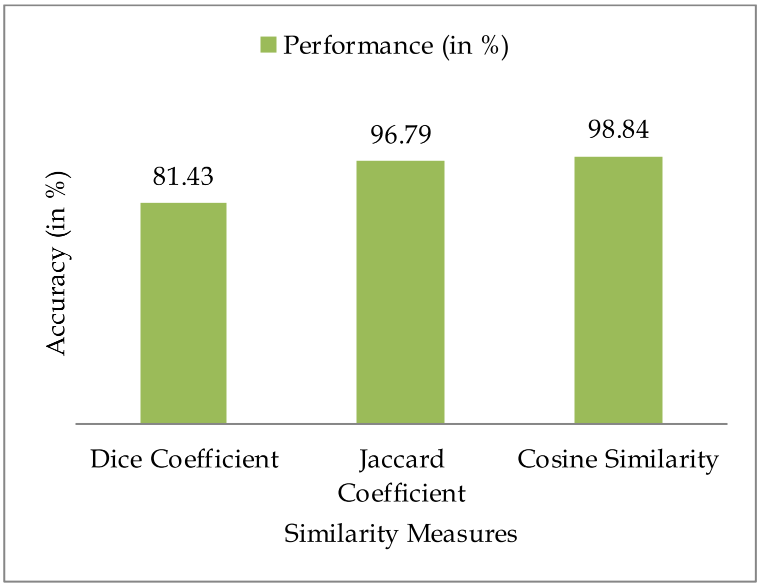


Fig. 43: Performance of different similarity measures using Genetic Approach

Table 10 shows the overall performance analysis of the document retrieval system. VSM is compared with GA using three fitness function and their results are compared in Table 11.

Table 11: Evaluating the Performance of Retrieval system

Technique	**VSM (Cosine)**	**Ontology assisted VSM (Cosine)**	**GA**		
			Dice	**Jaccard**	**Cosine**
Performance (in %)	93.893	95.7	81.43	96.79	**98.84**

From analysis, it is clear that GA with fitness function based on cosine angle provides better results than others.

4.7 Summary

Ontology assisted VSM and genetic based retrieval system returns an ordering of document over the collection of document for the required speech query. Genetic approach with cosine similarity provides better performance than other relevance retrieval approach. VSM based document retrieval system is analyzed using 387 queries over 243 documents and the quality of the speech was measured using MOS which is collected from 43 persons with four different scales. The system shows an accuracy of 95.87% as F1 score and 96.06% as G-measure. GA based retrieval system is evaluated using 20NG dataset with 19,999 documents and 734 user preferred document collection and shows an accuracy of 98.84% cosine based similarity measure.

5. Text to Speech Conversion

In this work, both phoneme and syllable utterances are considered. In SS, text document is given as input to the system which possibly consists of non-standard formats like digits/integers, numerical expressions, cardinal suffixes etc. Syllabic texts are extracted from the normalized form of the given input. Text normalization, syllabic transcription and phonemic transcriptions are discussed in Section 5.2. Fig. 44 shows the architecture of speech synthesis system.

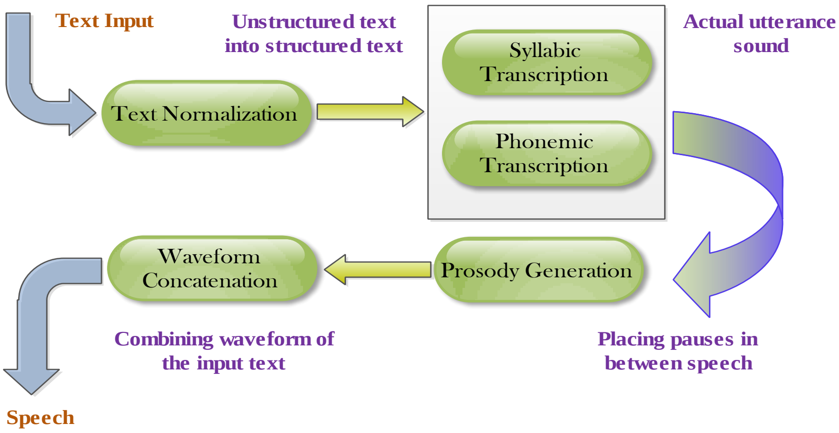


Fig. 44: Framework for the speech synthesis system

Concatenation of speech synthesis can be created by concatenating pieces of recorded speech that are pre-recorded into the database, whereas for converting the text into speech sound the system must generalize all the speech units uniquely (Mahwash Ahmed and ShibliNisar, 2014). Any text may contain special elocution that should be stored in lexicon form. For this conversion the system uses text normalization which will be possessed internally within the synthesizer (Shreekanth.T et. al. 2014). Concatenative speech synthesis converts the written text into phonemic or syllabic representation and then it converts the generated phonemic/syllabic representation into waveform (Carlson, R., &Granström, B., 2007). These waveforms are combined to fabricate as sound. Naturalness in synthetic speech generated by concatenative speech synthesis increases when the number of concatenation points required for creating a waveform is minimal.

5.1 Text Normalization

Text to Speech Synthesizer mechanism works internally by synthesizing words. However, input text documents contains words comprised of written elements such as statistics, date, time, abbreviations, numbers, symbols etc., If the text consist of abbreviations and numbers then the system must determine how these non-standards should be read out (Jong Kuk Kim, 2009). Any text that has a special pronunciation should be stored in lexicon such as abbreviation, acronyms, special symbols etc. All diverse elements must be first converted into general or actual utterances and then only the system can synthesized as speech. Such conversion of diverse units into actual utterances which takes place internally within the synthesizer is expressed as text normalization.

In general, text normalization is the conversion of text that includes non-standard word such as statistics, abbreviations, misspelling into normal words. Each literal symbols are recognized as units. These symbols should be separated from adjacent text with white space. Speech unit can be either fixed size diphones or variable length units such as syllables and phonemes.

Table 12: General Text Normalization

Type	Text Normalization Representation
Digits/ Integer	0-9
Alphabetic character alpha_char	a-z, A-Z
Phrase or Sentence word_char	<alpha_char> | - | _ | 0 -9
Cardinal suffix	st, nd, rd, th
Numerical Expression	Integer expression, Cardinal suffix, Floating expression

Open close brackets, equals, greater-than, dollar, percent, Pound sign, comma etc. are termed as Literal Symbols. Phrases in a context consist of literal symbols and numerical representation. The explanation for all the literals and numerical representation has to be declared initially for the conversion of non-structural to structural representation.

5. 2. Phonemic/Syllabic Transcription

Phonemic transcription attempts to depict the individual dissimilarity that arises between speakers of a language. Phonemic transcription aspires to record the phonemes that a speaker utilize rather than the real spoken variants of those phonemes that are created when a speaker converse an utterance (Mahwash Ahmed and ShibliNisar, 2014). A phoneme is an intangible linguistic unit that survives entirely in the brain of a speech producer; they could be symbolized by any arbitrary classification of symbols. The most widely accepted classification of symbols is the International Phonetic Alphabet (IPA). According to IPA English consist of 44 phonemes and are classified into three major sound categories namely 24 consonants, 8 diphthongs and 12 vowel sounds. Each Phoneme is a group of sound that is actually uttered (Jong Kuk Kim, 2009). These phonemic transcriptions are used by the speech synthesis system for the conversion of Text to Speech.

Table 13: Phonemic Transcription

Standard Representation	Phonetic Transcription
speech	spiːtʃ
transcription	trænˈskrɪpʃən
text	tekst
Cat	kæt
speech	spiːtʃ
Conversion	kənˈvərʒən

Syllabifications the process of segmenting stream of speech units into syllables. Speech is based on basic sound units which are inherently syllable units made from C,CV, CCV, VC and CVC combinations, where C is a consonant and V is a vowel. From perceptual results, it is observed that from four different choices of speech units: syllable, diphone, phone and half phone, the syllable unit performs better than all the rest and is a better representation for Indian languages. There are few rules for dividng words into syllables. There are four ways to split up a word into its syllables:

Rule 1: Divide the word between two middle consonants.

Rule 2: Usually divide before a single middle consonant.

Rule 3: Divide before the consonant before an "-le" syllable.

Rule 4: Divide off any compound words, prefixes, suffixes and roots which have vowel sounds.

According to rule 1, split up words have two middle consonants. For example consider the following words hap/pen, let/ter, din/ner. Happen consist of two consonants pp at the middle of the word; similarly for letter and dinner it has two middle consonants.

The only exceptions are the consonant digraphs. Never split up consonant digraphs as they really represent only one sound. The exceptions are "th", "sh", "ph", "th", "ch", and "wh".

According to Rule 2, when there is only one syllable, it usually divides in front of it. "o/pen", "i/tem", "e/vil", and "re/port" are few examples for Rule2. The only exceptions in these are those times when the first syllable has an obvious short sound, as in "cab/in".

Based on Rule 3, a word that has the old-style spelling in which the "-le" sounds like "-el", divide before the consonant before the "-le". For example: "a/ble", "fum/ble", "rub/ble" "mum/ble" and "thi/stle". The only exceptions in these are "ckle" words like "tick/le".

According to Rule 4, the word is split off into parts of compound words like "sports/car" and "house/boat". Divide off prefixes like "un/happy", "pre/paid", or "re/write". Also divide off suffixes as in the words "farm/er", "teach/er", "hope/less" and "care/ful". In the word "stop/ping", the suffix is actually "-ping" because this word follows the rule that when it is added with "-ing" to a word with one syllable, it doubles the last consonant and add the "-ing".

5.3 Prosody

Prosody is the pitch, volume and speed that words, sentences and phrases are spoken with. The system may also need to split the input into smaller chunks of output text to determine which words needs to be emphasized (Campbell N., 2006). The term prosody refers to both pitch and the placement of pauses in between speech for making synthetic natural speech sound. It is very easy for a human speaker to pause at any places in speech, but it’s complicated for the machine to fabricate sounds.

Prosody plays an important role in guiding listener for speaker attitude towards the message. Prosody consists of systematic perception and recovery of speaker intentions based on Pauses, Pitch, Rate and Loudness. Fig. 46 shows the architecture of prosody generation. Pauses are used to indicate phrases and separate the two words (Mahwash Ahmed and ShibliNisar, 2014). Pitch refers the rate of vocal fold cycle as function of time (Chalamandaris A, 2009). Rate denotes Phoneme/syllable duration, time and Loudness represents the relative amplitude or volume.

Initially, the engine identifies the beginning and ending of sentences. The pitch will be likely to fall near the end of a statement and rise for a question. Similarly, the machines starts speaking the small piece of syllable/phoneme and it fall on to the last word, and then pauses are placed in between the sentences or phoneme/syllable for clear reading. So a prosodic system for a text to speech must provide suitable pauses for the speech and also adequate information to make the pitch sound realistic. All the Text to Speech Engines have to convert the list of syllabic/phonemes and their volume, pitch and duration into digital audio.

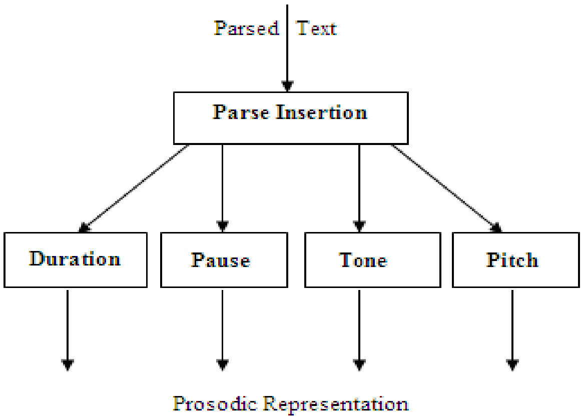


Fig. 45: Architecture of Prosody

TTS Engine generates the digital audio by concatenating pre-recorded sound clips, which are stored in a database. The combined pre-recorded smallest unit of sound is given to the TTS Engines which speaks the sentence loud.

5.3.1 Duration

Rule based method is commonly used method for computing duration for the phrase or sentence. Time duration between the sentences or phones decides the clarity of the speech so that durational assignment plays a vital role in text to speech conversion. Each Phone is pronounced in various duration by the user. The duration of phone d is expressed as

$$d = d_{min} + r(d' - d_{min}) \tag{25}$$

Where,

d = Average duration of the phone

d_{min} = Minimum duration of the phone

r = correction

Whereas the correction r is calculated by,

$$r = \prod_{i=1}^{N} r_i \tag{26}$$

5.3.2 Pause and Pitch

Pauses are mainly used in running text which is generated in the form of utterance output. In usual systems, the reliable location which is designated to insert pause is the pronunciation symbols. For every full stop or commas in a sentence the pause has to be placed for absolute reading of phrase in between those sentences.

That is text may allow pausing for some duration when it finds a comma or full stop in the sentence. So silence sound "Sil" is placed for few milliseconds then Connection "C" has to be made for the continuation for the phrases. Generally, speech synthesis engines need to express their usual pitch patterns within the broad limits specified by a Pitch markup (Chalamandaris A, 2009).

5.3.3 Tone

The prosodic parameters in tones are used to generate the voice output. The tone is determined by calculating "TILT". "TILT" is directly calculated from F0. From acoustic aspect, it is directly represented by the shape of fundamental frequency (F0) contour. The tone shape is represented by,

$$tilt = \frac{|Arise| - |Afall|}{|Arise| + |Afall|} \tag{27}$$

Where,

A_{rise} =Amplitude of rise (in Hz)

A_{fall} =Amplitude of fall (in Hz)

5.4 Concatenation

A high quality speech synthesizer concatenates speech waveforms referenced by a large speech database. Speech quality is further improved by speech unit selection and concatenative smoothing (Tabet. Y. and Boughazi. M., 2011). Synthesized speech can be produced by combining pieces of pre-recorded phonemes/syllables that are stored in a database. The Intelligibility of the speech is very important quality for synthesized speech. Concatenative synthesis is based on the concatenation or combining the progressive segment of recorded speech together. In general, concatenative synthesis produces the most natural-sounding synthesized speech.

Concatenative TTS System produces very natural sounding speech. Since, they simply join pre-recorded segment or units of phonemes/syllables to form sentences (Mahwash Ahmed and ShibliNisar, 2014). Speech generated by this approach inherently possesses natural quality. The system generates speech by searching for appropriate combinations of syllabic/phonemic sound in a large database of human speech. The required sound fragments are found in database and hence joining or modifying certain speech unit can be avoided. The best combinations are found and they are concatenated which is shown in Fig. 46.

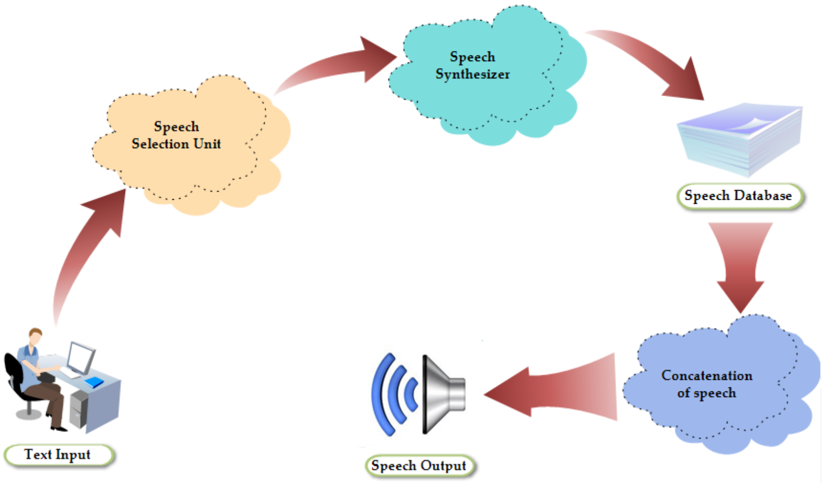


Fig. 46: Concatenative Speech Synthesis

The following steps have to be followed by the Synthesis machine to convert text into speech.

Step 1: The text dialogue is normalized as well as sentences are converted into words.

Step 2: Phonemic/syllabic text representation is obtained from the normalized text where phonemic representation falls under 44 sound categories.

Step 3: Phonemic text representation into phonemic clips and syllabic text representation into syllabic transcription are selected from speech database

Step 4: The phonemic/syllabic clips duration and pitch are changed according to their respective position.

Step 5: The modified sound clips are concatenated to form the individual words.

Step 6: These isolated words are combined with respective pauses within each word.

5.5 Smoothing

While concatenating the phoneme/syllable sound clips, discontinuities may exist between the speeches and the mismatch in spectra of the speech units causes this discontinuity (David T. Chappell and John H.L. Hansen, 2002). To avoid such circumstance smoothing methods can be applied. For smoothing optimal coupling method is proposed in this work. Initially optimal concatenating point (OCP) is to be identified. OCP is identified by choosing the minimal zero crossing rate of the hamming window on the selected portion of boundaries. If the given syllabic unit/phoneme unit lies at the beginning of the boundary then it is set from 5n/6th position, else the boundary is set from the beginning to n/3rd position. Hamming windows are generated to process on the selected portions which helps the system to find the concatenating point.

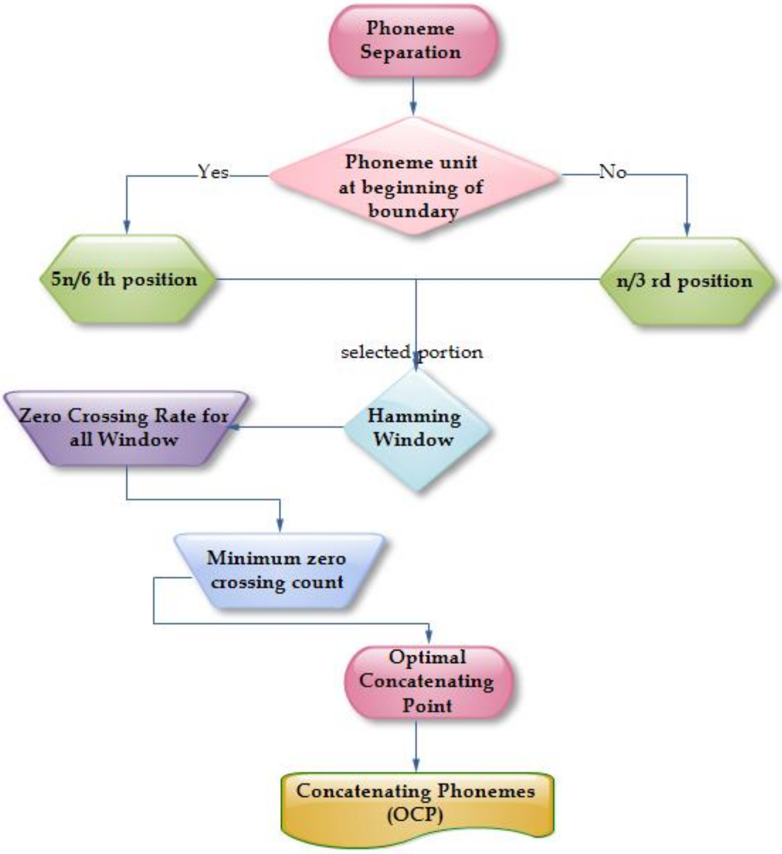

Phoneme Separation
Phoneme unit at beginning of boundary
Yes
No
5n/6 th position
n/3 rd position
selected portion
Hamming Window
Zero Crossing Rate for all Window
Minimum zero crossing count
Optimal Concatenating Point
Concatenating Phonemes (OCP)

Fig. 47: Flowchart for Optimal Coupling Algorithm

Zero crossing counts for all the hamming windows are analyzed and the point of minimum count is considered as optimal concatenating point. Finally the phonemes/syllabic sound clips are combined at the chosen points to generate the smoothened speech output as shown in Fig. 47.

5.6 Database

A set of 400 isolated words, 2800 sentences and 1750 paragraphs of speeches are considered for evaluation from which 180 isolated words, 1760 sentences and 960 paragraphs are selected as training set from the speech database and 220 isolated words, 1040 sentences and 790 paragraphs are selected as testing set which are not included or trained in the training set. Training is given for every data used in the training stage. A set of 1760 sentences and 960 paragraphs are selected from the speech database which is not tested in the training set. Table 14 shows the dataset for SS system. Rank of the voice quality decides the quality of the synthetic voice.

Table 14: Dataset for Speech Synthesis System

No. of Isolated word	400
No. of Sentences	2800
No. of Paragraphs	1750
Total No. of Training Speech Samples	2900
Total No. of Testing Speech Samples	2050

5.7 System Evaluation

The quality of the synthetic voice is measured using formal listening tests. In the initial stage non-structural to structural conversion is performed using text normalization. Table 15 shows the sample conversion of normal structure to normalized structure.

Initially, the given input is normalized and paragraph or sentences are broken up into the isolated words. Further, phonemes present in the words are analyzed. After finding the phonemes, the appropriate position is also analysed. Finally, waveform of pronounced phonemic units the sound clips are combined together to form natural speech.

Table 15: Text Normalization for non-structure representation

Expression Type	Normal Text (Unstructured)	Normalized Text (Structured)
Date	06/08/1988	Sixth August Nineteen Eighty Eight 6<digit>, th<Cardinal suffix>, august<word>,1988 <digit>
Tel Number	9876543210	Tel{Nine, Eight, Seven, Six, Five, Four, Three, Two, One, Zero} Individual Numbers has to be read (int)
Vehicle Number	TN 45	T<word_char>, N<word_char> Forty Five <digit>
Time	Hour: Minute: sec 09:45:49	Time{Nine Hours Forty Five Minutes and Forty seconds}

After concatenating the phoneme sound clip, it is noted that discontinuities exists between the speech outputs. To overcome such problems, optimal concatenating point is identified by choosing the minimal zero crossing count of the hamming window in the selected boundary.

The concatenated speech is tested by forty three people. The average scores given by these persons are considered for ranking and this score is termed as Mean Opinion Score (MOS). Three categories of speeches are considered for measuring the quality of speech. Speech synthesis using 3 minutes, 5 minutes and 10 minutes of speech were heard by different users.

The quality of the speech is measured by the rank (ranges from 1 to 4 grades) given by these 43 users. Rank is measured based on 4 different scales namely average (A), precision (PR), pleasantness (PL) and naturalness (NL).

Fig. 49, 50 and 51shows the Snapshots for text to speech conversion.

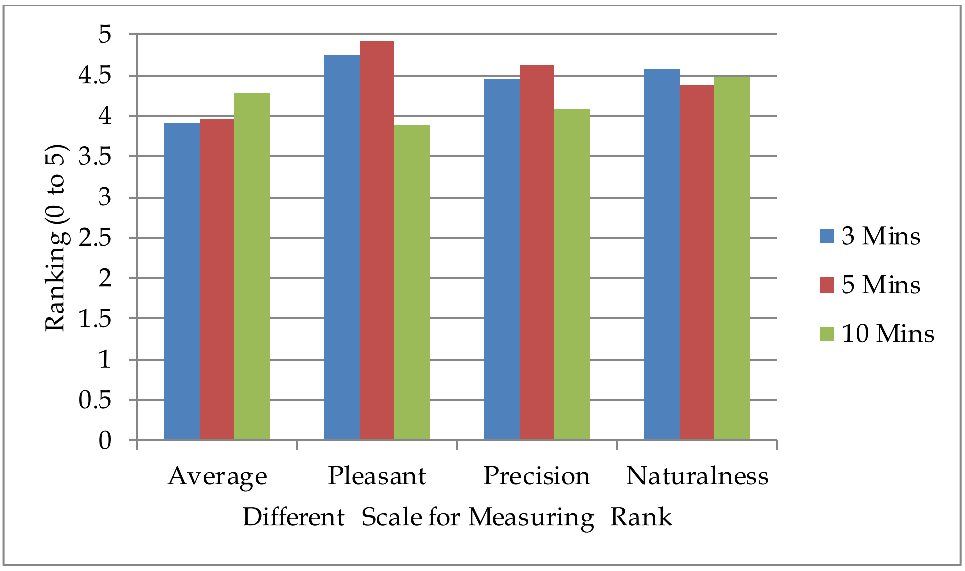
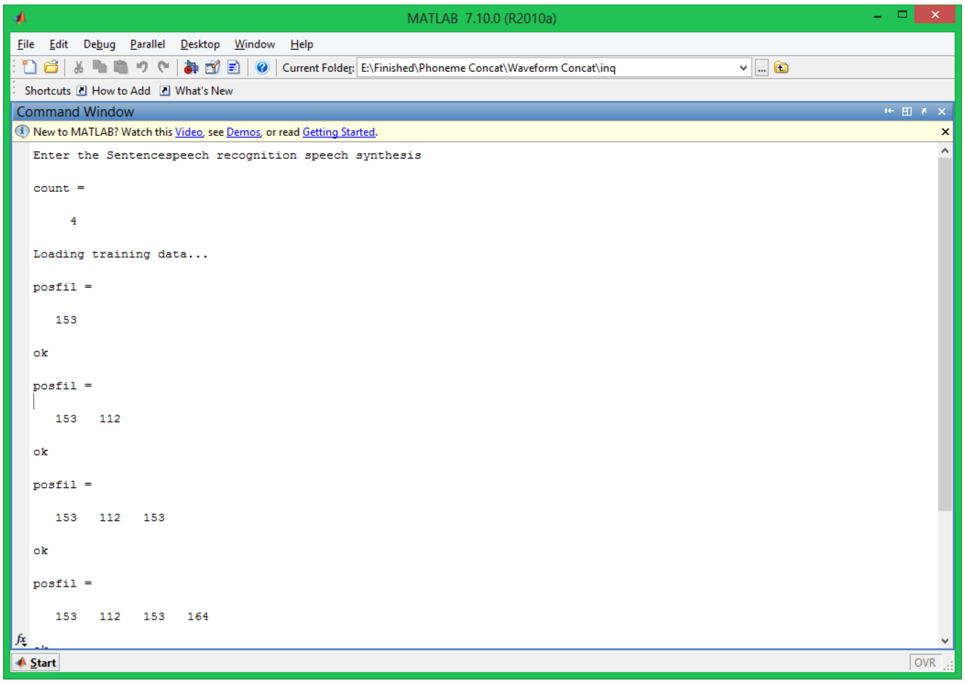


Fig. 48: Mean Opinion Score for Synthesized Speech


MATLAB 7.10.0 (R2010a)
File
Edit
Debug
Parallel
Desktop
Window
Help
Current Folder: E:\Finished\Phoneme Concat\Waveform Concat\inq
Shortcuts
How to Add
What's New
Command Window
New to MATLAB? Watch this Video, see Demos, or read Getting Started.
Start
OVR

```
Enter the Sentencespeech recognition speech synthesis

count =

     4

Loading training data...

posfil =

   153

ok

posfil =

   153   112

ok

posfil =

   153   112   153

ok

posfil =

   153   112   153   164
```

Fig. 49: Snapshot for Text to Speech (Identifying the position of query input)

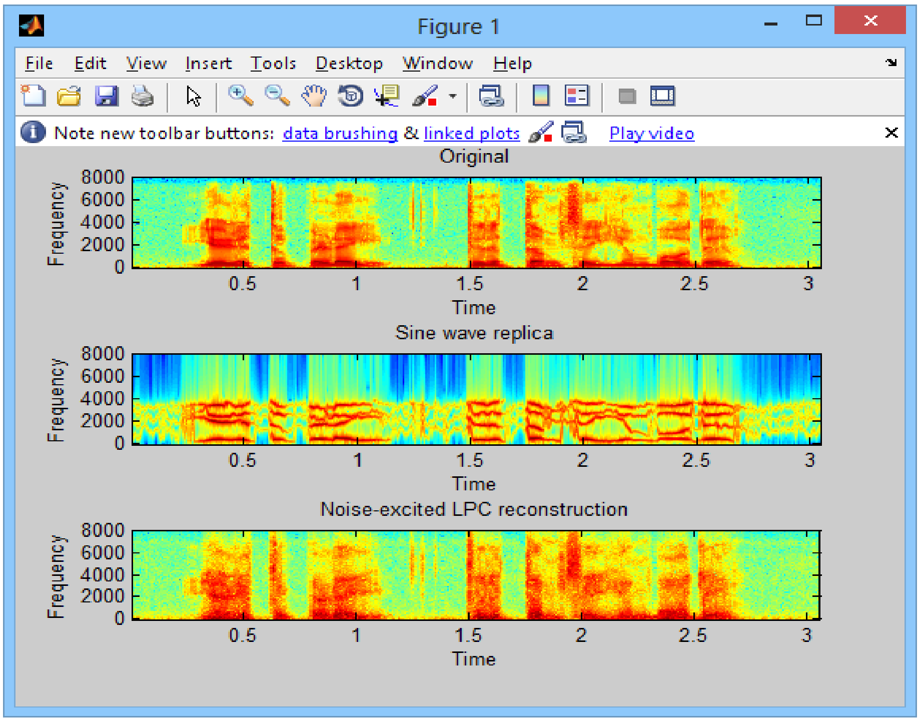
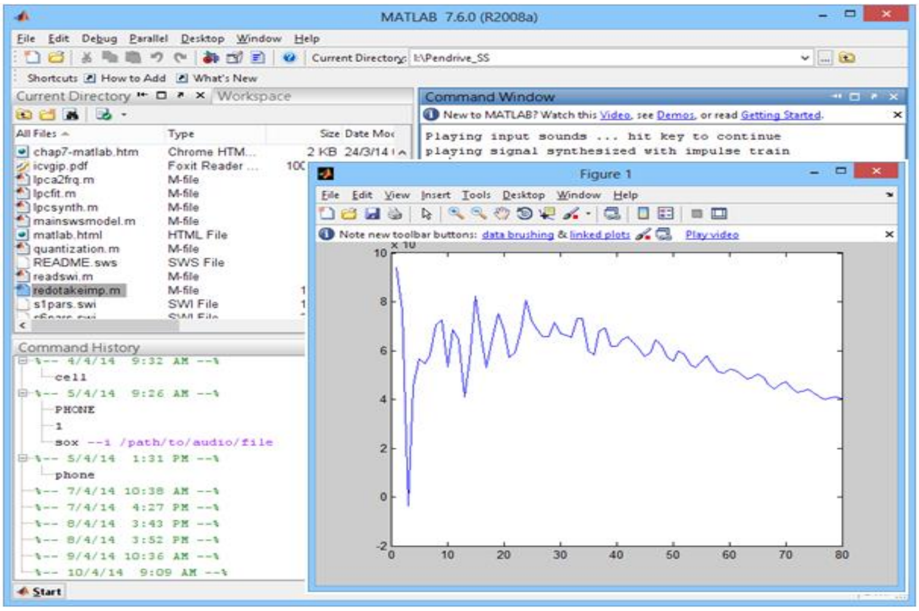


Fig. 50: Snapshot of generated Sine wave and Noise-excited Reconstruction for the text query


MATLAB 7.6.0 (R2008a)
File Edit Debug Parallel Desktop Window Help
Current Directory: I:\Pendrive_SS
Shortcuts How to Add What's New
Current Directory
Workspace
Command Window
New to MATLAB? Watch this Video, see Demos, or read Getting Started.
Command History
Figure 1
File Edit View Insert Tools Desktop Window Help
Note new toolbar buttons: data brushing & linked plots Play video
x 10
10
8
6
4
2
0
-2
0
10
20
30
40
50
60
70
80
Start

All Files	Type	Size	Date Mod
chap7-matlab.htm	Chrome HTM...	2 KB	24/3/14
icvgip.pdf	Foxit Reader ...	10(	
lpca2frq.m	M-file		
lpcfit.m	M-file		
lpcsynth.m	M-file		
mainswsmodel.m	M-file		
matlab.html	HTML File		
quantization.m	M-file		
README.sws	SWS File		
readswi.m	M-file		
redotakeimp.m	M-file	1	
s1pars.swi	SWI File	1	

```
Playing input sounds ... hit key to continue
playing signal synthesized with impulse train
```

```
%-- 4/4/14  9:32 AM --%
   cell
%-- 5/4/14  9:26 AM --%
   PHONE
   1
   sox --i /path/to/audio/file
%-- 5/4/14  1:31 PM --%
   phone
%-- 7/4/14 10:38 AM --%
%-- 7/4/14  4:27 PM --%
%-- 8/4/14  3:43 PM --%
%-- 8/4/14  3:52 PM --%
%-- 9/4/14 10:36 AM --%
%-- 10/4/14  9:09 AM --%
```

Fig.51: Snapshot for playing the sound of input text with impulse train

5.8 Summary

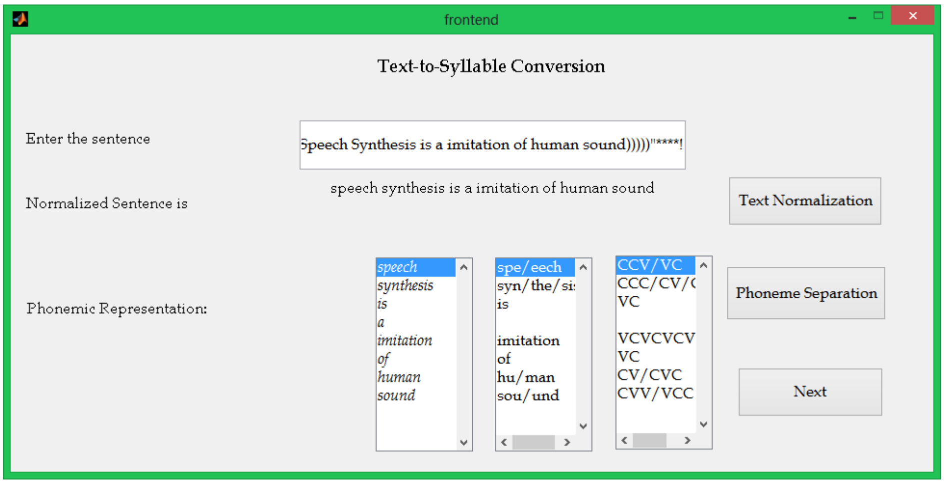
Speech synthesis system is evaluated using three categories of speeches for measuring its quality. MOS for speech synthesis using 3 minutes, 5 minutes and 10 minutes of speech were heard by different users. The quality of the speech is measured by the MOS (ranges from 1 to 4 grades) given by these 43 users. Rank is measured based on the 4 point different scales.

6 Experimental Results of Speech based Spoken Document Retrieval

The performance of spoken document was evaluated using Mean Opinion Score (MOS). Datasets were increased to 6700 documents and forty three graduate students were invited to evaluate the performance of keyword based spoken document retrieval. These 43 persons were requested to subjectively evaluate the results using MOS. MOS ranges from 1 to 10 different levels of grade with two categories: speech based document retrieval (SDR) and favourite (FAV); the evaluator selects the level of grade based on the naturalness and pleasantness of speech output. SDR is decided by the relation between query and the spoken document. FAV depends on the person's favourite topic of interest. Table 16 illustrate the overall performance of the speech based spoken document retrieval for system.

Table 16: Subjective Evaluation of Speech based Spoken Information Extraction on 10-point scale

	SDR	FAV
MOS	9.24	8.87



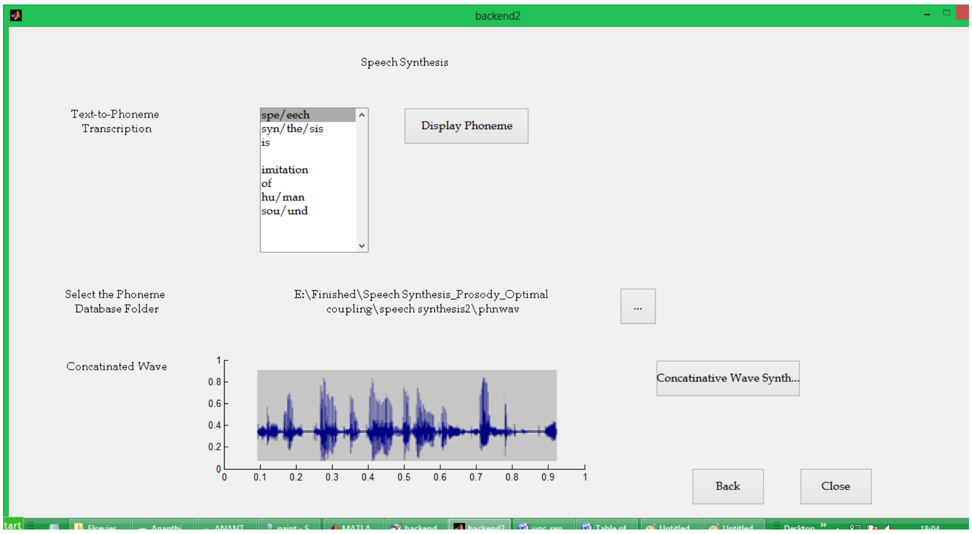
Enter the sentence: Speech Synthesis is a imitation of human sound)))))"****!

Normalized Sentence is: speech synthesis is a imitation of human sound

Phonemic Representation:

```
speech       spe/eech     CCV/VC
synthesis    syn/the/si   CCC/CV/(
is           is           VC
a
imitation    imitation    VCVCVCV
of           of           VC
human        hu/man       CV/CVC
sound        sou/und      CVV/VCC
```

Fig. 52: Snapshot of phoneme separation in speech synthesis



Text-to-Phoneme Transcription:

```
spe/eech
syn/the/sis
is

imitation
of
hu/man
sou/und
```

Select the Phoneme Database Folder: E:\Finished\Speech Synthesis_Prosody_Optimal coupling\speech synthesis2\phnwav

Fig. 53: Snapshot of SS system with its phonemic transcription and speech wave

6.1 Summary and outcome of the project:

Speech based spoken document retrieval system is implemented using speech recognition, document retrieval and speech synthesis system. The main intention of developing this system is to satisfy the need of the visually impaired persons while using the search engines. The computational efficiency of the speech recognition system, document retrieval system and speech synthesis was evaluated individually and collectively their results are analyzed. A system has been developed to convert spoken word into text using HMM modeling technique. Voice Activity Detection (VAD) is used for separating individual words out of the continuous speeches. Features for each isolated word are extracted and 11,000 models were trained successfully. HMM is used to model each individual utterance. Each isolated word segment from the test sentence is matched against the 11,000 models generated by the SR system for finding the semantic representation of the test input speech. The recognition system shows an average accuracy of 96.65% for HMM.

Ontology assisted VSM and genetic based retrieval system returns an ordering of document over the collection of document for the required speech query. Genetic approach with cosine similarity provides better performance than other relevance retrieval approaches. VSM based document retrieval system is analyzed using 387 queries over 243 documents and the quality of the speech was measured using MOS which is collected from 43 persons with four different scales. The system shows an accuracy of 95.87% as F1 score and 96.06% as G-measure. GA based retrieval system is evaluated using 20NG dataset with 19,999 documents and 734 user preferred document collection. The system shows an accuracy of 98.84% for cosine based similarity measure. Speech synthesis system is evaluated using three categories of speeches for measuring its quality. MOS for speech synthesis using 3 minutes, 5 minutes and 10 minutes of speech were heard by different users.

The quality of the speech is measured by the MOS (ranges from 1 to 4 grades) given by these 43 users. Rank is measured based on the 4 point scale. Speech based spoken document retrieval system is analyzed for 11,000 speech models for recognition and 4950 utterances for synthesis with 6700 documents using MOS on a 10 pt scale. MOS are collected from 43 persons (both visually impaired and normal persons) with two categories SDR and FAV and their results are analyzed. By implementing this type of retrieval system the impaired persons can make use of the internet facilities effectively.

REFERENCES

Ajimi Ameer, SreeKumar.K and Minu K.K. (2014), ‘Content Based Image Retrieval of User’s Interest Using Feature Fusion and Optimization Using Genetic Algorithm: A Survey’, *International Journal of Computer Science and Information Technologies,* vol. 5 (6), pp.:7885-7888.

Al-HaddadS.A.R., SamadS.A., Hussain A, Ishak K.A. and NoorA.O.A. (2009), ‘Robust Speech Recognition Using Fusion Techniques and Adaptive Filtering’, *American Journal of Applied Sciences 6 (2)*: 290-295.

Anusuya, M.A., Katti, S.K. (2011): ‘Front end analysis of speech recognition: A review’, *Int J Speech Technology,* 14: 99 -145.

Anubha Jain, Swati V. Chande and Preeti Tiwari (2014), ‘Relevance of Genetic Algorithm Strategies in Query Optimization in Information Retrieval’, *International Journal of Computer Science and Information Technologies,* Vol. 5 (4), 5921-5927.

Belkin, N.J. and Marchetti, P.G. (2001) ‘Determining the functionality and features of an intelligent interface to an information retrieval system,, *J.L. Vidick (ed.), proc. of the 13th International Conference on Research and Development in Information Retrieval,* Brussels, Presses Universitaires de Bruxelles, pp-151-177.

Berlin Chen, Yi-Wen Chen, Kuan-Yu Chen, Hsin-Min Wang and Kuen-Tyng Yu (2014), Enhancing Query Formulation for Spoken Document Retrieval, *Journal of Information Science and Engineering 30,* pp.:553-569.

Berlin Chen, Pei-Ning Chen and Kuan-Yu Chen (2011), Query Modeling for Spoken Document Retrieval, *IEEE,* pp.: 389-394.

Bishnu Prasad Das, Ranjan Parekh (2012): Recognition of Isolated words using features based on LPC, MFCC, ZCR and STE with Neural Network Classifiers. *International Journal of Modern Engineering Research (IJMER),* Vol.2, Issue.3, 854-858.

BjörnSchuller, Zixing Zhang, Felix Weninger and Felix Burkhardt (2012) ‘Synthesized speech for model training in cross-corpus recognition of human emotion’, *Int J Speech Technol,* 15, pp-313–323.

Campbell N. (2006) ‘Conversational speech synthesis and the need for some laughter’, *IEEE Transactions on Audio, Speech, and Language Processing,* 17(4), 1171–1179.

Campbell, N., Hamza, W., Hog, H and Tao, J. (2006) ‘Editorial special section on expressive speech synthesis’, *IEEE Transactions on Audio, Speech, and Language Processing,* 14, 1097–1098.

Carlson, R., and Granström, B. (2007) ‘Rule-based Speech Synthesis’*, in Benesty, J., Sondhi, M. M., & Huang, Y. (Eds.),Springer Handbook of Speech Processing,* 429-436, Springer Berlin Heidelberg.

Chalamandaris A., Tsiakoulis P., Karabetsos S. and Raptis S. (2009), ‘An efficient and robust pitch marking algorithm on the speech waveform for TD-PSOLA’, *ICSIPA, IEEE,* 397 - 401.

Chapelle, O., Vapnik, V., Bousquet O., Mukherjee S. (2002): Choosing multiple parameters for support vector machines,*Machine Learning,* 46, 131-159.

David T. Chappell and John H.L. Hansen (2002), ‘A comparison of spectral smoothing methods for segment concatenation based speech synthesis’, *Speech Communication* 36, 343-374.

Davide Buscaldi, Paolo Rosso, José Manuel Gómez-Soriano and Emilio Sanchis (2009), ‘Answering questions with an n-gram based passage retrieval engine’, *J Intell Inf Syst.*

Duan K., Keerthi S., and Poo A. (2003): Evaluation of simple performance measures for tuning SVM hyperparameters. *Neuro computing*, 51, 41-59.

Geir Solskinnsbakk and Jon AtleGulla (2010), 'Combining ontological profiles with context in information retrieval', *Data & Knowledge Engineering*, 69, 251–260.

Goldberg, D.E.. (2003), 'Genetic Algorithms in Search, Optimization and Machine Learning', *Pearson Education, New Delhi,* ISBN-10: 0201157675.

Govind D. and Mahadeva Prasanna S.R. (2013), 'Expressive speech synthesis: a review', *Int J Speech Technol.*, 6: pp. 237–260.

Hocine Bourouba, Mouldi Bedda, Rafik Djemil (2006): Isolated words recognition system based on Hybrid approach DTW/GHMM. *Informatica 30,* 373-384.

Jinn-TsongTsai (2014), 'Optimized weights of document keywords for auto-reply accuracy', *Neurocomputing,* 124, Elsevier, 43-56.

John Lafferty and Cheng xiang Zhai (2001), 'Document Language Models, Query Models, and Risk Minimization for Information Retrieval', *SIGIR '01 proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval,* New Orleans, Louisiana, USA, 111-119.

Jong Kuk Kim, HernSoo Hahn and Myung Jin Bae (2009), 'On a Speech Multiple System Implementation for Speech Synthesis', *Wireless Pers. Comm.,* 49, 533–543.

José A. González, Antonio M. Peinado, Ning Ma, Angel M. Gómez, and Jon Barker (2013), 'MMSE-Based Missing-Feature Reconstruction with Temporal Modeling for Robust Speech Recognition', *IEEE Transactions on Audio, Speech and Language Processing,* Vol. 21, No. 3., 624-635.

Kuan-Yu Chen and Berlin Chen (2010), 'A Study of Topic Modeling Techniques for Spoken Document Retrieval', *Proceedings of the Second APSIPA Annual Summit and Conference,* pp.: 237–242.

Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer, 'Estimating Hidden Markov Model Parameters so as to Maximize Speech Recognition Accuracy', *IEEE Transactions on Audio, Speech and Language Processing,* Vol. 1, No. 1, 77-83.

LI Guoliang, FENG Jianhua and ZHOU Lizhu (2009), 'Keyword Searches in Data-Centric XML Documents Using Tree Partitioning', *TSINGHUA Science and Technology,* vol. 14, no.1, pp.:7-18.

Louis S. Wang (2009), Relevance Weighting of Multi-Term Queries for Vector Space Model, CIDM, *IEEE,* 396-402.

Mahwash Ahmed and ShibliNisar (2014),'Text-to-Speech Synthesis using Phoneme Concatenation', *International Journal ofScientific Engineering and Technology,* Vol. No.3 (2), 193-197.

Mark S. Hawley, Stuart P. Cunningham, Phil D. Green, Pam Enderby, Rebecca Palmer, Siddharth Sehgal, and Peter O'Neill (2013), 'A Voice-Input Voice-Output Communication Aid for People With Severe Speech Impairment', *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* Vol. 21, No. 1, 23-31.

Nadia, L. (2014), 'Design and Implementation of Information Retrieval System based Ontology', *ICMCS, IEEE,* 500-505.

Peiyun Zhang and RongjianXie, (2009), 'Ontology-based Unstructured Information Organization and Retrieval', *World Congress on Software Engineering, IEEE,* 408-411.

Philomina Simon and Siva Sathya (2009), 'Genetic Algorithm for Information Retrieval', *IEEE.*

Sabato Marco Siniscalchi, TorbjørnSvendsen and Chin-Hui Lee (2013), 'A Bottom-Up Modular Search Approach to Large Vocabulary Continuous Speech Recognition', *IEEE Transactions on Audio, Speech and Language Processing,* Vol. 21, No. 4, 786-797.

Sandeep Kumar, S. Bhattacharya, Vishal Dhiman and Shuvashree Mohapatra (2013), 'Performance evaluation of a wavelet-based pitch detection scheme', *Int J Speech Technol.* 16:pp.431–437.

Shreekanth.T, Udayashankara.V and ArunKumar.C (2014), 'An Unit Selection based Hindi Text To Speech Synthesis System Using Syllable as a Basic Unit', *IOSR Journal of VLSI and Signal Processing (IOSR-JVSP,*Volume 4, Issue 4, Ver. 2, 49-57.

Stephan Philippi and Jacob Köhler (2004), 'Using XML Technology for the Ontology-Based Semantic Integration of Life Science Databases', *IEEE Transactions on Information Technology In Biomedicine*, vol. 8, no. 2, pp. 154-160.

Sudhakar Sangeetha and Sekar Jothilakshmi (2013), 'Syllable based text to speech synthesis system using auto associative neural network prosody prediction, *Int J Speech Technol.,* 17: pp. 91–98.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn and Supachanun Wanapu (2013), 'Using of Jaccard coefficient for keywords similarity, *Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS),* vol. 1.

Thiang, Wijoyo S. (2011): Speech recognition using linear predictive coding and artificial neural network for controlling movement of mobile root, *International Conference on Information and Electronics Engineering (IPCSIT),* vol. 6, 179-183.

Thiruvengatanadhan R, Dhanalakshmi P, and Palanivel S., (2014), "GMM Based Indexing and Retrieval of Music Using MFCC and MPEG-7 Features," *Emerging ICT for Bridging the Future, Springer Advances in Intelligent Systems and Computing,* Volume 1, pp. 363-337.

Ting KM. (2002) 'An instance-weighting method to induce cost-sensitive trees. *IEEE Transactions on Knowledge Data Eng.,* 14, 659–665.

Turney, P. and Pantel, P., (2010), 'From frequency to meaning: Vector space models of semantics', *Journal of Artificial Intelligence Research37,* 141–188.

Utpal Bhattacharjee (2013): A comparative Study of LPCC and MFCC Features for the Recognition of Assamese Phonemes, *International Journal of Engineering Research and Technology (IJERT),* Vol.2, Issue: 1, January 2013, 1-6.

Vikas Thada and Dr. VivekJaglan (2013), 'Comparison of Jaccard, Dice, Cosine Similarity Coefficient to find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm', *International Journal of Innovations in Engineering and Technology,* vol.2, issue 4, 202-205.

Wafa Maitah, Mamoun. Al-Rababaa and Ghasan. Kannan (2013), 'Improving the effectiveness of Information retrieval system using Adaptive genetic algorithm', *International Journal of Computer Science & Information Technology (IJCSIT),* Vol. 5, No. 5, pp.:91-105.

Yi-Wen Chen, Kuan-Yu Chen, Hsin-Min Wang and Berlin Chen (2013), Effective Pseudo-Relevance Feedback for Spoken Document Retrieval, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp.: 8535 - 8539.

Zolnay, A., Schulueter, R. Ney, H. (2005): Acoustic feature combination for robust speech recognition, *IEEE Transactions on Acoustics, Speech and Signal Processing,* 457-460.

LIST OF PUBLICATIONS

[1] S. Ananthi, P. Dhanalakshmi, “Literature Review on Speech Based Information Retrieval through web for visually impaired”, *International Journal of Advanced Research in Computer Science (IJARCS),* ISSN No. 0976-5697 Volume 3, No. 7, Nov-Dec 2012, pp.: 1-6.

[2] S. Ananthi, P. Dhanalakshmi, “Survey about Speech Recognition and Its Usage for Impaired (Disabled) Persons”, *International Journal of Scientific & Engineering Research (IJSER),* ISSN 2229-5518 Volume 3, Issue 13, Februvary-2012, pp.: 1-7.

[3] S. Ananthi, Dr. P. Dhanalakshmi, “Speech Recognition System and Isolated Word Recognition based on Hidden Markov Model (HMM) for Hearing Impaired”, *International Journal of Computer Application (IJCA),* ISSN 0975 8887, July- 2013 pp.: 30-34.

[4] S. Ananthi and P. Dhanalakshmi, “Speech Synthesis Techniques for Text to Speech Conversion Based on Concatenative Speech Synthesis”, *Eighth International Multi-Conference on Information Processing (IMCIP-2014), Digital Image and Signal Processing* - Elsevier Science and Technology Publications, July 2014, pp.: 360-369. (Elsevier Proceedings)

[5] S. Ananthi and P. Dhanalakshmi, “SVM and HMM Modeling Techniques for Speech Recognition Using LPCC and MFCC Features”, *Advances in Intelligent Systems and Computing Volume 327,* Springer International Publishing, 2015, pp.: 519-526, 2015. (Springer)

[6] S. Ananthi, P. Dhanalakshmi, “Syllable based Concatenative Synthesis for Text to Speech Conversion”, *Smart Innovation, Systems and Technologies 33, Springer*, 2015, pp.: 65-73. (Springer)

[7] S. Ananthi and P. Dhanalakshmi, “A Novel approach for Text Information Retrieval using Vector Space Model”, *International Journal of Applied Engineering Research,* Vol. 9, No. 21, pp.: 4865-4870, 2014. (Annexure – II)