Journal of Engineering Science and Technology
Vol. 11, No. 6 (2016) 806 - 825

CONVERTING RETRIEVED SPOKEN DOCUMENTS INTO TEXT USING AN AUTO ASSOCIATIVE NEURAL NETWORK

J. SANGEETHA*, S. JOTHILAKSHMI

Department of computer science and Engineering, Annamalai University, Annamalai Nagar, Chidambaram - 608002
*Corresponding Author: sangita.sudhakar@gmail.com

Abstract

This paper frames a novel methodology for spoken document information retrieval to the spontaneous speech corpora and converting the retrieved document into the corresponding language text. The proposed work involves the three major areas namely spoken keyword detection, spoken document retrieval and automatic speech recognition. The keyword spotting is concerned with the exploit of the distribution capturing capability of the Auto Associative Neural Network (AANN) for spoken keyword detection. It involves sliding a frame-based keyword template along the audio documents and by means of confidence score acquired from the normalized squared error of AANN to search for a match. This work benevolences a new spoken keyword spotting algorithm. Based on the match the spoken documents are retrieved and clustered together. In speech recognition step, the retrieved documents are converted into the corresponding language text using the AANN classifier. The experiments are conducted using the Dravidian language database and the results recommend that the proposed method is promising for retrieving the relevant documents of a spoken query as a key and transform it into the corresponding language.

Keywords: Spoken Document Retrieval, Spoken keyword spotting, Mel frequency cepstral coefficients, Auto associative neural networks, Continuous speech recognition, Automatic speech segmentation, Zero crossing rate, Short time energy.

1. Introduction

There is now prevalent use of Information Retrieval (IR) techniques to access information stored in electronic texts. One of most broadly used examples of the IR, is in internet search engines. However, there is also much information enclosed in documents that are not originally created as text but are verbal. One

Nomenclatures

a	adjustable parameter
CS	Speech corpus
fl	l^{th} frame
M	frame shift
Mel(f)	mel scale frequency
Mf	number of filters in the mel filter bank
N	frame size
o	Output vector
$s'(n)$	Preemphasized signal
$s(n)$	input speech signal
SK	Keyword signal
S_{max}	Global maximum
t_s	threshold
$w(.)$	window function
Wp	P^{th} analysis window

Greek Symbols

α	Scaling factor.
$\sum$	covariance matrix

Abbreviations

AANN	Auto Associative Neural Network
ASR	Automatic Speech Recognition
ATWV	Actual Term-Weighted Value
BNEWS	Broadcast News
CTS	Conversational Telephony Speech
FOM	Figure Of Merit
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
KS	Keyword Spotting
MFCC	Mel Frequency Cepstral Coefficient
MTWV	Maximum Term Weighted
OCC	Occurrence-Weighted Value
REALCONFSP	REAL time CONFerence room meetings
STD	Spoken Term Detection
WER	Word Error Rate
WRR	Word Recognition Rate

such area is the audio associated with radio and TV news broadcasts. If these audio sources could be transcribed spontaneously then the information they contain can be indexed and relevant portions of broadcasts retrieved using conventional IR techniques. Finding appropriate information in oral documents is a challenging task for recent multimedia information systems [1] because speech contents (spoken documents) are problematic to search, summarize or browse, which is a huge difference from text-based contents. We should listen to the complete part of a spoken document to comprehend it. This drawback has prohibited us from exploiting spoken documents on the Internet. To facilitate easier admission of spoken documents, spoken document retrieval methods have been developed in [2], [3].

There are a variety of advantageous data organization and retrieval techniques that can be applied to gatherings of There audio documents. These comprise document clustering (where collections of topically associated documents are bundled together), document link detection (where document couples which are topically correlated or linked are recognized), and query-by example document retrieval (where documents which are typically connected to an example or query document are itemized and graded). The spoken document retrieval methods assist us to start a search query based on keywords and search spoken document that match the query. The conventional method for the retrieval problem is to make use of an Automatic Speech Recognizer (ASR) incorporated with the typical information retrieval method. However, ASRs tend to produce transcripts of spontaneous speech with the significant word error rate, which is an undesirable trait of standard retrieval system. To victorious over such a constraint, we propose a method for spoken document retrieval based on spoken keyword spotting using the auto associative neural networks.

Speech recognition technology has remarkable potential as it is an essential part of future intelligent devices, where automatic speech recognition and text to speech synthesis are used as the elementary means of communicating with humans. It will streamline the Herculean task of typing and will eliminate the conventional keyboard. This speech technology enhances a lot in manufacturing and control applications where there is occupation for hands and eyes. Disabled, elderly and blind people will no longer need to be away from the internet and the information technology revolution. Recently, there has been a huge increase in the number of recognition applications for practice over telephones, including, operator assistance, automated dialling and remote data access services; such as financial services, for voice dictation systems like medical transcription applications. Such tantalizing applications have initiated research in automatic speech recognition since 1950's. Speech is a natural and simple communication method for human beings. A speech interface to the computer is the next big step that computer science needs to take for general users. Speech recognition will play an important role in taking technology to them. The need is not only for speech interface, but speech interface in local languages. It is an extremely complex and difficult job to make a computer respond to spoken commands in local languages in India. Recently there has been a momentous need for continuous speech recognition system to be developed in the local languages in India.

In this proposed work spoken query keyword is given. Acoustic keyword spotting aims at identifying any specified keyword in spoken sounds. Keyword spotting is a technologically significant problem, playing an essential role in audio indexing, voice mail retrieval, voice command detection, spoken term detection & document retrieval and speech data mining applications. Then the retrieved documents are clustered together and fed into the continuous speech recognition system. The proposed ASR system comprises of three steps namely pre-processing, feature extraction and classifications. In the pre-processing step, the input signal is pre-processed through the steps such as pre emphasis filter, framing and windowing, in order to remove the background noise and to enrich the signal. The best filtered and the enriched signal from the pre-processing step is taken as the input for the further process of ASR system. The speech features being the most essential segment in speech recognition system, are analysed and extracted via Mel Frequency Cepstral Coefficients (MFCC). These feature vectors are given as the input to the classifiers such as an auto associative neural network for classifying and recognizing the languages. Experiments are carried out with Dravidian languages such as Tamil and Malayalam speech signals. Fig. 1 shows the overview of spoken document retrieval system.

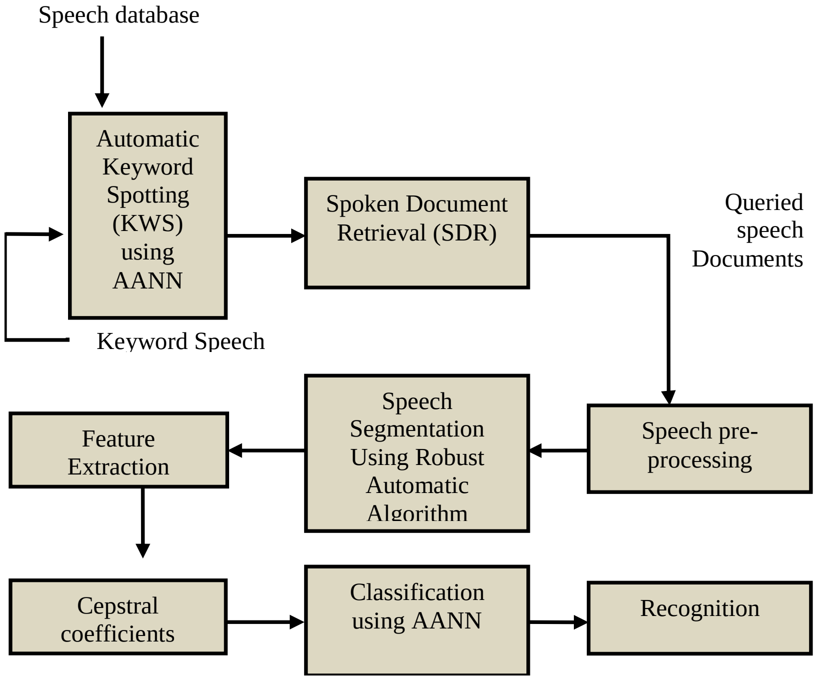

Fig. 1. System overview of the proposed system.

The main benefit of this proposed system is its ability to achieve hands-free computing. It also offers huge social benefits for people with disabilities who have difficulties in using a keyboard and the mouse. Thus, it has become an attractive alternate choice for many users to manage applications through speech rather than a mouse or keyboard. The main application of this system includes voice dialling, call routing, automatic transcriptions, information searching, data entry, speech-to-text processing and aircraft etc.

The rest of the paper is organized as follows: Section 2 presents the related work. A brief description about the method of extracting the features for spoken document retrieval and continuous speech recognition from the speech signal is described in Section 3. Auto associative neural networks model for capturing the distribution of acoustic feature vectors is given in Section 4. The proposed algorithm for spoken document retrieval system and continuous speech recognition is presented in Section 5. Section 6 presents the performance measures for the proposed system. Section 7 presents the experimental results. Section 8 gives the conclusions and describes the future work.

2. Related Work

Several approaches to this problem have been proposed in the literature [4]. Investigating the task of spotting predefined keywords in continuous speech has both practical and scientific motivations. Even in situations where little access to non-lexical linguistic constraints is provided (e.g. spotting native words in an unfamiliar language). Several computational approaches to this problem have been proposed. One of the first keyword spotting strategies, proposed by [5], involved sliding a frame-based keyword template along the speech signal and using a nonlinear dynamic time warping algorithm to proficiently search for a match. While the word models in later approaches changed significantly, this sliding model strategy was used in other approaches [6], [7].

An unsupervised learning framework has been proposed [8] to address the problem of detecting spoken keywords. Without any transcription information, a Gaussian Mixture Model (GMM) is trained to label speech frames with a Gaussian posterior gram. Given one or more spoken example of a keyword, they use segmental dynamic time warping to compare the Gaussian posterior grams between keyword samples and test utterances. A standard Hidden Markov Model (HMM) based method is the key word filler model. In this case, an HMM is constructed of three components: a keyword model, a background model, and a filler model. The keyword model is tied to the filler model, which is typically a phone or broad class loop, meant to represent the non-keyword portions of the speech signal. Finally, the background model is used to normalize keyword model scores.

A Viterbi decode of a speech signal is performed using this keyword-filler HMM, producing predictions when the keyword occurs. Variations of this approach are provided by [9], [10], [11]. The main research effort is focused on defining specialized confidence measures that maximize performance. Examples include [12], [13], [14] and [15]. While these systems do not require a predefined vocabulary, they rely on language modelling and are thus highly tuned to the training environment. Continuous speech recognition is still a demanding field of research in the area of digital signal processing due to its versatile applications. In spite of the improvements made in this area, machines cannot tie the performance of human beings in terms of accuracy and speed particularly in the case of speaker independent speech recognition systems. Since speech is the most important means of communication among people, research in automatic speech recognition and speech synthesis by machine has attracted a great deal of devotion over the past five decades [16]. Recent technological developments have made much progress in the recognition of complex speech patterns. But much more investigation and improvement is desirable in this field. The speech recognition system typically performs two important operations: signal modelling and pattern matching [17].

During signal modelling, speech signal is transformed into a set of parameters by a procedure called feature extraction. Pattern matching is the task of identifying parameter set from memory which narrowly matches the parameter set attained from the input speech signal also known as classification. Among these steps, feature extraction is a key, because enhanced feature is good for enlightening recognition rate. Recognition accuracy is an imperative measure to calculate the performance of a speech recognition system. There are numerous techniques available in literature to improve the efficiency of speech recognition systems. The modelling accuracy is to reduce the HMM conditional-independence assumption, and condition the distribution of each examination of the earlier studies in addition to the state that generates it. This technique is known as conditional Gaussian HMMs or autoregressive HMMs. However, it has been shown that the conditional Gaussian HMMs frequently do not provide an advantage if the dynamic features are used. There are other forms of approaches which explore the utilization of more difficult HMM structures, such as multiple-path modelling [18]. This formulation comprises of multiple parallel paths, each of which may be the reason for the acoustic variability from a specific source. The multiple-path prototype may over-correct the trajectory folding problem connected with the GMM-HMM, as there are acceptable mixture paths and they are minimized exponentials. Most of these systems have only been validated on certain simple recognition actions using a small number of parallel paths. But to develop a model that is principally robust to speaker and environmental alterations is quiet a challenging problem.

There have been certain noticeable advances in discriminative training such as Maximum Mutual Information (MMI) estimation [19], Minimum Classification Error (MCE) training [20], and Minimum Phone Error (MPE) training [21], in large margin approaches (such as large-margin estimation [22], [23], large-margin MCE [24],[25] and boosted MMI [26]]), as well as in novel acoustic models (such as Conditional Random Fields (CRFs) [27], hidden CRFs [28] and segmental CRFs [29]). The hidden Markov model technique is frequently considered as inaccurate to model heterogeneous data sources. The mixture segments that are attained in diverse acoustic conditions for one sound can be joined to match at a high probability with the speech observations from another sound, a problem denoted to as trajectory folding [30].

3. Feature Extraction

Mel frequency cepstral coefficients have proved to be one of the most successful feature representations in speech related recognition tasks [32]. The mel-cepstrum exploits auditory principles, as well as the de-correlating property of the cepstrum. The computation of MFCC is shown [33] in Fig. 2 and described as follows.

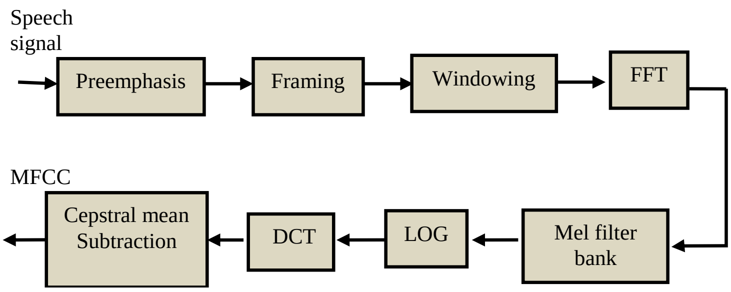

Fig. 2. Extraction of MFCC from speech signal.

3.1. Preemphasis

The digitized speech signals (n) are put through a low order digital system, to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing. The output of the preemphasis network, is related to the inputs (n), by the difference equation

$$\hat{s}(n) = s(n) - \alpha s(n-1)$$

The most common value for α is around 0.95.

3.2. Frame blocking

Speech analysis usually assumes that the signal properties change relatively slowly with time. This allows examination of a short time window of speech to extract parameters presumed to remain fixed for the duration of the window. Thus to model dynamic parameters, the signal must be divided into successive windows or analysis frames, so that the parameters can be calculated often enough to follow the relevant changes. In this step the preemphasized speech signal, ŝ (n) is blocked into frames of N samples, with adjacent frames being separated by M samples. If we denote the lth frame speech by xl(n), and there are L frames within the entire speech signal, then

$$x_l(n) = \hat{s}(Ml + n), n = 0,\ldots, N-1, l = 0,\ldots, L-1$$

3.3. Windowing

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and the end of the frame. The window must be selected to taper the signal to zero at the beginning and end of each frame. If we define the window as w (n), 0 ≤ n ≤ N - 1 then the result of windowing the signal is

$$\bar{x}_l(n) = x_l(n)w(n), 0 \le n \le N-1$$

The Hamming window is used for this work, which has the form

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1$$

3.4. Computing spectral coefficients

The spectral coefficients of the windowed frames are computed using Fast Fourier Transform, as follows:

$$X(k) = \sum_{n=0}^{N-1} \bar{x}_l(n) \exp^{-jk(2\pi/N)n}, \quad 0 \le n \le N-1$$

3.5. Computing mel spectral coefficients

The spectral coefficients of each frame are then weighted by a series of filter frequency response whose center frequencies and bandwidths roughly match those of the auditory critical band filters. These filters follow the mel scale whereby band edges and center frequencies of the filters are linear for low frequency and logarithmically increase with increasing. These are called as mel-scale filters and collectively a mel-scale filter bank . As can be seen, the filters used are triangular and they are equally spaced along the mile scale which is defined by

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

Each short term Fourier transform (STFT) magnitude coefficient is multiplied by the corresponding filter gain and the results are accumulated.

3.6 Computing MFCC

The Discrete Cosine Transform (DCT) is applied to the log of the mel spectral coefficients to obtain the MFCC as follows:

$$x(m) = \sqrt{\frac{2}{\mathrm{M}}} \sum_{i=0}^{M-1} E(i) \cos\left(\frac{(2x+1)m\pi}{2N}\right), m = 1, \ldots, M$$

where M is the number of filters in the filter bank, finally, cepstral mean subtraction is performed to reduce the channel effects.

4. Overview of the Proposed Work

The propose work consists of three important steps such as spoken keyword spotting, spoken document retrieval and automatic speech recognition. The spoken keyword spotting and spoken document retrieval involves the technical essentials offered in aforementioned sections. It is assumed that the acoustic features of the user's search speech keyword and the speech files present in the database has been extracted from the speech signal.

4.1. Proposed Spoken document retrieval algorithm

The outline of the algorithm is summarized as follows: After attaining the speech features for every single frame of the given search keyword speech signal, AANN model is trained to capture the distribution of this keyword speech signal. As well the features are gained for each frame of the speech signal present in the speech database which is to find the given keyword is presented or not. At first a block of frames in the input speech signal such that the number of frames in the block is equal to the number of frames of the search term keyword signal are selected starting from the first frame. This chunk of feature vectors is used to test the model. If the search word corresponding to the block of frames is same as the search keyword then the confidence score for the chunk will be very high. If the word corresponding to the block of frames is absolutely dissimilar from the keyword, the feature vectors from the block possibly will not fall into the distribution and the model gives low confidence (probability) score.

Similarly, the next possibility is the word corresponding to the block of frames is partly similar to the search keyword. If this is the case, the confidence score of the block will be amid the above two values. After obtaining the confidence score for the current block, the block is shifted by a fixed number of frames to the right. Then the entire process is reiterated for this fresh block and the confidence score is found. In the same way the confidence scores are measured up to the tail end of the block reaches the last frame of the input speech frames. From the confidence score, the global maximal positions are the locations for the search keyword in the input signal and they are detected using a threshold. If the keyword is present in the audio file, the corresponding audio file is extracted and stored separately. The above process is repeated for the total number of audio files present in the database in order to retrieve the spoken documents which is the search keyword is present. Finally all the retrieved files are organized in an order based on the number of occurrences of the keyword and clustered together for the user.

There is a collection of speech files in the corpus is CS = $\{cs_t : t = 1, 2,. . . . m\}$ where m is the total number of speech files in the corpus. Given the speech features of the each input speech signal present in the speech database is SF= {sfi: i = 1, 2, . . . ,n} where i is the frame index and n is the total number of frames in the input speech signal. As well the speech features of the keyword signal SK= $\{sk_j: j = 1, 2. . . z\}$ where j is the frame index and z is the total number of frames present in the keyword signal. The proposed algorithm for retrieving the speech files in accordance with the given search speech keyword term is summarized as follows:

1) From n frames, z numbers of frames are cautiously selected, and considered as analysis window W. WP is the pth analysis window which is given by

$$Wp = \{IS_l\}, p\leq l < m+ p$$

2) AANN is trained by means of the frames in SK and the model captures the distribution of this particular block of data. Then the feature vectors in WP are given as input to the AANN model and the output of the model is matched with the input to calculate the normalized squared error e_k. The normalized squared error (ek) for the feature vector y is given by

$$e_k = \frac{||y-0||}{y^2}$$

where o is the output vector given by the model. The error ek is transformed into a confidence score s using

$$s = \exp(-e_k)$$

3) The average confidence score is computed by adding the confidence score of the individual frames and the end result is divided by the total number of frames in the block. The experiments have been conducted with the weighted sum of the frame scores within the block and there is no development in the performance. If the word occurs at WP is entirely dissimilar from the word given in SK, the average confidence score for this WP will be very low. Similarly, if the word occurs at Wp is absolutely identical as the word given in SK then the average confidence score will be very high. The next possibility is that the word occurs at WP is partially similar to the word given in SK. If this is the case, the average confidence score will be in between the above two values.

4) The value of p is incremented by a fixed number of the frames and the testing is done in this new analysis window. This process is repeated until z + p reaches m.

5) Identifying keywords from the confidence score by applying a threshold. The threshold (ts) is computed from the confidence score as follows

$ts = aS_{max}, 0.5 < a < 0.9$

Where S_{max} is the global maximum confidence score and a is the adjustable parameter.

6) After spotting the search keyword in the speech file, the corresponding speech file is retrieved from the database and stored separately.

7) Step 1 to step 6 is repeated for all the audio files present in the database in order to retrieve all the audio files related to the spoken keyword. Then they are arranged in an order based on the number of occurrences of the keyword and clustered together.

4.2. Continuous speech recognition

The proposed continuous speech recognition system comprises of three stages namely pre-processing, segmentation and classification.

4.2.1 Signal pre-processing

It is very critical to pre-process the speech signal in the applications where silence or background noise is completely undesirable.

- Stop Band Filter: A band-stop filter works to screen out frequencies that are within a defined range, providing easy passage only to frequencies outside of that range. It is also called as band elimination, band reject, or notch filters. Placing a low-pass filter in parallel with a high-pass filter can make it as a band-stop filter. The limit of frequencies that a band-stop filter [7] blocks is known as the 'stop band', which is bound by a lesser cut-off frequency and a higher cut-off frequency. The frequency of maximum attenuation in it is called the notch frequency. In order to enhance the performance, the stop band filter has been used in this research work.
- Framing: In most processing tools, it is not appropriate to consider a speech signal as a whole for conducting calculations. A speech signal is often separated into a number of segments called frames. A continuous speech signal has been blocked into N samples, with adjacent frames being separated by M (M < N). In our work, after the Pre-emphasis, filtered samples have been converted into frames, having a frame size of 25 msec. Each frame overlaps by10 msec.
- Windowing: The window w (n), determines the portion of the speech signal that is to be processed by zeroing out the signal outside the region of interest. To reduce the edge effect of each frame segment, windowing is done. Rectangular window has been used in this work.

4.2.2. Speech segmentation

Automatic speech segmentation is a necessary step which is used in speech recognition and synthesis systems. Speech segmentation is breaking continuous streams of sound into some basic units like words, phonemes or syllables that can be recognized. The general idea of segmentation can be described as dividing something continuous into discrete, non-overlapping entities [32]. Segmentation can be also used to distinguish different types of audio signals from large amounts of audio data, often referred to as audio classification [33]. Automatic speech segmentation methods can be classified in many ways, but one very common classification is the division to blind and aided segmentation algorithms. A central difference between aided and blind methods is as to how much the segmentation algorithm uses previously obtained data or external knowledge to process the expected speech. The algorithm for automatic speech segmentation is as follows.

- Short term energy and zero crossing rates are computed for the pre-processed frames.
- Some threshold value which is dynamically generated has been taken and signals having a value less than this threshold value has been changed to zero as signal having syllable will have a data value more than the threshold value.
- Then signal has been checked for value not equal to zero and greater than some particular value and that point will be marked as starting location of the boundary.
- After getting the starting location, the zero values of signal have been checked and if there are suitable numbers of continuous zeros then it has been defined as the end of the boundary. Once an endpoint has been detected, we can precede analysing signal from the endpoint of the first one looking for the starting position of next one.

The strategy is evolved for a speech recognition task, with a view to identify the spoken utterances of specific words in retrieved documents. The vocabulary includes the words present in the speech corpus. The experiments are conducted using the database created, from which eighty percentage samples are used for training purpose and the remaining twenty percentage samples for testing the performance of the methodology. The recognition task is achieved by using the distribution capturing capability [44] of the AANN.

5. Performance Measures

The purpose of this research is to identify keywords within audio and retrieve the documents based on the keyword detection and concerting into the corresponding language text. Unlike ASR, which typically considers the correct recognition of all words equally important, we are interested in the trade-off of precision and recall. We use the following metrics to evaluate the systems presented in this work. The Figure Of Merit (FOM) was originally defined by [40] for the task of Keyword Spotting (KS). By optimizing the FOM [41], [42] accuracy of the Spoken Term Detection (STD) can be increased. It gives the average detection rate over the range [1, 10] false alarms per hour per keyword. The FOM values for individual keywords can be averaged in order to give an overall figure. The NIST STD 2006 evaluation plan [43] defined the metrics occurrence-weighted value (OCC) and Actual Term-Weighted Value (ATWV) and a Maximum Term Weighted (MTWV). These three metrics have been adopted and their description follows.

For a given set of terms and some speech data, let $N_{correct}(t), N_{FA}(t) and N_{true}(t)$ represent the number of correct, false alarm, and actual occurrences of term t respectively. In addition, we denote the number of non-target terms (which gives the number of possibilities for incorrect detection) as $N_{NT}(t)$. We also define miss and false alarm probabilities, $P_{miss}(t)$ and $P_{FA}(t)$ for each term t as:

$$P_{miss}(t) = 1 - \frac{N_{correct}(t)}{N_{true}(t)}, \quad P_{FA}(t) = \frac{N_{FA}(t)}{N_{NT}(t)}$$

In order to tune the metrics to give a desired balance of precision versus recall, a cost CFA for false alarms was defined, along with a value V for correct detections. The occurrence-weighted value is computed by accumulating a value for each correct detection and subtracting a cost for false alarms as follows:

$$OCC = \frac{\sum_{f \in terms} [VN_{correct}(t) - C_{FA} N_{FA}(t)}{\sum_{f \in terms} [VN_{correct}(t)}$$

Whilst OCC gives a good indication of overall system performance, there is an inherent bias towards frequently occurring terms. The second NIST metric, the actual term-weighted value is arrived at by averaging a weighted sum of miss and false alarm probabilities, $P_{miss}(t) and P_{FA}(t)$ over the terms:

$$ATWV = 1 - \frac{\sum_{f \in terms} [P_{miss}(t) + \beta P_{FA} N_{FA}(t)}{\sum_{f \in terms} 1}$$

where $= \frac{c}{y}(P_{prior}(t)^{-1} - 1)$. The NIST evaluation scoring tools sets a uniform prior term probability $P_{prior}(t) = 10^{-4}$ and the ratio $\frac{c}{y}$ to be 0.1 with the effect that there is an emphasis placed on recall compared to the precision in the ratio 10:1.The third term MTWV is over the range of all possible values of threshold. It ranges from 0 to +1.

In this work, we also present the results in terms of FOM and OCC. However, rather than giving the ATWV values which give point estimates of the miss and false alarm probabilities, we present these results graphically in order to show the full range of operating points. For all results, tuning for the parameters using the developed algorithm is performed on STD development set according to the metric which is used in evaluation. For all measures, higher values indicate better performance.

The performance of speech recognition systems is usually specified in terms of accuracy, error rate and speed. Accuracy may be measured in terms of performance accuracy which is usually rated with word error rate (WER), whereas speed is measured with the real time factor.

- Word Error Rate (WER)

Word error rate is a common metric of the performance of a speech recognition or machine translation system. The general difficulty of measuring performance lies in the fact that the recognized word sequence can have a different length from the reference word sequence (supposedly the correct one). The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level. This problem is solved by first aligning the recognized word sequence with the reference (spoken) word sequence using dynamic string alignment. Word error rate can then be computed as

$$WER = \frac{S + D + I}{N}$$

where S is the number of substitutions, D is the number of the deletions, I is the number of the insertions, and N is the number of words in the reference

When reporting the performance of a speech recognition system, sometimes word recognition rate (WRR) is used instead.

$$WRR = 1 - WER$$

6. Experiments and Results

6.1. The databases

The experiments have been conducted over a corpus which is composed of broadcast news and conversations recorded from various channels like BBC, NDTV, Doordharsan News and real time recorded speech. It includes three different source types. one hour of broadcast news (BNEWS), 30 minutes of conversational telephony speech (CTS) and one hour of real time conference room meetings (REALCONFSP). For the experiments, we have processed the query set that includes 150 queries. Each query is a phrase containing between one to five terms, common and rare terms, terms that are in the manual transcripts and those that are not. The dataset is divided into the development corpus and evaluation corpus. The development corpus is utilized for training the structure and fine-tuning the parameters which is composed of two hour speech from the above three source types each including 150 search terms. The evaluation corpus is composed of two and half's hour speech including 100 search terms which is for validation.

6.2. Feature extraction

The first 39 mel frequency cepstral coefficients, other than the zero th value are used to evaluate the proposed algorithm. Cepstral mean subtraction is performed to trim down the channel effects. The preferred properties of the speech signals are a sampling rate of 8 kHz, 16 bit monophonic PCM format. The frame rate is as same as the keyword frames/sec, where each frame is 16 ms in duration with an overlap of 50 percent between adjacent frames.

6.3. Parameter Tuning Phase

In order to adjust the parameters of the algorithm that defer the preeminent performance, several experiments have been performed on development corpus, whose results are provided in this subsection. The parameters to be tuned are: number of epochs (One epoch of training is a single presentation of all the training vectors to the network), adjustable parameter, the number of frame shift and hidden and compression layer present in the AANN structure. The MFCC feature vectors are extracted for all the speech frames as described in Section 3 and Section 6.2. For the given keyword feature vectors, the distribution of the feature vectors is captured using the AANN model as described in Section 4. The feature vectors of Wp are given as input to the AANN model and the average confidence score is measured as described in Section 5. Fig. 3 shows the progression of the confidence score when the number of epochs increases. There is no considerable change present in the confidence score curve even though the number of epochs was increased to 1000. Consequently the AANN models are trained for only 100 epochs. Fig. 4 shows the evolution of confidence score when the frame shift is changed from 1/2, 1/4, 1/8th, and 1/16th. It's evidently shown that the confidence score is better for the frame shift 1/16. So the ANNN models are tested for 1/16th frame shift. The progression of confidence score when the hidden and compression unit of AANN model changed from 39L 59N 10N 59N 39L, 39L 78N 20N 78N 39L and 39L 98N 30N 98N 39L is measured. It renders that there is not a significant change in confidence score by changing the structure of the AANN. So we have taken the structure as 39L 78N 20N 78N 39L of the AANN model.

It is not possible to obtain the same average confidence score for the same keyword query every time. To avoid the false keyword spotting the confidence scores which are greater than the threshold value are considered. Hence, after obtaining the global maxima of the confidence scores for the entire speech signal, the hypothesized keyword is validated by using the threshold. For calculating the threshold, adjustable parameter (a=0. 5) is used in this experiment. We have determined empirically a detection threshold _ per source type and hardest decision of the occurrences having a score less than _ is set to false; false occurrences returned by the system are not considered as retrieved and therefore, are not used for computing ATWV, MTWV, precision and recall. The value of the threshold per source type is reported in Table. 1 It is correlated to the accuracy of the Document retrieval

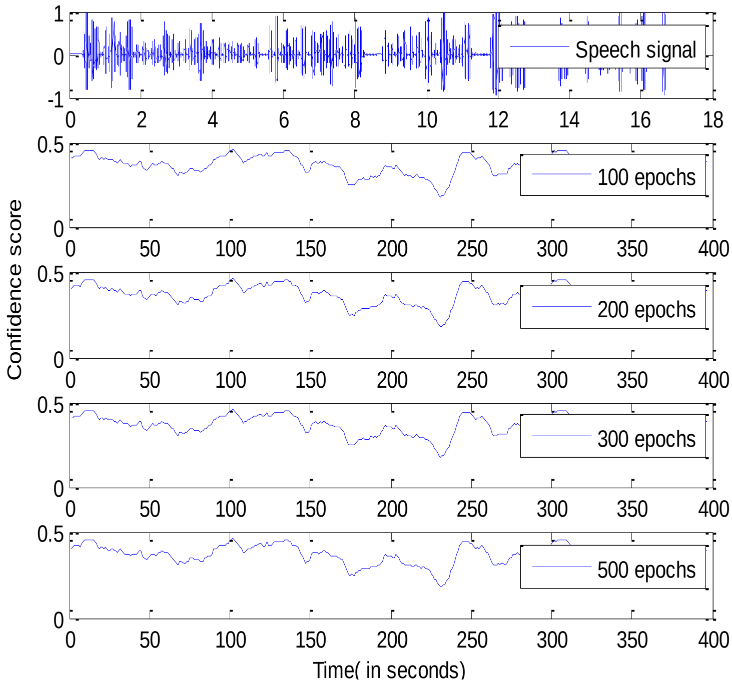

Fig. 3. Effect of epochs on the confidence score.

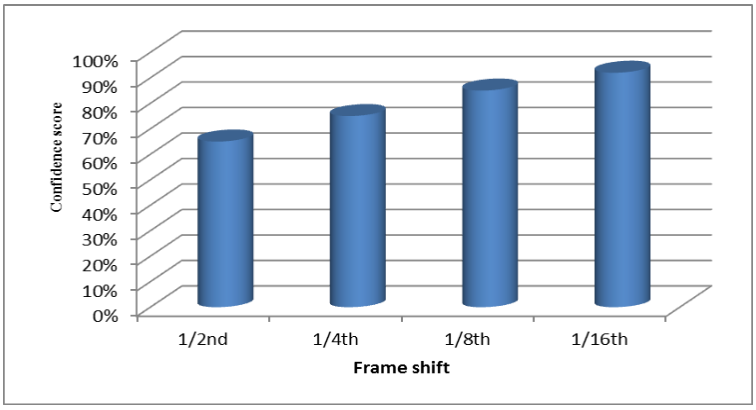

Fig. 4. Effects of frame shift in confidence score.

Table 1. Values of the threshold per source type.

BNEWS	CTS	REALCONSP
0.69	0.74	0.91

7. Evaluation Results

The experiments were conducted on the database described in Section 6.1. A set of 150 search queries were elected depending on their high frequency of occurrence and appropriateness as search terms for spoken document information retrieval, and evaluation (retrieving search terms) is performed on the test set.

7.1. Spoken term detection and keyword spotting results

- Recognition accuracy whilst the recognition accuracy is not the main focus of this work, it is an important factor in STD/KS performance. In Table II, we present the recognition accuracy results after providing the empirical tuned parameter values to the algorithm.

Table 2. Recognition accuracy for the source types.

	BNEWS	CTS	REALCONSP
Recognition accuracy	92.4%	91.6%	89.9%

- Evaluation in terms of FOM and OCC Table III shows that the evaluation in terms of the FOM, BNEWS renders better performance than the source types CTS and REALCONSP. Similarly in accordance with the term OCC, again the BNEWS provides the best performance.

Table 3. Results in terms of FOM and OCC for the source types.

	BNEWS	CTS	REALCONSP
FOM	85.9%	89.3%	81.7%
OCC	89%	87%	86%

7.2. Evaluation in terms of spoken document retrieval

For each found occurrence of the given query, our system outputs: the location of the term in the audio recording (begin time and duration), the score indicating how likely is the occurrence of the query, and a hard decision as to whether the detection is correct. We measure precision and recall by comparing the results obtained over the automatic transcripts (only the results having true hard decision) to the results obtained over the reference manual transcripts.

Table 4. ATWV, MTWV, Precision and recall per source type.

MEASURES	BNEWS	CTS	REALCONSP
ATWV	0.89	0.86	0.84
MTWV	0.90	0.88	0.85
Recall	0.83	0.79	0.76
Precision	0.82	0.78	0.77

Our aim is to evaluate the ability of the suggested retrieval approach to handle transcribed speech data. Thus, the closer the automatic results to the manual results are, the better the search effectiveness over the automatic transcripts. The results returned from the manual transcription for a given query are considered relevant and are expected to be retrieved with the highest scores.

7.3. Evaluation in terms of Continuous speech recognition

The recognition accuracy and the word error rate with the MFCC for proposed speech recognition system is presented in table and its analysis is presented in graph form shown in Table 5 and fig. 5.

Table 5. WRR and WER measure for CSR.

MEASURES	BNEWS	CTS	REALCONSP
ATWV	0.89	0.86	0.84
MTWV	0.90	0.88	0.85
Recall	0.83	0.79	0.76
Precision	0.82	0.78	0.77

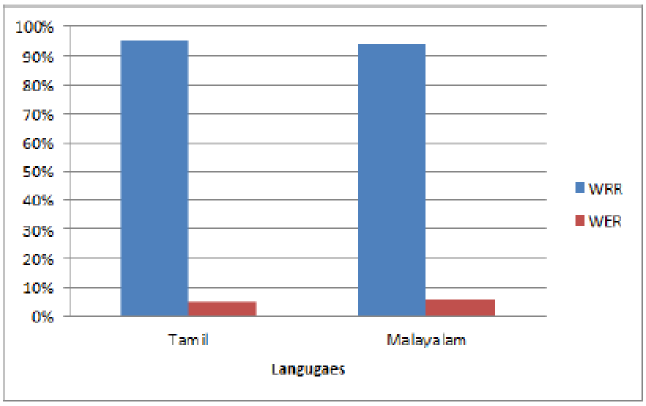

Fig. 5. WRR and WER for Dravidian languages.

8. Conclusion

In this paper, an alternate method for spoken document retrieval system is proposed using an auto associative neural network based keyword spotting and a significant effort has been carried out for recognizing the retrieved documents using acoustic features. The acoustic systems are an interesting compromise between complexity and performance. The distribution acquiring capability of the auto associative neural network uses this proposed work. The proposed method involves sliding a frame-based keyword template alongside the speech signal and by means of confidence score obtained from the normalized squared error of AANN to competently search for a match. This work formulates a new spoken keyword detection algorithm. Based on the spoken term as the key the relevant documents are retrieved. This module studies how the spoken document is retrieved based on the spoken query as the key can be performed efficiently over different data sources. The experiment reveals that all the measure provided the best performance for the source type BNEWS. It yields the overall performance at around 92% of the document retrieval. Then such a substantial effort has been carried out for recognizing the retrieved documents using acoustic features. To achieve this job, the desired feature extraction is done after performing required pre-processing techniques. The most extensively used MFCC is used to extract the substantial feature vectors from the enriched speech signal and they are given as the input to the AANN classifier. The adopted AANN classifier is trained with these input and target vectors. The results with the specified parameters were found to be agreeable considering the less number of training data. The more number of speech data to be trained and tested with this network in future. As the

Dravidian languages are alike in characteristics, designing a lesser amount of intricate system with the best performance is a challenging task. This work is the principal step in this track. Spoken document retrieval system will be in the direction to progress the effectiveness of the algorithm by optimizing the time taken to examine the entire audio file frame by frame and the document retrieval.

References

1. Schauble, P. (1997). Multimedia Information Retrieval Content-Based Information Retrieval from Large Text and Audio Databases. Kluwer Academic Publishers, Norwell, MA, USA.
2. Glavitsch, U.; and Schauble, P. (1992). A system for retrieving speech documents. *In Proc. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 168–176.
3. Garofolo, J.; Auzanne, G.; and E. Voorhees. (2000). The trec spoken document retrieval track: A success story. *In Proc. TREC 8*, 107 – 130.
4. Jansen, A.; and Niyogi, P. (2009). Point process models for spotting keywords in continuous speech. *IEEE Trans. Audio, Speech and Lang. Proc*, 17(8), 1457–1470.
5. Bridle, J.S. (1973). An efficient elastic-template method for detecting given words in running speech. *In Proc. of the Brit. Acoust. Soc. Meeting*.
6. Wilpon, J. G.; Rabiner, L. R.; Lee, C. H.; and Goldman, E. R. (1989). Application of hidden markov models for recognition of a limited set of words in unconstrained speech. *In Proc. of ICASSP*.
7. Silaghi, M. C.; and Bourlard, H. (2000). Iterative posterior-based keyword spotting without filler models. *In Proc. of ICASSP*.
8. Zhang.; Yaodong.; and James R. Glass. (2009). Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 398 – 403.
9. Wilpon, J. G.; Rabiner, L. R.; Lee, C.-H.; and Goldman, E. R. (1990). Automatic recognition of keywords in unconstrained speech using hidden markov models. *IEEE Transactions of Acoustic, Speech, and Signal Processing*, 38(11), 1870–1878.
10. Hofstetter, E. M.; and Rose, R. C. (1992). Techniques for task independent word spotting in continuous speech messages. *In Proc. of ICASSP*.
11. Szoke, I.; and Schwarz, P.; and Matejka, P.; and Burget, L.; and Fapso, M.; and Karafiat, M.; and Cernchy, J. (2005). Comparison of keyword spotting approaches for informal continuous speech. *In Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*.
12. Ed, J. A. (2002). *Topic Detection and Tracking: Event-Based Information Organization* , Norwell: Kluwer Academic Publishers.
13. Xu, W.; Liu, X.; and Gong, Y. (2003). Document clustering based on non-negative matrix factorization. *In Proc. of the 26thInternational ACM SIGIR*, New york and NY and USA, 267 – 273.
14. Junkawitsch, J.; Neubauer, L.; and Ruske, G. (1996). A new keyword spotting algorithm with pre-calculated optimal thresholds. *In Proc. of ICSLP*.

15. Thambiratnam, K.; and Sridharan, S. (2005). Dynamic match phone-lattice searches for very fast and unrestricted vocabulary kws. *In Proc. of ICASSP*.
16. Vimala, V.; and Radha. (2012). Efficient speaker independent isolated speech recognition for tamil language using wavelet denoising and hidden markov model. *In Proceedings of the Fourth International Conference on Signal and Image Processing*.
17. Wellekens, J. (1987). Explicit time correlation in hidden markov models for speech recognition., 384 – 386.
18. George, E.; Dahl.; and Dong Yu.; and Li Deng.; and Alex Acero. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions On Audio, Speech, And Language Processing*, 20(1), 30 - 42.
19. Kapadia, S.; and Valtchev, V.; and Young, S. J. (1993). Mmi training for continuous phoneme recognition on the timit database. *In ICASSP*, 2, 491-494.
20. Juang, B. H., and Chou, W.; and Lee, C. H.. (1997). Minimum classification error rate methods for speech recognition. *IEEE Trans. Speech Audio Process*, 5(3), 257 – 265.
21. Povey, (2003). *Discriminative training for large vocabulary speech recognition*. Ph.D. dissertation. Departmen of Engg. Cambridge University U.K.
22. Li, X.; and Jiang, H.; and Liu, C. (2003). Large margin hmms for speech recognition., 513– 516.
23. Jiang, H.; and Li, X. (2007). Incorporating training errors for large margin hmms under semi-definite programming framework. *In ICASSP*, 4, 629 – 632.
24. Yu.; and Deng, L.; and He, X.; and Acero, A. (2006). Use of incrementally regulated discriminative margins in mce training for speech recognition. *In ICSLP*, 2418– 2421.
25. Hifny, Y.; and Renals, S. (2009). Speech recognition using augmented conditional random fields. *IEEE Trans. Audio, Speech, Lang. Process*, 17(2), 354 – 365.
26. Morris, J., and Fosler-Lussier, E. (2006). Combining phonetic attributes using conditional random fields. *In Interspeech*, pp. 597– 600.
27. Heigold, G. (2012). *A log-linear discriminative modeling framework for speech recognition*. Ph.D. dissertation, Departmen of Engg., Aachen Univ germany.
28. Zweig.; and Nguyen, P.; (2010). A segmental conditional random field toolkit for speech recognition. *In Interspeech*, 2858-2861.
29. Amin Ashouri Saheli.; and Gholam Ali Abdali.; and Amir Abolfazlsuratgar. (2009). Speech recognition from psd using neural network. *In International MultiConference of Engineers and Computer Scientists*.
30. Hasnain, S. K.; and AzamBeg. (2008). A speech recognition system for urdu language. *In International Multi-Topic Conference*, 74– 78.
31. Davis. S. B.; and Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Process*, 28, 357–366.
32. Young, Steve, et al. (2002), *HTK book.*

33. Yegnanarayana, B.; and Kishore, S. P. (2010). AANN: An alternative to GMM for pattern recognition. *IEEE Trans. Neural Netw,* 15, 459–469.
34. Palanivel, S. (2004). *Person authentication using speech, face and visual speech. Ph.D. dissertation*, Department of Computer Science and Engg., Indian Institute of Technology Madras.
35. Kramer, M. A. (1991). Nonlinear principal component analysis using auto associative neural networks. *AIChE*, 37, 233–243.
36. Bourlard, H.; and Kamp, Y. (1988). Auto association by multi-layer rceptrons and singular value decomposition. *Biol. Cybernet*, 59, 291–294.
37. Bianchini, M.; Frasconi, P.; and Gori, M. (1995). Learning in multilayered networks used as autoassociators. *IEEE Trans. Neural Netw*., 6, 512–515.
38. S. Haykin. (1999). *Neural netwoks: A comprehensive foundation*. New Jersey: Prentice-Hall.
39. Kishore, S. P. (2000). *Speaker verification using autoassociative neural networks model*. M. S. thesis. Indian Institute of Technology Madras: Department of Computer Science and Engg.
40. Rohlicek, W.; Russell, S.; Roukos, and Gish, H.; (1989). Continuous hidden markov modeling for speaker independent word spotting. *In Proceedings of ICASSP*, 1, 627 – 630.
41. Wallace, R.; Vogt, R.; Baker, B.; and Sridharan, S. (2010). Optimizing figure of merit for phonetic spoken term detection. *In Proceedings of ICASSP*, 298 – 5301.
42. Wallace, R.; and Vogt, R.; and Baker, B.; and Sridharan, S. (2011). Discriminative optimization of the figure of merit for phonetic spoken term detection. *IEEE Transctions on audio, speech and language processing*, 19(6), 1677 - 1687.
43. NIST. (2006). The spoken term detection (STD) (2006). evaluation plan. Gaithersburg, MD, USA.http://www.nist.gov/speech/tests/std: National Institute of Standards and Technology.
44. Yegnanarayana, B. (1999). *Artificial neural networks*, New Delhi: Prentice-Hall.