IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

A Review of Keyword Spotting as an Audio Mining Technique

B.K. Deka[1*], P. Das[2]

Department of Computer Science & Engineering and Information Technology, Assam Don Bosco University, Guwahati, India

*Corresponding Author: brajendeka@gmail.com, Tel.: +91-90850-72455

Available online at: www.ijcseonline.org

Accepted: 24/Jan/2019, Published: 31/Jan/2019

Abstract— Speech is that the essential and therefore the most profitable ways for correspondence between people. Speech is an emerging technology and automatic speech recognition has created advances in recent years. It provides the flexibility to a machine for responding properly to spoken language. Keyword Spotting could be a very important strategy in audio mining that is employed to recover of all occurrences of a given keyword within the knowledge talked expressions. It has transformed into a fascinating and testing zone as the proportion of an audio substance in the web, telephone and diverse sources growing rapidly. It can be viewed as a subproblem of automatic speech recognition where only partial information has got to be extracted from speech utterances. KWS is closely associated with the task of speech transcription and offers several advantages for certain applications. The main aim of this study is to understand the various approaches used for keyword spotting of speech in order that we can find out the methods that provide better accuracy and performance. Additionally, we have quickly examined the Keyword spotting framework and Audio mining system in this paper.

Keywords—Audio Mining, Keyword Spotting, Automatic Speech Recognition, Audio Indexing.

I. INTRODUCTION

Humans interact with others effortlessly using speech. The speech of an individual differs and varies from each other with regard to some variables like age, gender, mental state, background sound, ascent, articulations and so on [1]. Speech Recognition (SR) is that the inter-disciplinary sub-field of methodologies and technologies that change the popularity and translation of speech into text by computers. It is additionally called automatic speech recognition ASR). It permits a computer to seek out the words that an individual into a microphone or telephone and convert it to written text [2]. The aim of ASR analysis is to permit a computer to acknowledge speech in the time period as human understands [3]. The words that are spoken by somebody freelance of vocabulary size, noise, speaker characteristics, and channel. The accuracy rate of speech recognition depends on the feature extraction techniques in so far as the model used.

The structure of the paper is given below: Section II explains Keyword Spotting System, Section III describes Audio Mining System, Section IV illustrates the Feature Classification Approach, Section V contains the Feature Extraction Method, Section VI contains the Evaluation and Implementation, Section VII contains Literature Review and Section VIII gives the conclusion.

II. KEYWORD SPOTTING SYSTEM

Keyword spotting (KWS) is a branch of audio mining that deals with identifying certain keywords in a long utterance. Keyword spotting is well suited to data mining tasks, process a large amount of speech such as real-time monitoring and to audio document indexing. The task of locating the occurrences of given keywords in a speech utterance is termed as keyword spotting (KWS). It plays a vital role in audio indexing and speech data mining applications [4]. The method is applied mainly for a large number of spoken documents must be searched to learn whether they contain some specific words. The fast detection of the words and information about the exact location eliminate a lot of human work in such tasks like audio data mining [5]. In present context when people are trying to make human-machine interaction seamlessly natural, the scope of this field is even bigger because human conversation contains not only irrelevant words but also non-intentional sounds like a cough, exclamations, and noise. If we can extract only the embedded information, computation can be much efficient and robust. Some examples of these kinds of systems can be voice dialling, voice command systems, audio search engines etc.

A. Challenges

- Spotting the keyword accurately and precisely in the place of the uttered text in a given audio file.

- The background noise is the most significant reason for the multilingual audio mining accuracy, which is high for the clean samples but deteriorates quickly for noisy samples.
- Based on the confidence level score to search the efficient matching of spoken keyword utterance. If the confidence level is high or low, then the measure of discriminative audio mining is considered or discarded.
- To investigate and novel techniques that can be used to improve the performance of keyword spotting in audio mining system for real-world applications.
- To improve keyword detection performance.
- Handling unrestricted out of vocabulary (OOV) words.

B. Applications

The applications of audio mining [6], [7] are:

- Keyword monitoring applications.
- Audio document indexing.
- Command controlled devices.
- Dialogue systems.
- Voice command control.
- Information retrieval.
- Music Information retrieval.
- Human-computer interaction.
- Indexing and looking multimedia system knowledge.
- Monitoring of phone services for the target keyword.
- Person authorization.
- Spoken password verification.
- Security.

III. AUDIO MINING SYSTEM

Audio mining is a speech recognition technique which is used for searching and analyzing audio files for occurrences of spoken words or phrases [8]. It can be used to search specific characteristics of keywords within the huge and heterogeneous audio files. Audio Mining can analyze and search the contents of an audio signal for identifying patterns and associations, retrieving keywords and information, monitoring keywords and indexing. Audio may be within the kind of radio, speech, music and so on, because of the continual, dynamic and non-structured nature of audio, audio files should be depicted with spectral coefficients for more process with data mining techniques. There are variant feature choices to represent speech characteristics in numeric forms.

A. Audio Mining Techniques

- Keyword Spotting System (KWS): This approach aims to detect the occurrences of keywords within the test spoken utterance [6].
- Wake-up-word detection (WUW): It is related to KWS, but it uses speech commands to activate or wake up other systems by an alerting signal [6].
- Spoken term Detection (STD): It is defined by NIST as audio mining which is employed for content-based indexing. Similar to keyword spotting which involves finding occurrences of specific keywords in a speech utterance, STD extends the same by finding a sequence of multiple words in the speech utterance [6].

B. Audio Mining Approaches

Audio mining is a speaker independent, speech recognition technique used to search or recognize audio or video files for occurrences of spoken words or phrases. There are two main approaches to audio mining as:

- Large vocabulary continuous speech recognition (LVCSR): LVCSR audio mining is called text-based indexing which is a two-step process. In the first step, this method converts speech to text and in the second step, it identifies keywords in the generated dictionary that can contain several hundred thousand entries. If the keyword is not present in the dictionary, the system will choose the most similar word it can find. This system is more complex and expensive to carry out [6].
- Phonetic recognition (PR): Phonetic audio mining is also called phoneme based indexing which doesn't convert speech to text but instead works only with sounds. Phonetic audio mining is also a two-step process. In the first step, audio is processed with a phonetic recognizer to generate a phonetic index file. In the second step, the system uses the generated dictionary of phonemes to compare a user's search term to the correct phonetic string [6].

The major differences between the above two approaches are listed in **Table 1** [8].

Table 1. Difference between Phonetic and LVCSR

Phonetic Audio Mining	***LVCSR Audio Mining***
Phonetic System works at the Phoneme level.	LVCSR System works at Words level.
The rate at that the audio content is indexed is persistently quicker than with LVCSR.	The rate at which the audio content can be indexed is slow.
During the search stage, the computational burden is larger for phonetic search systems than for LVCSR technique.	In this case, the search stage is typically simple and less time-consuming.
Phonetic recognition does not need the use of complex language models.	LVCSR approaches must use new language models, which leads to a greater computation load at the indexing stage for LVCSR approaches and results in much slower indexing speeds.
In the phonetic approach, an open vocabulary is maintained which means that searches for personal or company names can be performed without the need to reprocess the audio.	With LVCSR systems, any word that was not known by the system at the time, the speech was indexed can never be found. The LVCSR system has to be updated with a new dictionary that includes all the new words and all the audio has to be pre-processed again which is a time-consuming task.

IV. CLASSIFICATION APPROACH

There are several approaches have been developed over the years. The following approaches are highlighted in this paper:

A. Hidden Markov Model-Based Approaches

HMM is normally used for speech recognition along with keyword spotting. This model is built for both keyword and test utterance. The model apart from the keyword is referred to as a garbage model or filler model. The chance is calculated for each speech utterance to search if it is closer to the keyword. HMM-based keyword spotting suffers from that it requires a large amount of training data that is time-consuming and it requires language ability. Also, the addition of new keywords requires retraining. There were several attempts to beat these issues of HMM-based KWS. However, it once more needs training which might cause a disagreement for audio classification of recent words. These problems related to the HMM, in recent years, advances to the use of DTW based KWS. For text dependent applications, whose phrases or phonemes may be modified using multi-state left to right HMMs. For text independent applications, single state HMMs also known as Gaussian Mixture Models (GMMs) is used. Hence, HMM has been the most successful technique in several speech recognition areas and its application in keyword spotting is also very promising [10], [11], [13], [14], [16], [17].

B. Artificial Neural Network Based Approaches

ANN is a network of artificial neurons, inspired by the principles of biological neural networks. This network can be used in pattern recognition as an effective tool. Neural networks can be implemented either as a generative algorithm or discriminative algorithm. Recently Greedy based algorithm is proposed using neural networks which have shown better results than standard discriminative algorithms. In recent years keyword spotting is properly achieved by an Artificial Neural Network (ANN). In the ANN approach, the test audio is preprocessed to remove noises and then feature extraction is done using the cepstral method. The ANN is trained with the cepstral values to produce a set of last weights. During the testing process, these weights are used to mine the audio files. A Neural Network is constructed by highly interconnected processing units (neurons) which do simple mathematical operations. Neural Networks are characterized by their topologies, weight vectors and active functions which are used in the hidden layers and output layers [3], [9], [31].

V. FEATURE EXTRACTION METHODS

Feature extraction means to find the component of an audio signal that is good for identifying the content and discarding all other matter which carries information like background noise, emotions etc. The various techniques are used for feature extraction like Linear Predictive Coding (LPC), Linear Predictive Cepstral Coefficients (LPCC), Mel-frequency Cepstrum Coefficients (MFCC), Log Energy Filter Coefficient (LEFC), Per-channel energy normalization (PCEN). MFCC is the most popular and commonly used feature extraction technique in most of the application of speech signal. The audio signal may be a non-stationary signal that is its frequency and amplitude response variable with relation to time. MFCC was introduced by Davis and Mermelstein in the 1980s. MFCC can be defined as the short-term power spectrum of the human voice. It has greater significance in speech processing and it approximates the human system response accurately. The main principle of MFCC is filtered bank coefficient. After applying MFCC to the speech signal we get features as in the form of cepstral coefficients [18], [19], [20], [21], [22], [23].

1. Frame the signal.
2. Take FFT of the signal
3. Multiply each FFT size of the corresponding Mel frequency filter value.
4. Take the log of filter bank energies.
5. Take DCT on Mel log energy values (Cepstrum).

The estimated Mel for a given frequency f in Hz is calculated by the following formula:

$$\text{Mel}(f) = 2595 * \log10(1 + (f / 700))$$

The step by step process of MFCC is shown in the following Figure 1.

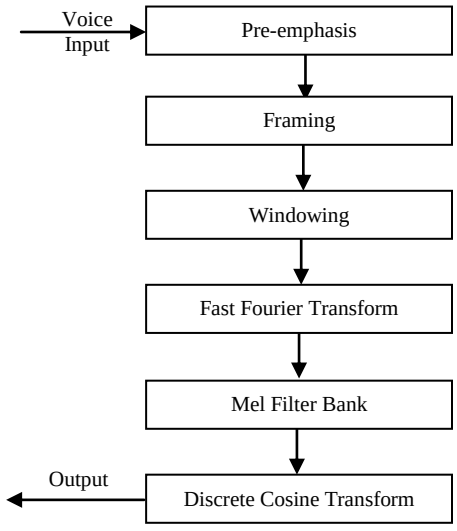

Figure 1. Block diagram of MFCC.

A. Supervised Audio Mining

Supervised Audio mining is the type of learning that takes place when the training instances are labelled with the correct result, which gives feedback about how learning is progressing. In this case, the classes to which the training

samples belong are known beforehand [24]. The different supervised approaches used for spoken term detection in audio mining are as follows [25]:

- Acoustic based keyword spotting.
- LVCSR (Large Vocabulary Continuous Speech Recognition) based.
- Sub-word recognizer based.
- Query-by-Example (text-based STD).
- Event-based.

B. Unsupervised Audio Mining

In unsupervised Audio mining, there is no desired output, so no error signal is generated. It refers to the problem of trying to find a hidden structure in unlabeled data. Here, the input vectors of similar types are grouped together during the training phase [24]. The different unsupervised approaches used for spoken term detection in audio mining are as follows [25]:

- Frame-based template matching.
- Segment based template matching.

VII. EVALUATION AND IMPLEMENTATION

A. Corpora and Tools

Speech corpora/corpus is a database of speech audio files and text transcriptions of these audio files in a format that can be used to create Acoustic models. According to the literature review, no standard dataset of an Assamese language is found of this research. In this system, keywords are Assamese keywords. They are to be trained for a specific speaker. Hence training data used for them is isolated utterances of the keywords. All utterances are essentially from a particular speaker. For each of the ten (10) Assamese words, 50 isolated utterances by 25 male and 25 female speakers in the age group ranging from 20 to 40 years are used for training. So, the total size of the keyword training database is 500. All speech data are digitized into 16-bit samples at a sampling rate of 16 kHz.

The keyword spotting in Audio Mining system will be evaluated for the speech data collected for the experimental purpose. Male and Female voices are recorded using a laptop, microphone in a noisy free environment and the performance measure will be evaluated in terms of FA/KW/H.

B. System Description

Initially, a keyword spotting system will be developed using HMM as a classifier and MFCC as the popular feature extraction technique on the recorded data to improve the accuracy and increased the scalability of automatic recognition and keyword detection system. The performance of the keyword spotting system will be evaluated using standard test database from TIMIT, Switchboard1 and Albayzin to address the OOV (out of vocabulary) words and to spot keywords in unconstrained Assamese speech which is expressed in terms of hit rate, miss rate and minimum false alarms (FAs).

C. System Performance

Initially, we are trying to implement the HMM-based approach and subsequently ANN based approach and then comparing the result with the existing approach to find the better performance rate of the keyword spotting in audio mining system. The performance rate is based on word error ratio (WER) of discriminative methods in different vocabulary size.

VIII. LITERATURE REVIEW

In paper [3], proposed the Malayalam Word Identification for Speech Recognition System which designed for the Malayalam language uses a syllable-based segmentation approach. It maintains a database of utterances which consists of a total of 100 words used in the agriculture domain. Multiple utterances of these words by 9 different speakers with 6 male and 3 females were recorded using an audio tool Audacity with a sampling frequency of 16000Hz. A speech corpus was consecutively developed which consists of these audio files along with the syllable transcription. The experimental results show that syllable based word identification system for Malayalam using HTK on the Linux platform was performed using MFCC feature extraction technique and HMM model. The training was conducted for 40 vocabularies of bi-syllable words. The system was trained to spot the utterances at intervals the training vocabulary with moderate accuracy. Experiments are carried out with real-time utterances of 100 words and obtained 96.4 % accuracy in ANN, which outperformed HMM.

In paper [4], investigates the speech recognition task of spotting predefined keyword in continuous speech has both practical and scientific motivations. Considering the keyword and background point process modelling of a sparse, event-based representation of the speech signal. Keywords can be spotted with accuracy levels comparable to Hidden Markov Model-based keyword spotting systems based on two speech corpora as TIMIT corpora and Boston University Radio News Corpora. The experiment was carried out by using MFCC as a feature extraction technique. The experimental results show that the figure-of-merit performance of keyword spotting in the high-precision regime is better predicted by the median duration of the keyword rather than simply the number of its constituent syllables or phonemes.

In paper [5], paying attention to Keyword Spotting System for Tamil Isolated Words using Multidimensional MFCC and DTW Algorithm. Tamil month names are recorded in a noisy environment by a female speaker. Each month name in Tamil has been repeated 25 times so that totally 300 words are

recorded using AUDACITY speech software. A pre-processing is created for not only noise reduction but also normalization. Then 12D, 26D, and 39D MFCCs are calculated for all utterances and are stored as templates. Finally, DTW is employed as a pattern matching rule in consequence of its speed and effectiveness in detecting similar patterns. The experimental results show that the number of occurrences of given keywords is identified and a maximum of 89.7% average accuracy of keyword spotting is obtained with 39-dimensional MFCC. The system is designed for Tamil language but all its modules except the spoken word references developed are language independent.
In paper [6], presents a satisfactory audio mining solution for various tasks like spoken document indexing and retrieval. KWS is a technologically relevant problem in audio mining to automatically detect keywords of interest in the spoken document and has been utilized in a broad variety of applications. Various speech files like MIT corpus and TIMIT corpus are used in the KWS approaches. There are variant feature extractions types like MFCC, LPC and VQ are used to find the speech features. The various approaches have been used in keyword spottings like DTW, HMM, Neural Network, VQ, SVM, and Hybrid methods.

In paper [9], presents a keyword spotting method based on BLSTM neural network and CTC token passing algorithm. The pre-processing section performed by the neural network maps every position of the associated input sequence to a vector, indicating the likelihood of every character presumably being written at that position. The CTC rule generates a token for each character and each position within the text line that stores the likelihood of that character being present at that position along with the likelihood of the simplest path from the start to that position. It was absolutely tested on 3 datasets of English written documents and gave promising results.

In paper [10], describes the approach in the development of Urdu keyword spotting system (KWS) which is based on filler models to account for non-keywords speech intervals. Here, 512 points FFT has used a feature extraction technique. Also, Sliding Model Method has been used to develop KWS which is implemented with Distance time Warping (DTW) and HMM as a filler model for developing the system. KWS has been developed on 5 keywords of Urdu using word boundary detection. Training and testing dataset consists of isolated words of 7500 and 3200 utterances respectively. The experimental result shows that the accuracy of the system has been found to be 98.1%.

In paper [11], highlights to measure the benefits of large training databases for non-English HMM-based keyword spotting based on experiments. Training and evaluation speech were taken from the Switchboard1 English telephone speech corpus, the Call-home Spanish telephone speech corpus, and the OGI Multilingual Indonesian telephone speech corpus. For each language, all utterances containing out-of-vocabulary words were removed. From the experiment, it is found that some gains in performance can be obtained through increased training database size, the magnitude of these gains may not necessarily justify the effort and incurred delay of constructing such databases. This has produced for the immediate development and deployment of non-English keyword spotting systems which can feasibly be developed with small training databases and still achieve performances close to that of a system trained using a very large database.
In paper [12], proposed a method to realize a speech-to-gesture conversion for communication between normal and speech impaired people. The keyword spotting was employed to recognize the keywords from input speech signals. Keywords are recognized from a speech by victimization the Hidden Markov Model (HMM) supported keywords recognizing and so the speech-to-gesture conversion was finally completed by enjoying the corresponding 3D gestures with OpenGL from the results of keyword recognizing. 13-Dimensional MFCC can be used as a feature extraction technique and Mean Opinion Score (MOS) method is used in the evaluation of synthesized speech quality to evaluate the accuracy of speech to gesture conversion. In the experiment, a total of 592 sentences of speech (mono-channel, 16-bit quantization precision, 16 kHz sampling rate, saving as the Microsoft Windows WAV format) are recorded under the office environment by the eight speakers including four women and four men. These sentences include numbers, alphabets, and common words. We randomly selected 481 sentences as the training speech and 111 sentences as the testing speech. The experimental result shows that the realized keyword spotting achieves 90.1% of average recognition rate on letters and numbers.

In [13], a fusion of spectral and cepstral features was used for a keyword spotting system. Bark scale based mostly energy and MFCC are used severally and together with acceptable weights for characteristic keywords in spoken utterances. DTW algorithm was used as a template matching algorithm to determine the similarity between reference keywords and unknown utterances. The performance of DTW shows an accuracy of higher than 81% for various speakers whereas the fusion of the 2 feature sets raises the score to over 86%, each supported a little set of utterances from the Call Home database. The result with the combination of those features was extremely comparable the individual results.

In paper [14] describes understanding the various approaches used for the recognition and detection of speech so that we can find out the methods that give better accuracy and performance. Here, MFCC, LEFC are widely used for feature extraction technique and HMM gives better performance compared to other existing models.

In the paper [18], presents a report on the development of a speaker independent, continuous transcription system for Malayalam. The system utilizes the Hidden Markov Model (HMM) for acoustic modelling and Mel Frequency Cepstral Coefficient (MFCC) for feature extraction. It is trained with 10 male and 11 female speakers in the age group ranging from 20 to 40 years. For testing the system, five unknown speakers are selected and speech data of five different sentences collected from them. For training, the speakers are requested to read the 20 sentences. During the training phase the HMM trainer creates acoustic models from three components; feature vectors, language models and dictionary. Word models that are engineered from language models and pronunciation wordbook are used for pattern matching within the testing part together with the syntax and linguistics of the language. The simulation results show that the transcription system for Malayalam achieved while working with medium size vocabulary. The system obtained a word recognition accuracy of 87.4% and a sentence recognition accuracy of 84% once tested with a group of continuous speech knowledge. The accuracy results are highly encouraging while considering the training data and vocabulary size.

In paper [19] demonstrates a speaker independent isolated speech recognition system for the Tamil language. The most versatile and thriving approach to speech recognition thus far has been a Hidden Markov Model that is enforced during this analysis work. MFCC is used for the extraction of the features from the speech waveform. The speech corpus containing 50 utterances of isolated speech collected from 10 females is used for training. The database consists of 5 repetitions of every word produced by each speaker. The experimental result displays high-quality word accuracy of 88% for trained and test utterances spoken by the speakers. The performance evaluation of the system gives 0.88 Word Error Rate (WER).

In [20], Multimedia Keyword Spotting (MKWS) was proposed where a keyword spotter for searching a spoken keyword in a multimedia file. Both HMM and DTW techniques are employed in this approach.HMM is employed to represent the spoken words in phoneme illustration. DTW is used for measuring similarities between the keyword and spoken utterance. MFCC is used for extracting audio features from the speech signal.

In paper [21], Bengali KWS has been developed on 12 keywords using filler modelling approach. Training dataset consists of 350 utterances of keywords and a subset of TIMIT English speech corpus has been used to develop a filler model. Test dataset consists of 240 speech utterances. The speaker dependent keyword spotter is based on Hidden Markov Model (HMM) where the keywords will be having whole word models trained as isolated words. The HMMs will be trained using both single Gaussian and then a GMM (Gaussian Mixture Density) based approach will be taken. MFCC is used for feature extraction technique to extract the feature from a speech in this project. The overall accuracy of the system has been found to be 95.83%.

In paper [25], proposed a sequence discriminative training framework for both fixed vocabulary and unrestricted acoustic KWS. By presenting word-autonomous telephone cross sections or non-keyword clear images to develop contending theories, viable and proficient succession discriminative preparing approaches are proposed for acoustic KWS. The experiments show that the proposed approaches are obtained consistent and significant improvement in both fixed vocabulary and unrestricted KWS tasks compared to previous frame-level deep learning based acoustic KWS methods.

In paper [26], proposed a keyword spotting approach for content-based video indexing and retrieval system for classifying and indexing broadcast news videos in Malayalam. Here auditory modality is used to identify keywords to analyze the content of the news videos. The audio signals are segmented into 10ms frames and applied the filter bank for extraction of eight- dimensional Log Energy Filter Coefficient (LEFC) feature. Dynamic Time Warping (DTW) of the template and the test audio data feature vector is performed in feature matching stage. The closeness separations measures are processed and it is standardized by the total of vitality in every layout. Best matches from all layouts are put away as per the separation measure and lastly, the ideal keyword will be spotted dependent on this separation measure. The experimental result exhibits that template based keyword spotting using the LEFC feature can be effectively used for the speaker dependent (76%) keyword spotting while it will be less effective with the speaker independent (61%) keyword spotting.

In paper [27], presents a sensitive keyword spotting system applied to voice alarm occasions. By sensitive keyword recovery, crisis cautioning frameworks can consequently screen and catch up the circumstance, discover the emergency happens, and after that give the alert. In this system, recorded 15 sentences in mandarin and simulated a set of conditions in dangerous surroundings in a studio environment (10 female and 10 male speakers). Each sentence is uttered 20 times and the test speech is uttered by the same speakers. The total number of the occurrences of the keyword is 1200. All speech data are digitized into 16-bit samples at a sampling rate of 16 kHz. The set of 26 Mel-frequency cepstral coefficients (MFCC) is used for feature extraction technique. Both sliding windows and Hidden Markov Model by HTK tools techniques are used in this approach. The experimental result shows that the mean

recognition rate of all keywords reached 94%. The recognition speed of the system is relatively high. Hence the proposed approach has rather good performance.

In paper [28], discuss the class of plosive sounds can be considered for detection and classification in continuous speech for keyword spotting based on acoustic matching. In this work, two databases were constructed, one each for training and testing the classification system. The training database consists of 5 male and 5 female speakers of standard Marathi were selected for the study. It shows the way to a data set of 80 x 10 x 2 = 1600 tokens (or 400 per stop consonant), recorded at a sampling rate of 16 kHz in silent condition. The testing database comprised of 16 Marathi sentences spoken by each of 1 male and 1 female speaker recorded at a sampling rate of 16 kHz. The sentences contained a high concentration of plosives especially UVA stops. In all, test data had 219 plosives. The results show that the Acoustic-phonetic features extracted in the vicinity of landmarks or speech events are shown to be reliable for the detection of unvoiced stops with high temporal accuracy.

In this paper [29], presents a perceptual approach which is based on the detection of some particularities of a keyword, the system produces a hit when a keyword is detected and returns the boundaries of the word. This alarm is confirmed by comparing the sequence inside the boundaries to Hidden Markov Models (HMM). International Association of phonetics (IPA) can be considered as the dataset which includes some narrative audio files from several sources with multiple speakers. This approach will improve keyword spotting systems, particularly, in terms of time since Out-Of-Vocabulary (OOV) words are not considered.

In paper [30], discuss a new scoring algorithm has been implemented for spotting keywords using large-vocabulary continuous speech recognition (LVCSR) system. This technique uses the N-best answers and their Viterbi alignments to compute the probability that each particular keyword is present in an utterance. The score for a putative hit is calculated by summing the probabilities for all hypotheses that contain the keyword and isolating by the aggregate of the considerable number of probabilities for all the hypotheses in the N-best list. The test set conversational speech from Switchboard Credit Card conversations contains 10 male and 7 female speakers, for a total of 1,928 utterances (1,065 male, 863 female). The experimental result shows that using a test set of conversational speech from Switchboard Credit Card conversations, to achieve an 81% figure of merit (FOM). Also, word recognition error rate on this same test set is 54.7%.

In paper [31], focus on implementing an artificial neural network (ANN) approach for audio data mining. An obtained audio is pre-processed to remove noise followed by feature extraction using the cepstral method as Fast Fourier Transform (FFT) and Linear Predictor Coefficients (LPC). The ANN model is trained with the cepstral values to produce a set of final weights. During the testing process, these weights are used to mine the audio file. The result shows that the ANN is able to produce only about 90% accuracy of mining due to less correlation of audio data.

In paper [32], presents an unsupervised learning framework to address the problem of detecting spoken keywords. Without any explain corpus, a completely unsupervised GMM learning framework is introduced to generate Gaussian posterior grams for keyword samples and test utterances. A modified segmental DTW is used to compare the Gaussian posterior grams between keyword samples and test utterances. After collecting the distort paths from the comparison of every pair of the keyword sample and the test utterance, used a voting based score merging strategy to give a relevance score to every test utterance for each keyword. The identification result is controlled by positioning all the test expressions as for their important scores. The experiment was conducted on both TIMIT and MIT speech corpus. A standard training set containing 57,351 utterances and a test set with 7,375 utterances used for these experiments. The vocabulary size of both the training and the test set is 27,431 words. The experimental result demonstrates that the feasibility and efficiency of unsupervised learning framework on the keyword spotting task which is encouraging and somewhat comparable to other methods that require more supervised training.

In paper [33] formulates the problem of keyword spotting as non-uniform error automatic speech recognition (ASR) problem and propose a model training methodology based on the non-uniform minimum classification error (MCE) approach. The main idea is to adapt the fundamental MCE criteria to reflect the cost-sensitive notion in that error on keywords is much more significant than errors on non-keywords in an automatic speech recognition task. Two challenging large-scale spontaneous conversational telephone speech (CTS) datasets in two different languages (English and Mandarin) are used in the proposed framework. For this experiment, Dynamic time warping (DTW), Hidden Markov models (HMMs) are widely used modelling technique for speech recognition. The experimental results show our framework can achieve consistent and significant spotting performance gains over both the maximum likelihood estimation (MLE) baseline and conventional discriminatively-trained systems with uniform error cost.

In paper [34] presents a new keyword spotting algorithm that allows keyword detection without representation of the non-keyword parts of an utterance by silence, garbage or filler models. Each point in time another Viterbi way is permitted to begin. So as to have the capacity to look at ways of

changed length, explicit standardized scores are utilized inside the hunt. Another methodology is presented which permits the count of word-explicit choice limits ahead of time. Starting with score distributions of phonemes, the Conditional probability density functions (pdf's) of keywords can be calculated and then applying different strategies decision thresholds can be fixed. Utilizing just equivalent length keywords, word-explicit choice edges don't turn out to be extremely viable. In this work, Multicom 94.4 corpus is used which contains 614 utterances of 2.13 hours duration. When using these 25 keywords, in Multicom 94.4 there are a total number of 941 keyword occurrences. The experimental result shows that the tests with spontaneous speech databases yielded 73.9% Figure-Of-Merit when using context-dependent HMMs. The detection rate at 10 fa/kW/h comes to 80%.

In paper [35], proposed a new method Cohort Word-level Verification (CVW) attempts to increase Isolated Word Verification (IWV) performance by incorporating higher-level linguistic and word level information into the selection of non-keyword models for verification. In this work, MFCC and Cepstral Mean Subtraction (CMS) are used for feature extraction technique. Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) were used to model the keywords and cohort words and the speech background model for the baseline verifier. TIMIT database was used for the evaluation. The first set Acoustically Dissimilar (AD) consists of 54 keywords and 180 non-keywords randomly selected from the TIMIT test set. All keywords used were short (6 phone length). The second set Acoustically Similar (AS) consists of 37 keywords randomly selected from the TIMIT test set. Keyword spotting was performed for the 37 keywords on the TIMIT test set. From the keyword spotter result set, 286 incorrect keyword spot results were selected as acoustically similar in keywords. The experimental result shows that when CVW is combined with the background model based IWV; this method yields a significant decrease in false alarm rate for a difficult test set containing non-keywords that are acoustically similar to the keywords.

In paper [36] introduces a keyword spotting approach to perform audio searching of uttered words in Arabic speech. For the dataset preparation, categorize the two tracks which are as follows: In the first track, more than 300 utterances were used as reference patterns. In the second track, 50 utterances were used as reference patterns. Keyword spotting is performed for the extracted reference to the corresponding audio file. The accuracy of the spotted utterances achieved 97%. The experiments showed that the use of the combined text and audio search has reduced the search time by 90% when compared with audio search only tested on the same content. The results are promising. The accuracy of the spot was around 84% in case of preaches and 88% in the case of the news.

In paper [37], investigates one of the large margin based keyword spotting approaches that use a discriminative method for training the keyword spotter and then evaluate the robustness of this approach in different noisy conditions. The performance of this method is comparing with an HMM-based keyword spotter which uses a generative training method in the same noisy conditions. The training portion of TIMIT datasets is divided into two disjoint parts containing 500, and 3196 utterances. The experimental results show that the large-margin based keyword spotter is more robust than the HMM-based system in noisy environments.

In paper [38] proposed a novel recognition algorithm for spotting keywords in a spontaneously spoken dialogue. The proposed algorithm derives the potential abilities of phoneme discrimination in LVCSR by relaxing the LM constraint and can accurately detect keywords in spoken dialogue using DTW between distinctive feature vectors with less computation time. Here, Spoken dialogue corpus "ETL Map Guidance Task" collected using a WOZ system, consisting of 100 utterances spoken by 14 unknown male and female speakers. MFCC is used as a feature extraction technique to extract the features in speech. The experimental result shows that the 20k Language Model of a general-purpose Large Vocabulary Continuous Speech Recognition (LVCSR) outperforms the Japanese sub-word Language Models and that Distinctive Features (DF) gives more robust performance than Confusion Matrix (CM).

In Paper [39] introduced a new robust frontend called per-channel energy normalization (PCEN), where the key idea is to replace the static log (or root) compression by automatic gain control (AGC) based dynamic compression. PCEN is conceptually simple, computationally cheap, and easy to execute. It significantly outperforms the widely used log-Mel frontend in noisy and far-field conditions. We have also formulated PCEN as neural network layers. This formulation does not just enable us to perform start to finish preparing yet, in addition, to sum up, PCEN to have frequency or time-frequency dependent parameters. The resulting model provides significant further improvements without increasing inference-time complexity. To finish up, this work represents an effort to embed signal processing components into a general-purpose neural network model, where the signal processing components can be seen as basic regularizations. The experimental result represents an effort to embed signal processing components into a general-purpose neural network model where the signal processing components can be viewed as structural regularizations.

In paper [40] explores the usefulness and transferability of deep features applied in the context of the problem of keyword spotting. Use a state-of-the-art deep convolution network to extract deep features. The optimal parameters are the impact of the choice of the hidden layer, the impact of

applying dimensionality reduction with a manifold learning technique as well as the choice of dissimilarity measure used to retrieve relevant word images. With extensive numerical results show that deep features lead to state-of-the-art KWS performance, even when the test and training set to come from different document collections.

In paper [41] describes the key concepts of a word spotting system for Russian based on large vocabulary continuous speech recognition. Additionally concentrating on the acoustic and dialect models, attempting probabilistic structures for grapheme-to-phoneme change, information-driven translation variety, presenting commotion pay and delay recognition into the front-end and at making explicit certainty measures to limit false alerts which are caused by regular words in the dialect display. The framework depends on CMU Sphinx open-source speech recognition stage and on the semantic models and calculations created by Speech Drive LLC. To test a 10-hour database including the recordings of dialogues of around 50 speakers has been recorded, 1183 different keywords are searched within the database. The experimental result shows that the high-quality automated keyword spotting system based on large vocabulary continuous speech recognition for online speech data analysis can be used both as an innovative stage to make successfully coordinated frameworks for checking and as a prepared to-utilize answer for screen worldwide data space.
In paper [42] proposes to recognize the one most dominant emotion in a short story. The Hybrid method is used which is a combination of keyword spotting technique and learning-based method. On the other hand, the learning-based method used three algorithms are Multinomial Logistic Regression, Support Vector Machine (SVM) and Naive Bayes Multinomial which are used instance sentences in short story. The total number of short stories for training data was 8 short stories. Meanwhile, the total number of short stories for data test was 30 short stories. The survey had been done to three persons to label the one most dominant emotion in each short story. The experimental result shows that the Hybrid method implemented a voting rule to determine the one most dominant emotion. The accuracy of the hybrid method resulted in this research was 65.71%.

In paper [43] presents a system for keyword detection in spontaneous speech using keywords predefined by a set of acoustic examples. Here to remove the phoneme models included in the keyword model from the filter models. In order to reduce the false alarms caused by keyword searching step, dynamic time warping (DTW) based template matching, Gaussian Mixture Models (GMM) and Hidden Markov Model (HMM) are used. Also, Mel Frequency Cepstral Coefficient (MFCC) is used for extracting the features. The keyword detection experiments demonstrate the effectiveness of the proposed methods by yielding improved detection performance compared to the baseline system.

In [44], Spanish KWS on 80 keywords has been developed using Albazyin database. Confidence Measure technique is applied to reduce the false alarms for keywords. Filler model can be modelled on a word level or phoneme level. Different filler models result in a different hit and false alarm rate. The hit and false alarm rate of the system have been found to be 84.33% and 41.44% respectively.

In paper [45], presents a novel approach to speech processing based on the principle of pattern discovery which can be used to automatically acquire lexical entities such as words and short multiword phrases directly from an untranscribed audio stream. An approach to unsupervised word acquisition utilizes a segmental variant of a widely used dynamic programming technique, which allows us to find matching acoustic patterns between spoken utterances. On a corpus of educational lecture material, clusters found using the computational technique for extracting words and linguistic entities from speech without supervision that exhibits high purity and plenty of the corresponding lexical identities are relevant to the underlying audio stream.

In paper [46] presented a new deep neural network based framework for keyword spotting which is trained to directly predict the keyword(s) or sub-word units of the keyword(s) followed by a posterior handling methodology manufacturing a final confidence score. This approach is fulfilling the need for keyword spotting with a little memory footprint, low machine value, and high exactitude. Experimental results show that the planned framework outperforms the quality of HMM-based system on each clean and noisy condition. Keyword recognition results reach 45% relative improvement with regard to a competitive Hidden Markov Model-based system, whereas performance within the presence of babble noise shows 39% relative improvement.

In paper [47] proposed a method for finding keywords in an audio database using a spoken query. The method is based on performing a joint alignment between a phone lattice generated from a spoken utterance query and a second phone lattice representing a long utterance needing to be searched. The Joint alignment procedure is implementing in a graphical models framework. TIMEOUT as well as on the Switchboard conversational telephone speech (CTS) corpus can be considered for evaluating the system. The experimental results show that a phone lattice representation of the spoken query achieves higher performance than using only the 1-best phone sequence representation. In [45], Spanish KWS on 80 keywords has been developed using Albazyin database. Confidence Measure technique is applied to reduce the false alarms for keywords. Filler model can be modelled on a word level or phoneme level. Different filler models result in a different hit and false alarm rate. The hit

and false alarm rate of the system have been found to be 84.33% and 41.44% respectively.

In paper [48], starts to build a system of voice recognition using backpropagation algorithm in neural networks by comparing the voice signal of the speaker with recorded voice signals in the database and extracting the main features of the voice signal using MFCC. Here a database has been created for a set of voices from 40 persons by taking 5 samples of each individual person, each of them recording a sample by pronouncing his name 5 times in the system training phase. This system has been tested for about 20 persons by taking a voice sample for each of them and makes the matching process to verify the identity of the persons. The DTW is playing a significant role in influencing in the identification accuracy. The backpropagation algorithm has been tested for the voices recognition and has achieved a success rate of about 95%.

In paper [49], proposed an approach to recognize spoken English words corresponding to digits zero to nine in an isolated way by different male and female speakers. The endpoint detection, framing, normalization, Mel Frequency Cepstral Coefficient (MFCC) and DTW rule are used to process speech samples to accomplish the popularity. The results show that the algorithm managed to recognize almost 90.5% of the English digits for all recorded words.

In paper [50] presented a low-resource approach and results for Wake-up-word (WUW) spotting approach detecting only one personalized keyword in a continuous audio signal which is based on template matching using dynamic time warping and other measures. The recognition of the WUW is performed by a combination of distance measures supported an easy background level classification. For evaluation, a database is recorded with three different background noise levels, four different speaker distances to the microphone and ten different speakers. It consists of 480 keywords embedded in continuous audio information. The overall performance of this challenging database is a recall of 59.6% and a precision of 99.7%.

In [51], proposed methods for using the development of a practical keyword spotting system. The system was designed for Czech language but all its modules, except for the acoustic model trained on Czech phonemes are language independent. Here focused mainly on the optimization of a speed of the system because, in applications like telephone call monitoring for state security services, short processing time is one of the main requirements. The experimental result shows that the performance is evaluated on recordings of Czech spontaneous telephone speech using rather large and complex keyword lists.

In paper [52] discusses the different available word spotting techniques for document images. Word-based approaches adopt touching characters easily and analyze the shapes of the words without explicit character recognition. An inventive approach to search digital document images and retrieve relevant results is using keyword spotting or simply word spotting which gives better results as well as less complex than OCR.

In paper [53] presents line-based keyword spotting based on Hidden Markov Model (HMM) which simulates the keywords in model space as a sequence of character models and uses filler models for background or non-keyword text. The use of filler models improves the retrieval result as non-relevant words are handled appropriately by reducing their value from the overall value. The technique is verified on written documents of English, Arabic, and Devanagari.

In paper [54] discusses two methods for keyword spotting in printed Sanskrit documents, one which is recognized based and other which is recognition free. The first approach is script specific that uses a Devanagari OCR based mostly Block Adjacency Graph (BAG) scheme for word recognition. It includes a BAG based mostly technique that uses a graph to keep up the general character structure. The second approach may be a moment based mostly word matching technique that maintains a script invariant illustration of all word pictures. Word matching is performed exploitation the circular function similarity. A significance feedback technique is employed to refine the word spotting results.

In paper [55], highlights two new methods namely Degree of Match (DM) and Degree of Stability (DS) to significantly reduce the false alarm rate of a keyword spotting system based on the keyword/garbage model. The KWS system was evaluated on a dataset which was recorded from 20 speakers (10 male and 10 female). The evaluating dataset contains 2000 utterances, where 1000 of them consist of 10 keywords and the other 1000 utterances are non-keywords. The evaluating data set further divided into a development set (10%) and a testing set (90%). The proposed methods are based on the observed uncertainty during decoding when non-keywords are present to the system. The experiments show that these 2 ways will significantly cut back the false alarm rate from 75.05% to 5.71% and presently false reject rate will increase from 1.04% to 5.71%.

In paper [56], a methodology for fusing spectral and prosodic information using combined error optimization is proposed. The projected fusion algorithmic program is evaluated on 2 tasks, keyword recognition, and keyword spotting. The integrated feature set is employed in an exceedingly Hidden Markov Modelling (HMM) framework together with an original sliding syllable protocol is employed for looking out

the keyword within the keyword spotting task. Search retrieval efficiency using the proposed method indicates reasonable improvements over conventional methods. In addition to this work, a late fusion of prosodic information at the model level by integrating the prosodic models will be investigated. Keyword spotting experiments are conducted on the Hindi language database developed for this purpose. Experiments on keyword recognition and keyword spotting are conducted to evaluate the performance of the projected fusion methodology. The experimental results obtained in terms of Word Error Rate (WER) and receiver operating characteristics indicate a reasonable improvement over the use of a single feature set like MFCC.

In this paper [57], Proposed a BLSTM based system that performs recognition scheme for the Indian script of Devanagari at the word level. The experimental results show that more than 20% improvement in word accuracy and over 9% reduction in character rate while comparing the traditional OCR system.

In paper [58], described a new technique to the reduction of false alarms in one order of magnitude without any decrease of recognition accuracy. An experimental analysis of the keyword spotting system shows that approx. 82 % of all words are literally detected at the same time with another false alarm of properly detected keyword.

In paper [59], introduced four acoustic confidence measures which are derived from the output of a hybrid HMM/ANN large vocabulary continuous speech recognition system. These confidence measures supported native posterior chance estimates computed by an ANN are evaluated at each phone and word levels using the North American Business News corpus.

In paper [60], described the modifications mode to a connected word speech recognition algorithm based on Hidden Markov models (HMM's) which allow it to recognize words from a predefined vocabulary list spoken in an unconstrained fashion, The novelty of this approach is that creation of statistical models of both the actual vocabulary words and the extraneous speech and background. An HMM-based connected word recognition system is then accustomed realize the simplest sequence of background, extraneous speech and vocabulary word models for matching the particular input. Word recognition accuracy of 99.3% on purely isolated speed and 95.1% when the vocabulary word was embedded in unconstraint extraneous speech, were obtained for the five-word vocabulary using the proposed recognition algorithm.

In paper [61], describes a Hidden Markov Model-based utterance verification system using the framework of statistical hypothesis testing. The two major problems on the way to design keyword and string rating criteria are self-addressed. For keyword verification, completely different various hypotheses are planned, supported the lots of opposing keyword models and a general acoustic filler model. For string verification, different measures are proposed with the objective of detecting non-vocabulary word strings and possibly erroneous strings. This paper motivates the requirement for discriminative hypothesis testing in verification. Once the planned verification technique was integrated into a progressive connected digit recognition system, the string error rate for valid digit strings was found to decrease by 57% once setting the rejection rate to 5%. Moreover, the system was able to appropriately reject over 99.9% of non-vocabulary word strings.

IX. CONCLUSIONS AND FUTURE SCOPE

The review of Keyword Spotting as an Audio Mining technique in speech processing for different spoken language has been made and all the papers give different innovative ideas. Major works in the recognition and detection of speech carried out in Hidden Markov Model (HMM), this gives better performance compared to other existing models. The HMM approaches need a huge quantity of supervised training information. Due to this reason, Dynamic Time Wrapping (DTW) is mostly used in Keyword spotting system. Also, the Mel Frequency Cepstral Coefficients (MFCC) can be used as the better choice for the feature extraction process. In future work, we are trying to develop a Speaker Independent Keyword Spotting System in both Assamese and English language using MFCC and Artificial Neural Network classification tool in the speech processing system.

REFERENCES

[1] G. Hemakumar, P. Punitha, "*Speech Recognition Technology: A survey on Indian Languages*", International Journal of Information Science and Intelligent System, Vol. **2**, No.**4**, **2013**.

[2] A. Katyal, A. Kaur, J. Gill, "*Automatic Speech Recognition : A Review*", IJCST, Vol. **4**, No.**3**, pp.**71–74**, **2014**.

[3] M.M. Kumar, E. Sherly, W.S. Varghese, "*Isolated Word Recognition System for Malayalam using Machine Learning*", In the Proceedings of the 12th International Conference on Natural Language Processing, **2015**.

[4] A. Jansen, P. Niyogi, "*Point process models for spotting keywords in continuous speech*", IEEE Transactions on Audio, Speech and Language Processing, Vol. **17**, No.**8**, pp.**1457–1470**, **2009**.

[5] K. A. Senthildevi, E. Chandra, "*Keyword Spotting System for Tamil Isolated Words using Multidimensional MFCC and DTW Algorithm*", IEEE International Conference on Communication and Signal Processing (ICCSP 2015), pp. **550–554**, **2015**.

[6] E. Chandra, K.A. Senthildevi, "*Keyword Spotting: An Audio Mining Technique in Speech Processing – A Survey*", IOSR Journal of VLSI and Signal Processing, Vol. **5**, No.**4**, Ver. **II**, pp.**22–27**, **2015**.

[7] A.J.K. Thambiratnam, "*Acoustic Keyword Spotting in speech with applications to data mining*", PhD Thesis Published in Queensland University of Technology, **2005.**

[8] M.K. Mand, D. Nagpal, "*An Analytical Approach for Mining Audio Signals*", International Journal of Advanced Research in Computer and Communication Engineering, Vol. **2**, No.**9**, pp.**3645–3647**, **2013**.

[9] V. Franken, A. Fischer, R. Manmatha, H. Bunke, "*A novel word spotting method based on recurrent neural networks*", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. **34**, No.**2**, pp. **211–224**, **2012**.

[10] S. Irtza, K. Rehman, S. Hussain, "*Urdu Keyword Spotting System using HMM*", Conference on Language and Technology, Karachi, Pakistan, **2014**.

[11] K. Thambiratnam, T. Martin and S. Sridharan, "*A study on the effects of limited training data for English, Spanish and Indonesian keyword spotting*", Proceedings of the 10th Australian International Conference on Speech, Science and Technology, **2014**.

[12] N. Zhao, H. Yang, "*Realizing Speech to Gesture Conversion by Keyword Spotting*", IEEE Transactions, **2016**.

[13] K. Gopalan, T. Chu, X. Miao, "*An Utterance Recognition Technique for Keyword Spotting by Fusion of Bark Energy and MFCC Features*", Proceedings of the 9th World Scientific and Engineering Academy and Society (WSEAS) International Conference on Signal, Speech & Image Processing and 9th WSEAS International Conference on Multimedia, Internet & Video technologies, **2009**.

[14] K.S. Kavya, "*Automatic Recognition and Detection of Words in Speech- A Review*", International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET), **2017**.

[15] J.S.R. Alex, N. Venkatesan, "*Spoken Utterance Detection Using Dynamic Time Warping Method Along With a Hashing Technique*", International Journal of Engineering and Technology (IJET), Vol. **6**, No.**2**, **2014**.

[16] P. Karmacharya, "*Design of Keyword Spotting System Based on Segmental Time Warping of Quantized Features*", MS Thesis, Temple University.

[17] R. P. Ramachandran, R. J. Mammone, "*Modern Methods of Speech Processing*", Vol. **327**, Springer, **1995**.

[18] C. Kurian, K. Balakrishnan, "*Automated Transcription System for Malayalam Language*", International Journal of Computer Application, Vol. **19**, No.**5**, pp.**5–10**, **2011**.

[19] C. Vimla, V. Radha, "*Speaker-Independent Isolated Speech Recognition System for Tamil Language using HMM*", Procedia Engineering, Vol. **30**, pp.**1097–1102**, **2012**.

[20] J. Patel, K. S. Maurya, S. Kulkarni, V. Sakore, S. Khonde, "*Multimedia Keyword Spotting (MKWS) Using Training and Template-Based Techniques*", International Journal of Emerging Technology and Advanced Engineering, Vol. **4**, Issue.**2**, **2014**.

[21] S. Das, P.C Ching, "*Speaker Dependent Bengali Keyword spotting in unconstrained English Speech*", A Project report, Indian Institute of Technology Guwahati, India, **2005**.

[22] M. Lindasalwa, M. Begam, I. Elamvazuthi, "*Voice Recognition Algorithm using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques*", Journal Of Computing, Vol. **2**, Issue.**3**, pp. **138-143, 2010**.

[23] K. Dhameliya, "*Feature Extraction And Classification Techniques for Speaker Recognition: A Review*", IEEE International Conference on Electrical, Electronics, Signal, Communication and Optimization, pp.**1-4**, **2015.**

[24] P. Mahana, G. Singh, "*Comparative Analysis of Machine Learning Algorithms for Audio Signals Classification*", IJCSNS International Journal of Computer Science and Network Security, Vol. **15**, No.**6**, **2015**.

[25] Z. Chen, Y. Qin, K Yu, "*Sequence Discriminative Training for Deep Learning based Acoustic Keyword Spotting*", arXiv:1808.00639v1 [cs.CL], **2018**.

[26] A. Kadan, P. Vivek, V.L. Lajish, "*A Keyword Spotting Approach for Content-Based Indexing and Retrieval of Malayalam News Videos*", Conference Paper NSA-2015, **2015**.

[27] C. Zhu, Q.J. Kong, L. Zhou, G. Xiong, F. Zhu, "*Sensitive Keyword Spotting for Voice Alarm Systems*", In the Proceedings of 2013 IEEE International Conference on Service Operations and Logistics, and Informatics, pp.**350–353**, **2013**.

[28] V. Karjigi, B. Patel, P. Rao, "*Identification of stop consonants for acoustic keyword spotting in continuous speech*", Journal of Intelligent Systems, Vol. **22**, No.**3**, pp.**215-228**, **2013**.

[29] H. Bahi, N. Benati, "*A New Keyword Spotting Approach*", IEEE transactions, **2009**.

[30] M. Weintraub, "*LVCSR Log Likelihood Ratio Scoring For Keyword Spotting*", IEEE International Conference on Acoustics, Speech and Signal Processing, **2002**.

[31] S. Shetty, K.K. Achary, "*Audio Data Mining Using Multi-perceptron Artificial Neural Network*", IJCSNS International Journal of Computer Science and Network Security, Vol. **8**, No.**10**, pp.**224–229**, **2008**.

[32] Y. Zhang, J.R. Glass, "*Unsupervised Spoken Keyword Spotting via Segmental DTW on Gaussian Posterior grams*", IEEE Automatic Speech Recognition and Understanding Workshop, **2009**.

[33] C. Weng, B. Juang, "*Discriminative Training Using Non-Uniform Criteria for Keyword Spotting on Spontaneous Speech*", IEEE/ACM Transactions on Audio, Speech and Language Processing, **2015**.

[34] J. Junkawitsch, L. Neubauer, H. Höge, G. Ruske, "*A New Keyword Spotting Algorithm with Pre-Calculated Optimal threshold*", IEEE Conference Paper, **2002**.

[35] K. Thambiratnam, S. Sridharan, "*Isolated Word Verification Using Cohort Word Level Verification*", EUROSPEECH, **2003**.

[36] M. Awaid, A.H. Kandil, S.A. Fawzi, "*Audio Search Based on Keyword Spotting in Arabic Language*", International Journal of Advanced Computer Science and Applications, **2014**.

[37] S. Tabibian, A. Shokri, A. Akbari, B. Nasersharif, "*Performance Evaluation for an HMM-based Keyword Spotter and a Large-margin based one in Noisy Environments*", In Proceeding of the Computer Science, (ELSEVIER), **2011**.

[38] T. Nitta, S. Iseji, T. Fukuda, H. Yamada, K. Katsurada, "*Key-word Spotting Using Phonetic Distinctive Features Extracted from Output of an LVCSR Engine*", ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, **2003**.

[39] Y. Wang, P. Getreuer, T. Hughes, R.F. Lyon, R.A. Saurous, "*Trainable Fronted for Robust and Far-Field Keyword Spotting*", IEEE transactions, **2017**.

[40] G. Retsinas, G. Sfikas, B. Gatos, "*Transferable Deep Features for Keyword Spotting*", Multidisciplinary Digital Publishing Institute Proceedings, Vol. **2**, No.**89**, **2018**.

[41] V. Smirnov, D. Ignatov, M. Gusev, M. Farkhadov, N. Rumyantseva, M. Farkhadova, "*A Russian Keyword Spotting System Based on Large Vocabulary Continuous Speech Recognition and Linguistic Knowledge*", Journal of Electrical and Computer Engineering, **2016**.

[42] W. Amelia, N.U. Maulidevi, "*Dominant Emotion Recognition in Short Story Using Keyword Spotting Technique and Learning-based Method*", IEEE Transaction, **2016.**

[43] W. Li, A. Billard, H. Bourlard, "*Keyword Detection for Spontaneous Speech*", Conference Paper of IEEE 2nd International Congress on Image and Signal Processing, **2009**.

[44] J. Tejedor, J. Colás, "*Spanish Keyword Spotting System Based on Filler Models, Pseudo N-Gram Language Model and a Confidence Measure*", IV Jornadas en Tecnologia del Habla, **2006**.

[45] A.S. Park, J.R. Glass, "*Unsupervised Pattern Discovery in Speech*", IEEE Transactions on Audio, Speech and Language Processing, Vol. **16**, No.**1**, **2008**.
[46] G. Chen, C. Parada, G. Heigold, "*Small-Footprint Keyword Spotting Using Deep Neural Networks*", IEEE International Conference on Acoustics, Speech and Signal Processing, **2014**.
[47] Lin, Hui, A. Stupakov, J.A. Bilmes, "*Spoken keyword spotting via multi-lattice alignment*", INTERSPEECH, **2008**.
[48] A.H. Mansour, G.Z.A. Salh, H.H.Z. Alabdeen, "*Voice recognition Using Back Propagation Algorithm in Neural Networks*", International Journal of Computer Trends and Technology, Vol. **23**, No.**3**, **2015**.
[49] M. Limkar, R. Rao, V. Sagvekar, "*Isolated Digit Recognition Using MFCC and DTW*", International Journal on Advanced Electrical and Electronics Engineering, Vol. **1**, Issue.**1**, pp.**2278-8948**, **2012**.
[50] A. Zehetner, M. Hagmuller, F. Pernkopf, "*Wake-Up-Word Spotting or Mobile Systems*", Proceedings of the 22nd European Signal Processing Conference, IEEE, **2014**.
[51] J. Nouza, J. Silovsky, "*Fast Keyword Spotting in Telephone Speech*", Radio Engineering, Vol. **18**, No.**4**, **2009**.
[52] B. Varghese, S. Govilkar, "*A Survey on Various Word Spotting Techniques for Content-Based Document Image Retrieval*", International Journal of Computer Science and Information Technologies, Vol. **6**, No.**3**, pp.**2682-2686**, **2015**.
[53] Wshah, Safwan, G. Kumar, V. Govindaraju, "*Script Independent Word Spotting in Offline Handwritten Documents Based on Hidden Markov Models*", IEEE International Conference on Frontiers in Handwriting Recognition, **2012**.
[54] A. Bhardwaj, S. Setlur, V. Govindaraju, "*Keyword Spotting Techniques for Sanskrit Documents*", Sanskrit Computational Linguistics, Springer, Berlin Heidelberg, **2009**.
[55] Q. Chen, W. Zhang, X. Xu, X. Xing, "*Improved Keyword Spotting based on Keyword/Garbage Models*", IEEE Transactions, **2017**.
[56] L. Pandey, K. Chaudhary, R.M. Hegde, "*Fusion of Spectral and Prosodic Information using Combined Error Optimization for Keyword Spotting*", IEEE Conference on Communication, **2017**.
[57] I. Chen, C. Lee, "*A Hybrid HMM/DNN Approach to Keyword Spotting of Short Words*", INTERSPEECH, **2013**.
[58] S. Lubos, T. Jan, "*Keyword Spotting Result Post-processing to Reduce False Alarms*", Proceeding of Recent Advances in Signals and Systems, **2009**.
[59] G. Williams, S. Renals, "*Confidence Measures for Hybrid HMM/ANN Speech Recognition*", In Proceedings of Eurospeech, pp.**1955-1958**, **1997**.
[60] J.G. Wilpon, L.R. Rabiner, C.H. Lee, E.R. Goldman, "*Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models*", IEEE Transactions on Acoustics, Speech and Signal Processing (ASSP), Vol. **38**, No.**11**, pp.**1870-1878**, **1990**.
[61] M.G. Rahim, C.H. Lee, B.H. Juang, "*Discriminative Utterance Verification for Connected Digits Recognition*", IEEE Transactions on Speech and Audio Processing, Vol. **5**, No.**3**, **1997**.
[62] V.K. Jain, S. Tripathi, "*Speech Features Analysis and Biometric Person Identification in Multilingual Environment*", Int. J. Sc. Res. in Network Security and Communication, Vol. **6**, Issue.**1**, **2018**.
[63] Ketan Sarvakar, Urvashi K Kuchara, "*Sentiment Analysis of movie reviews: A new feature-based sentiment classification*", International Journal of Scientific Research in Computer Science and Engineering, Vol. **6**, Issue.**3**, pp**. 8-12**, **2018**.

Authors Profile

Mr B. K. Deka passed Bachelor of Science with Mathematics Major from Gauhati University, Guwahati, Assam in 1999 and Master of Computer Application from Indira Gandhi National Open University, New Delhi in the year 2006. He is currently pursuing PhD in Department of Computer Science and Engineering & Information Technology in Assam Don Bosco University, Guwahati, Assam. He has been working as an Assistant Professor, Department of Computer Science, NERIM Group of Institutions, Guwahati, Assam since 2010 till date. He has published more than 6 research papers in reputed international conferences including IEEE and it's also available online. His main research work focuses on Speech Processing, Data Mining. He has 8 years of teaching experience and 4 years of Research Experience.

Dr Pranab Das received his PhD in Computer Science & Engineering from Rajiv Gandhi University, Arunachal Pradesh in the year 2016. He is currently working as an Assistant Professor in the Department of Computer Science and Engineering & Information Technology, Assam Don Bosco University, Guwahati since 2009. He is a member of IAENG & ICST since 2011. He has published many research papers in reputed international journals and conferences including IEEE and is available online. His main research work focuses on Speech Processing, Machine Learning, Audio Mining. He has 12 years of teaching experience and 4 years of Research Experience.

