IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)
Volume 5, Issue 4, Ver. II (Jul - Aug. 2015), PP 22-27
e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
www.iosrjournals.org

Keyword Spotting: An Audio Mining Technique in Speech Processing – A Survey

Dr. E. Chandra[1], K.A.Senthildevi[2]
[1](Professor, Department of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu, India)
[2](Assistant Professor, Department of Computer Science, Gobi Arts & Science College, Tamil Nadu, India)

Abstract: *Audio mining is a branch of data mining that is used to search and analyze the contents of audio signal automatically. Keyword spotting (KWS) is an important audio mining technique which searches audio signals for finding the occurrences of given keyword within the input spoken utterance. KWS provides a satisfactory audio mining solution for various tasks like spoken document indexing and retrieval. The research in audio mining has received increasing attention due to the increase in amount of audio content in the Internet, telephone call conversation and other sources. KWS is classified according to the type of input speech content and the method used for spotting. A number of approaches have been used in keyword spotting like DTW, HMM, Neural Network, Vector quantization and other approaches. KWS has been utilized in a broad variety of applications. A majority of such applications relate to audio indexing and phone call routing. In this paper, various audio mining methods and keyword spotting techniques are discussed.*
Keywords: *Audio mining, DTW algorithms, feature extraction, keyword spotting.*

I. Introduction

With the voluminous increase in the amount of audio data, there is a need to explore new methods for accessing and mining these data. Recently, a number of researches have been developed in audio data to reduce manual accessing time and effort. Before processing the audio files, it must be represented with symbolic features to facilitate functioning of data mining. Data mining techniques can be applied to the speech features for finding relevant patterns and associations, retrieving information, monitoring keywords, indexing audio and etc. [1].

Audio mining is a technique which is used for searching and analyzing audio files for occurrences of spoken words or phrases [2]. Audio mining can be used to search specific characteristics of keywords within the huge and heterogeneous audio files. As a part of audio mining, keyword spotting is used to automatically find the occurrences of keywords of interest in spoken documents. The method is also used to get valuable information from huge quantities of speech records. The technology needs only to specify single or multiple instances of keywords to find in the speech corpus. Keyword spotting also provides necessary statistical information for various applications [3]. It is used in the on-line (e.g. real-time stream monitoring) and also offline (e.g. data mining and indexing) applications.

II. Audio Mining System

The main objective of audio mining technology is to search through speech for identifying specific characteristics. Audio Mining can analyze and search the contents of an audio signal for identifying patterns and associations, retrieving keywords and information. Audio can be in the form of radio, speech, music, etc. Due to the continuous, dynamic and non-structured nature of audio, audio files must be represented with spectral coefficients for further processing with data mining techniques. There are variant feature types to represent speech characteristics in numeric forms. The audio mining system can run at high speed that is several times faster than that of traditional systems. Hence large quantities of audio or speech can be searched in a short time.

2.1 Classification Of Audio Mining System

Audio mining techniques can be roughly classified into three techniques, namely Keyword Spotting System (KWS), Wake-up-word (WUW) detection, and Spoken term Detection (STD). KWS approach aims to detect the occurrences of keywords within the test spoken utterance. WUW is related to KWS, but it uses speech commands to activate or wake up other systems by an alerting signal.

Spoken Term Detection (STD) which is defined by NIST as an audio mining is employed for content based indexing. STD is aimed at open-vocabulary search over large collections of spoken documents. 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. However, keyword spotting is considered as a part of STD and both of them have been addressed in this survey.

2.2 Design Of Audio Mining System

Audio mining system consists of two main phases: training and template matching [3]. During the training phase, the input speech is preprocessed and training vectors are generated from the speech signal. The training vectors extract the spectral features for distinguishing different classes of words. The feature vectors are quantized and codebook is generated for further processing. During the template matching phase, the given test keyword is matched with the generated codebook using template matching algorithm. Similarity distance is calculated and the occurrences of keyword is found with the given threshold value. The phases of an audio mining system are shown in Fig.1 and Fig. 2.

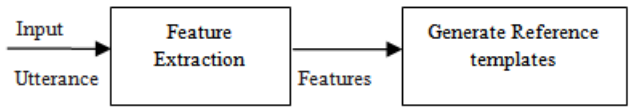

Fig.1. Feature Extraction Phase

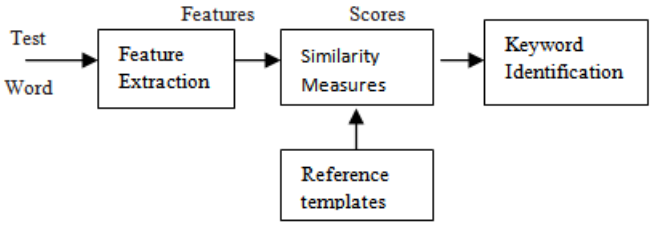

Fig.2. Template Matching Phase

2.3 Audio Mining Principles

To mine audio data, first it has to be converted into text or it has to be divided as phonemes or used directly. There are three basic principles that have been developed over the years for audio mining [2], [4].

2.3.1 LVCSR Audio Mining

In LVCSR audio mining, the entire test audio data is first transcribed into text and then the text transcription is used for searching the keywords. Hence LVCSR audio mining is a two-step process. In first phase, this method converts speech to text and in second phase, it identifies keywords in the generated dictionary that can contain several hundred thousand entries. If the keyword is not in the dictionary, the system will choose the most similar word it can find. LVCSR systems are more complex and expensive to implement.

2.3.2 Acoustic Audio Mining

In acoustic audio mining, the keyword search can be performed directly from untranscribed audio stream. Speech signals are represented as bark based energy or mel scale coefficients or they are represented as phoneme posteriorgrams generated by an acoustic model. This approach is easy to implement and provides some pronunciation tolerance.

2.3.3 Phonetic Audio Mining

Phonetic audio mining is phoneme based indexing method. This method doesn't convert speech to text but divides into phonemes. Phonetic audio mining is also a two-step process. In the first step, audio is processed (indexed) with a phonetic recognizer to generate a phonetic index file. In second step, system uses the generated dictionary of phonemes to compare user's search term to the correct phonetic string. This approach combines the advantages of LVCSR based keyword spotting and acoustic keyword spotting.

2.4 Terminology Used For Audio Mining

A number of different terms are used to refer audio mining. These include: audio mining, audio indexing, phonetic searching, phonetic indexing, speech indexing, audio analytics, speech analytics, word spotting, and information retrieval [2].

2.5 Difference Between Speech Recognition And Audio Mining

Speech technology is used to recognize the words or phonemes that are spoken in an audio or video file and an automatic speech recognition system is first trained with the entire content of audio file but audio mining searches can then be carried out to locate specific words and phrases within the audio. Keyword spotting is concerned only with matched words and their counts. There is no need to transcribe the entire audio file. Keyword spotting can be viewed as a sub-problem of automatic speech recognition, where only partial information (the keyword) has to be extracted from speech utterances [5]. Keyword Spotting is closely related to the task of speech transcription and offers many advantages for certain applications.

III. Keyword Spotting

Speech keyword spotting is an excellent technology in audio mining [6]. It is a retrieval of all instances of a given keyword in spoken utterances. Keyword spotting is well suited to data mining tasks that process large amount of speech such as real-time keyword monitoring and to audio document indexing. Keyword spotting is a technologically relevant problem, playing an important role in audio indexing and speech data mining applications [7]. KWS is similar to speech recognition but it ignores the additional signal information around the words of interest [8]. As this is a specific application of automatic speech recognition, most of the approaches used for ASR could also be used for KWS system with a little modification.

A keyword spotting system that simply detects occurrences of topical keywords will be more efficient than a fully-fledged LVCSR engine. Keyword spotting is a powerful and relevant technology and if it is used appropriately, it will bring with reduced computational requirements, increased scalability and potentially higher accuracies [6].

3.1 Types Of Keyword Spotting

There are several types of keyword spotting methods. Broadly, it is classified as speaker- dependent KWS and speaker - independent KWS. Based on the input speech data, KWS is also classified as

- Keyword spotting in continuous speech
- Keyword spotting in isolated speech

Keyword spotting in continuous speech is used when the input spoken keywords may not be separated from other words and Keyword spotting in isolated speech is used when keywords can be separated from other words [9]. The input for these systems may either be text or a spoken utterance. Systems that use the audio signals as test input are specifically termed as Spoken Keyword Spotting, or Spoken Term Detection [10].

3.2 Keyword Spotting Processes

The KWS has the following sub-processes [11], [12]:

a. Data Selection
b. Preprocessing
c. Feature Extraction
d. Vector Quantization
e. Template Matching
f. Decision Making

Various speech files like MIT corpus and TIMIT corpus are used in the existing KWS approaches. The spoken files are first preprocessed to reduce noise effects and are represented with spectral coefficients for further processing with data mining techniques. There are variant feature types like MFCC, LPC and etc., used to represent speech characteristics in numeric forms. Vector Quantization is done on the speech features of input speech and keyword for dividing them into finite number of cluster by using various clustering algorithms. Each cluster represents a similar sound. After quantization, the speech input and keyword are represented by a codebook with sequence of cluster indices. Similarity score between input speech and keyword are calculated and are compared with template matching algorithm for identifying occurrences of keyword given.

3.3 Application Of Keyword Spotting System

Major applications using keyword spotting technologies [3], [6] are:

i. Keyword monitoring applications
ii. Audio document indexing or searching
iii. Information retrieval
iv. Indexing and searching multimedia data
v. Route multimedia files and streams according to their content
vi. Music Information Retrieval
vii. Human-computer interaction

viii. Voice-command control
ix. Telecommunications services
x. Spoken password verification
xi. Security/defense.
xii. Command controlled devices
xiii. Dialogue systems
xiv. Monitoring of telephone services for target keywords.
xv. Person authorization

IV. Approaches In Implementing Keyword Spotting

The core objective of a keyword spotting system is to detect all instances of a given keyword within the given spoken utterance. For this purpose, various approaches have been developed over these recent years. The approaches are:

- Template Matching
- Hidden Markov Model
- Artificial Neural Network and
- Other Approaches

4.1 Template Matching Based Approaches For Keyword Spotting

Template matching algorithm such as Dynamic Time Warping (DTW) technique is widely used in KWS. Unlike other techniques, this does not need training of any models and also any addition of new keyword does not call for a retraining. In the basic approach, keywords are stored as templates in the database and matched against the unknown spoken utterance for similarity. Based on the similarity distortion, conclusion is drawn that a keyword is found or not. Because of the recent advancement in the computing power, DTW based system is mostly used in keyword spotting system [13].

4.2 HMM-Based Approaches For Keyword Spotting

HMM is commonly used for speech recognition, has also been used for keyword spotting. In this approach, HMM model is built for both keyword and test utterance. The model other than the keyword is referred as garbage model or filler model. The probability is calculated for each speech utterance to search if it is closer to the keyword. HMM model is used in most of the previous work of keyword spotting [8]. HMM based keyword spotting suffers from that it requires large amount of training data. It is time consuming and it requires language expertise. Also addition of new keywords requires retraining. There were many attempts to overcome these problems of HMM-based KWS. However, it again requires training which can cause a problem for audio indexing of new words. These problems related with the HMM, in recent years leads to the use of DTW based KWS [11], [13].

4.3 Neural Network Based Approaches For Keyword Spotting

In recent years keyword spotting is properly achieved by Artificial Neural Network (ANN). In ANN approach, the test audio is preprocessed to remove noises and then feature extraction is done using cepstral method. The ANN is trained with the cepstral values to produce a set of final weights. During testing process, these weights are used to mine the audio files. A Neural Network is constructed by highly interconnected processing units (neurons) which perform 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 [11], [14].

4.4 Other Approaches For Keyword Spotting

Besides the methods discussed above, there are some other methods for KWS based on Support Vector Machines (SVM), Iterative clustering, Self organizing Maps, Learning Vector Quantization, and Hybrid methods combining two or more methods are used. The methods obtained slight performance differences with the traditional methods but none of the methods significantly outperformed than the above discussed methods.

V. Related Works On Keyword Spotting With Modified Dtw Algorithms

A number of recent research in keyword spotting have been developed with Segmental Dynamic Time Warping (S DTW) algorithm which is used to find pattern matching between spoken utterances. Several speech representation techniques are used with Segmental DTW algorithm by various researchers for spotting keywords in speech.

Segmental DTW (SDTW) algorithm was first proposed in unsupervised pattern discovery in speech by A S Parked al [15]. SDTW takes two input spoken utterances and finds pattern matching between pairs of subsequences. This algorithm serves as the foundation for keyword spotting system in template based approaches. The approaches are discussed below.

Angeura ed al [16] proposed a task to find the recurrent occurrences of spoken words which was inspired by [15]. In his work, a speech summarization method was developed using Unbounded DTW algorithm. First recurrent acoustic patterns were discovered using this algorithm, and then they were clustered and ranked according to their number of occurrences in the input spoken database.

In paper [17], spoken term detection system based on queries by example was developed. The system used template matching approach (SDTW) and phonetic posteriorgram representation of speech signal. Audio snippets are used as queries instead of using word or phone strings. SDTW algorithm is used to calculate similarities between phonetic posteriorgrams of test utterance and spoken query word. The system is first tested with single query and then it was improved with multiple queries. The result obtained by the authors in this work is better than the result obtained in their previous work with DHMM.

In paper [18], Keyword spotting system was developed in completely unsupervised GMM method. Without any transcription, each speech frame was represented as Gaussian posteriorgram and SDTW was used to compare the Gaussian Posteriorgram between keywords and speech file. The system was initially tested with TIMIT Corpus and then was evaluated with MIT lecture corpus.

The above work was reviewed by Karmacharya ed al [19]. He also developed a keyword spotting system by quantizing the speech MFCC feature vectors by K means clustering algorithm and by using SDTW algorithm for comparing the quantized vectors of keywords and test speech utterances [20]. His work is similar to the work in [18], but he used quantized features instead of Gaussian Posteriorgram. The system performance was evaluated with the Call-Home and the switchboard corpus. Maximum accuracy 90% was achieved in this system that is comparable with the previous work done with the same SDTW algorithm.

In the above papers [17], [18], [20], the spoken keywords are compared with spoken documents frame by frame. But in paper [21], spoken keyword and spoken document are first segmented by clustering algorithm and then they are compared with segment– based DTW algorithm. Hierarchical agglomerative clustering algorithm was used to segment speech signals and clusters were formed in tree structure. In this method, the number of redundant data was reduced. The authors also proposed a two-pass framework for keyword spotting [22]. They used segment-based DTW in first pass to locate hypothesized spoken term and the frame-based DTW is used in second pass to expand acoustic variation in the query. They achieved lower computation load than that in the former work.

In [23], a fusion of spectral and cepstral features was used for a keyword spotting system. Bark scale based energy and MFCC are used independently and in combination with appropriate weights for identifying keywords in spoken utterances. DTW algorithm was used as template matching algorithm to determine similarity between reference keywords and unknown utterances. The result with the combination of these features was highly comparable with the individual's results.

In [24], a keyword spotting system for searching mathematics formula was designed with Gaussian Mixture Model. The experiments were conducted using SDTW and also with Lower Bound IP algorithms. Both performance showed better results.

In paper [25], the author proposed KWS based on MEaRBS rule segmentation. He used quasi-syllable segments with classical DTW algorithm. He compared MEaRBS method with KWS algorithm based on sliding window and the experiments showed that his proposed method performed 5% to 25% better performance than the classical method.

In [26], the author evaluated various acoustic features in the context of spoken term discovery with DTW-based framework. He used two broad feature categories. That is, features representing spectral properties like MFCCS, PLPs and features derived from high resources like FDLP based features and MLP based features are used in his evaluation.

Authors in [27], used DTW based distance histogram for computing similarity threshold for every keyword and test utterance pair in keyword spotting system. He obtained high recall ratio than a HMM - based approach.

Recently, Multimedia Keyword Spotting (MKWS) was proposed in [28]. The paper presented a keyword spotter for searching a spoken keyword in a multimedia file. Both HMM and DTW techniques are used in this approach. HMM is used to represent the spoken words in phoneme representation. And DTW was used for measuring similarities between the keyword and spoken utterance.

VI. Conclusion

In this paper, audio mining and its branch keyword spotting technologies are analyzed. Audio mining is performed in three different methods. Keyword spotting is a technologically relevant problem in audio mining to automatically detect keywords of interest in spoken document, and has been utilized in a broad variety of applications. Based on the input speech data, KWS is broadly classified as Keyword spotting in continuous speech and Keyword spotting in isolated speech. It plays an important role in audio mining. As a part of audio mining, keyword spotting is achieved with various approaches such as DTW, HMM, ANN and etc. HMM based keyword spotting suffers from number of problems. The HMM approaches require a large amount of supervised training data. Due to these reasons, DTW is mostly used in keyword spotting systems.

References

[1]. Zheng-Hua Tan, "Audio and Speech Processing for Data Mining," Multimedia, Aalborg University, Denmark.
[2]. Manpreet Kaur Mand, Diana Nagpal, Gunjan, "An Analytical Approach for Mining Audio Signals", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013,ISSN 2319-5940.
[3]. Phonexia Keyword Spotting, White paper, www.phonexia.com
[4]. Anna M. Kruspe, "Keyword Spotting in a CAPELLA SINGING," 15th International Society for Music Information Retrieval Conference (ISMIR 2014)
[5]. Guoguo Chen, "Low Resource Keyword Spotting," Department of Electrical and Computer Engineering Johns Hopkins University Baltimore, Maryland
[6]. Thambiratnam, Albert J. K," Acoustic keyword spotting in speech with applications to data mining." PhD Thesis Published in Queensland University of Technology, 2005
[7]. Jansen, A., Niyogi, P.: Point process models for spotting keywords in continuous speech. Audio, Speech, and Language Processing, IEEE Transactions on 17(8) (2009) 1457{1470 IEEE.
[8]. Ramachandran, R.P., Mammone, R.J.: Modern methods of speech processing. Volume 327. Springer (1995)
[9]. Leverkuhn ,'What Is Keyword Spotting, 'Last Modified Date: 01 December 2013, What Is Keyword Spotting.htm
[10]. M R Srikanth, "Speaker Verification and Keyword Spotting Systems for Forensic Applications," PhD Thesis, IIT. Madras, December 2013
[11]. Piush Karmacharya, "Design of Keyword Spotting System Based on Segmental Time Warp ing of Quantized Features," MS thesis ,Temple University.
[12]. K.A.Senthildevi, Dr. E.Chandra, " Keyword spotting in Isolated MFCCs and DTW algorithm," 4th IEEE conference on communication and signal processing, 2015.
[13]. John Sahaya Rani Alex, Nithya Venkatesan, "Spoken Utterance Detection Using Dynamic Time Warping Method Along With a Hashing technique ," International Journal of Engineering and Technology (IJET)
[14]. S.Shetty, and K.K. Achary, "Audio Data Mining Using Multi- perceptron Artificial Neural Network," International Journal of Computer Science and Network Security, vol.8, pp.224-229, Oct. 2008
[15]. Park and J. Glass,"Unsupervised pattern discovery in speech", in IEEE Trans. ASLP, 1558–1569, 2008.
[16]. Xavier Anguera, "Spoken word cloud: Clustering Recurrent patterns in speech," in International Workshop on content based Multimedia Indexing, June 2011.
[17]. Zhang and J. R. Glass, "Unsupervised spoken keyword spot- ting via segmental dtw on gaussian posteriorgrams," in ASRU, 2009
[18]. Yaodong Zhang and James R Glass, "Unsupervised Spoken Keyword Spotting via Segmental DTW on Gaussian Posterior- grams," in Proc. ASRU, Merano, Italy, 2009, pp. 398–403
[19]. Piush Karmacharya, "A short Literature Review on Unsupervised Spoken Keyword Spotting via Segmental DTW on Gaussian Posteriorgrams".
[20]. Timothy J. Hazen, Wade Shen, Christopher M White, and A Phonetic Posteriorgram Representation, "Query-By- Example Spoken Term Detection Using Phonetic Posterior- gram Templates," in ASRU, 2009, pp. 421–426.
[21]. Chun-an Chan and Lin-shan Lee, "Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warp- ing," in Interspeech, 2010.
[22]. Chun-an Chan and Lin-shan Lee, "Integrating frame-based and segment-based Dynamic Time Warping for unsupervised spoken term detection with spoken queries," IEEE, ICASSP 2011.
[23]. K. Gopalan, Tao Chu, and Xiaofeng Miao, "An Utterance Recognition Technique for Keyword Spotting by Fusion of Bark Energy and MFCC Features".
[24]. Mullier,"Keyword spottingin audio for access maths".
[25]. Mindaugas Greibus and Laimutis Telksnys,"Speech Keyword Spotting with Rule Based Segmentation".
[26]. Aren Jansen and Benjamin Van Durme, "Efficient Spoken Term Discovery Using Randomized Algorithms," in ASRU,2011.
[27]. M. S. Barakat, C. H. Ritz & D. A. Stirling, "Keyword spotting based on the analysis of template matching distances," in 5th International conference.
[28]. Jigar Patel, Kailash Singh Maurya,Sameer Kulkarni, Vaibhav Sakore, Shraddha Khonde,"Multimedia Keyword Spotting(MKWS) Using Training And Template Based Techniques".