Audio Classification: A Comprehensive Survey of Research

Arshi Khan*, Prof. Garbita Gupta**, Dr. Smita Shandilya*** and Dr. Shishir K. Shandilya****
*-****Department of CSE, Bansal Institute of Science & Technology, INDIA

Abstract: Data mining is define as to extract knowledge from large amount of data. Data mining has a research application in the field of audio, speech processing & spoken word language, so as to get useful data from large amount of data. In this paper we have described about audio mining to extract useful audio signals for classification of audio data. Various audio features like Mel frequency Cepstral Coefficient (MFCC), Linear Predictive Coefficient (LPC), Compactness, Spectral Flux (SF), Band Periodicity (BP), Zero Crossing Rate (ZCR) etc are used to classify audio data into various classes. Various classification algorithms such as Naive Bayes, SVM and PNN are used to classify audio data into defined classes. Using various performance parameters such as True Positive (TP) Rate, False Positive (FP) Rate etc., results of various classification algorithms are compared.

Keywords: DataMining, Audio Mining, Audio Classification, Classification algorithms.

Introduction

Huge amount of data is increasing day by day and is available on the internet as to mine large data various data mining techniques are available like classification, clustering, pattern analysis etc. Audio data is available from various sources in the form of newscast, telephonic conversation and recording in office meetings, business meetings. As in past era where it is not possible to analyze, recognize and how to interpret digitized data, for this large companies or firms used to do the manual process for analyzing written transcript of audio data. Therefore to recognize, analyze & interpret huge data we have at present large capacity for storage, better audio classification algorithms, faster microprocessor processing so as to get audio data with the help of audio mining.

Audio Mining

In order to save a large storage space MP3 format is the best tool used for audio data as better performance in sound quality and compression techniques. One of the compression techniques in audio data is PCM format where no header is required for saving the data. To classify data into number of classes various audio formats came into existence. Audio mining is widely used to detect audio files on the basis of spoken words occurrence. Search Technique is applied to calculate words which are occurred frequently in audio data. Audio mining works on 2 indexing mainly:-

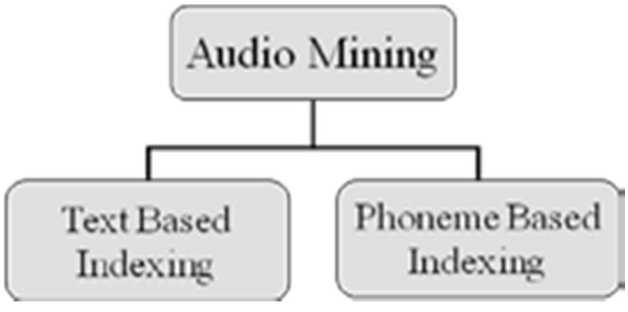

Fig. 1 : Audio Mining Approaches

1. Text Indexing:- Here large data is converted into useful form by applying speech recognition algorithm in the form of LVCSR library. Users put a query and according to that words are searched from that library and we count the most frequent words occurred in the library, so as to make searching faster and reliable with better performance.
2. Phoneme based indexing:- It doesn't convert text data but do the audio data conversion into audio signal in the form of digitized way. It uses several numbers of phonemes from a dictionary to search user defined phoneme from the library.

Audio Data Conversion
The sound or audio information is accessible in many formats. It has been watched that PCM format is a reasonable configuration for processing of information, as it contains sound information and no header data is added to it. PCM contain information in uncompressed form so, it is easy to read. Here Sample Format is physically set to 16bits and Bit rate to 16000 Hz (Samples every second) for feature extraction. These values are chosen from audio data because any sample rates less than this will results least or degraded quality of audio and it would be very much difficult to get useful information out of it. Likewise, higher specimen rate may bring about additional overhead in calculation.

Audio Size
The span of sound media is an imperative consideration in applications including sound. The capacity required for the sound information, and the time required for download and transmission from the web are critical concerns. For the achievement of sound applications, it is vital to make media questions little in sizes.
Long however not precisely repeated sounds, would require tremendous capacity. In numerous applications, there is a requirement for a basic sound of subjective length, such sounds are moderately monotonic, straightforward in structure, and have repeated yet potentially factor sound patterns.

Audio Texture
Audio Texture [3] is an effective technique for creating sounds from illustration clips. Audio surface gives a productive method for integrating constant, perceptually important, yet non-dull sound stream from a case sound clip. Perceptually significant suggests that the combined sound stream is perceptually like the given example clip. For producing sound surfaces a two-organize strategy is proposed-

Analysis Stage
In this phase by removing its building patterns or identically discovering design breakpoints the illustration clip is examined and divided into sub-clips. This step depends on the comparability measures between every two frames as indicated by their Mel recurrence cepstral coefficients (MFCCs).

Synthesis Stage
In this stage the arrangement to play the sub-clips or building patterns is chosen. Variable impacts can be consolidated into the building patterns to avoid repetitiveness of the combined sound stream.

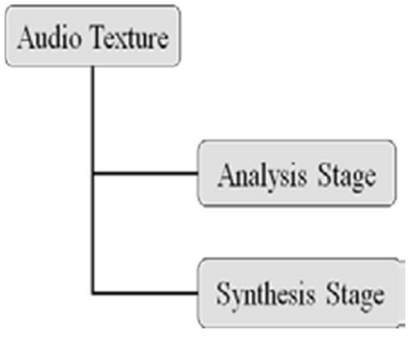

Fig. 2: Method to generate Audio Texture

Methodology Used

Audio Classification
Classification groups information to pre-characterized classes. The grouping application constructs a model from the prepared classes and uses that model to arrange new objects into one of the predefined classes consequently. It is a best approach among the most generally utilized information mining strategies. The primary concentration of data mining procedures is to mine important information. Speech recognition is one of the classification characterization applications. Time arrangement coordinating and order have gotten much consideration in Speech recognition research field . The aim of a Speech recognition framework is to interpret speech signals into the textual format. The classification is grouping of similar data is the first step in feature extraction. In Speech recognition, the standard examination strategy is to isolate the speech pattern into frames and apply an feature

extraction technique on each frame. Following order calculations [4] are chosen to effectively arrange sound information:

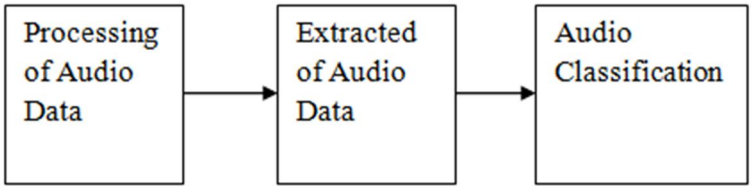

Fig. 3 : Audio Classification Process

Classification Model

Fig. 9 demonstrates the audio characterization model. Audio components are extracted from sound information for arrangement. Sound information is partitioned into preparing and testing set. A negligible list of capabilities is extracted from sound features. A classifier model is fabricate utilizing insignificant list of capabilities and preparing information and utilizing a test set, the classifier model is further tested. Naive Bayes, ID3, J48, FT and LibSVM calculations are utilized to accurately group audio information. The outcomes acquired from the classifier model are further analyzed by utilizing execution parameters.

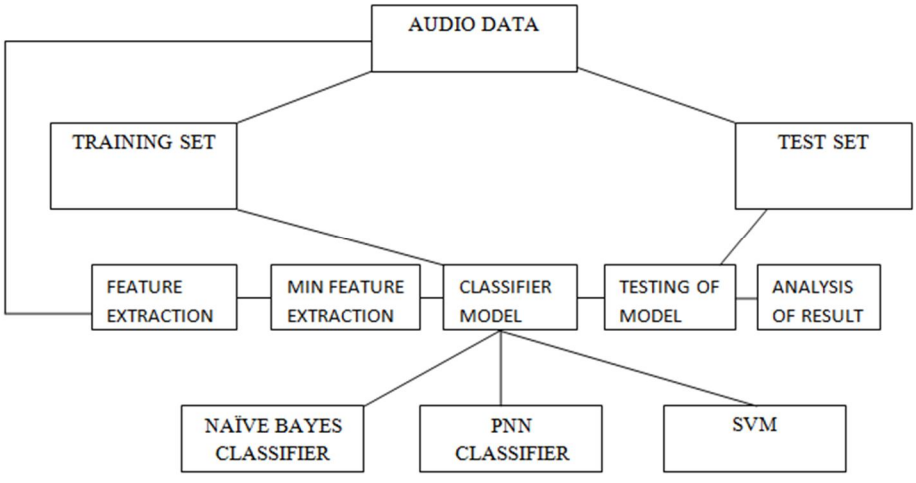

Fig. 4 : Audio Model

Naïve Bayes Classifier

Bayesian classifiers are measurable classifiers. They help in predicting class participation probabilities, for example, a likelihood that a given tuples has a place with a specific class. Bayesian order depends on Bayes Theorem. Naive Bayes classifier is the most simplest of the Bayesian classifiers. Naive Bayesian classifiers depend on the presumption that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence.

Naïve Bayes classifier accept that the presence (or absence) of a specific component of a class is irrelevant to the nearness (or nonattendance) of some other element. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Regardless of the possibility that these elements rely on upon each other or upon the presence of alternate components, a Naive Bayes classifier considers these properties to independently add to the likelihood that this organic product is an apple.

An advantage of the Naive Bayes classifier is that it just requires a little measure of preparing information to estimate the parameters, for example, means and variance of the factors essential for characterization. Since independent variables are expected, only the variances of the variables for each class should be resolved and not the whole covariance matrix.

Support Vector Machines (SVM)

Created by Vladimir Vapnik and his collaborators at AT&T Bell Labs in the mid 90's, Support Vector Machines (SVM) is an arrangement of related directed machine learning approaches utilized for characterization and regression.SVM is an idea in software engineering for an arrangement of related administered learning strategies that analyze data and recognize patterns.. It is utilized for characterization and relapse investigation. SVM which consist of an arrangement of information

and try to predicts, for each given input, two possible classes are there in an individual from, which makes the SVM a non-probabilistic paired direct classifier. As such a set of training data as an example is given, each set is termed to mark as one of two classifications, a SVM which prepares a calculation that assembles a model to assigns new cases into one classification or another. A SVM is termed to represent a model consist of patterns in a space, which mapped so that the points of the different classifications are separated by clear gape that is so wide as could be expected under different situations. Various new examples are then mapped into that same space of time and prediction is made which belong to a category based on which side of the gap they fall on.

SVM constructs a hyperplane or set of hyperplanes in a high- or infinite- dimensional space, which can be used for classification, regression, or other tasks. A good separation is occurred in terms of largest distance by the hyperplane to the nearest training data points belong to any class. since the larger the margin the lower the generalization error of the classifier. Whereas the original problem may be stated in a finite dimensional space, it often happens that in that space the sets to be discriminated are not linearly separable. For this reason many assumptions are made and then it was proposed that the original finite-dimensional space must be mapped into a much higher- dimensional space, presumably making the separation easier in that space. SVM schemes use a mapping into a larger space so that cross products may be computed easily in terms of the variables in the original space, making the computational load reasonable. The cross products in the larger space are defined in terms of a kernel function $K(x,y)$ selected to suit the problem The hyperplanes in the higher dimensional space are characterized as the set of points whose inward item with a vector in that space is consistent. The vectors characterizing the hyperplanes can be chosen to be linear combinations with parameters αi of images of feature vectors that occur in the data base.

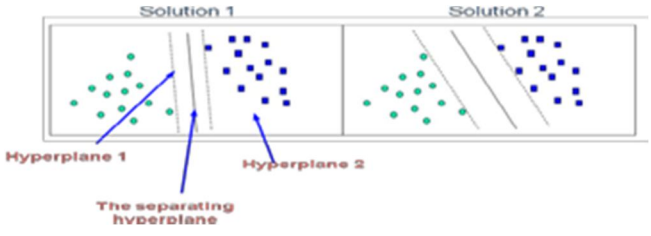

Fig. 5 : A 2-D Hyperplane in SVM

PNN Classifier

A probabilistic neural network (PNN) is predominantly a classifier which maps any input patterns to various arrangements and can be constrained into a more broad capacity approximator. A PNN is a usage of a measurable calculation statistical algorithm called kernel discriminant analysis in which the operations are sorted out into a multilayered feedforward connect with four layers:Input layer

1. Pattern layer
2. Summation layer
3. Output layer

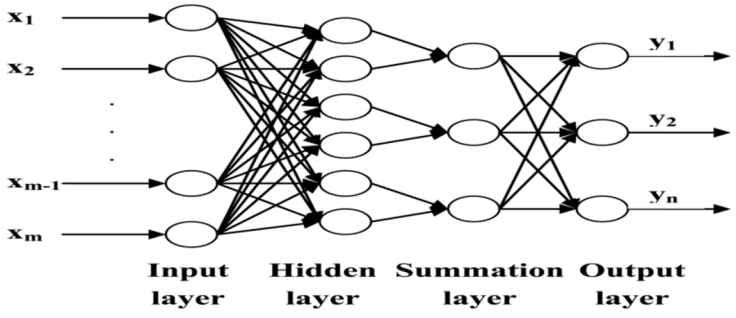

Fig. 6 : PNN Classifier

Clustering

Clustering provides an attractive mechanism to consequently discover some structure in expansive or large dataset that would be generally hard to compress. Clustering of sequences or time arrangement is an approach to gathering an accumulation of groupings based on their similarity. There are a variety of methods for clustering. As in sequence clustering, ARMA models or Hidden Markov Models are widely used. The other broad class in sequence clustering

uses pattern alignment-based scoring or similarity measure to compare sequences. In speech applications, a data mining system with clustering algorithm is used to find useful patterns from speech database.

Search and Retrieval

Searching for sequences in large databases is an important task in data mining. Sequence search and retrieval techniques play an important role in interactive explorations. In content-based retrieval, the task is to search a large database of sequential data and retrieve from it sequences or subsequences similar to the given query sequence. In speech or audio applications, the individual elements of the sequences may be feature vectors of real numbers. When the sequential elements are feature vectors, Euclidean distance may be used for measuring similarity between two elements.

In speech or audio signals, similar wounding patterns may give feature vectors that have large Euclidean distances and vice versa. An elaborate treatment of distortion measure for Speech and Audio signals can be found. In speech applications, Dynamic Time Warping (DTW) is a systematic and efficient method that identifies which correspondence among feature vectors of two sequences is best when scoring the similarity between them.

Evolution

According to Paul MC Fedries [9] states that a process for extraction of sentences or a phrases or index words & search the words in audio file for audio mining data.

Hall [8] states an audio mining is a combination of various fields as pattern discovery, classification, and speech recognition, NLTP which translates or connect audio data into human readable format or textual format. In addition various search algorithms are also used to classify, cluster, pattern format or transcribe the audio data into computer based text data.

Dragon System [10] states that for searching keywords, phrases, sentences in audio data it is cumbersome to convert audio or conversation into a searchable format. It is hard to listened long hour's audio conversation so to reduce the listening time various speech recognition techniques are developed which is reliable, consume less time and enhance productivity.

According to Neal Leavitt in large audio data users can randomly pick the useful data instead of accessing the whole data .To extract number of features audio signals are used to discover pattern mining like music, song, business, recording, official recording, political recording, tune and melody, listening to pitches, rhythms.

Data mining is an important tool to extract useful data from large set of volume13].

Speech Recognition aim is to develop techniques and system to give input and machine give output [14].

Speech Recognition is the method of converting speech signal to number of frequent words with the help of an algorithm implemented on a computer [15].

Voice or Speaker recognition is the method of analyzing the translated speech [16] to refer the recognition system to trained the speaker.

Audio data is to be mined broadcast recorded meetings of office, business and telephonic data [17].

Voice data mining (VDM) is a multi-lingual voice processing system that is proficient in mining specific keywords from a large audio repository. It deals with the need to organize, search and retrieve collection of spoken documents such as recorded telephony conversations, TV or radio archives in an effective, efficient manner [18].

Surveillance video often consist of events that are not occurred previously, and consist of target for unsupervised discovery of patterns, which in this case are termed as events [20].

The feature extraction step in sequence recognition applications typically generates, for each pattern such as a speech utterance, a sequence of feature vectors that must then be subjected to a classification step. Sequence classification applications use pattern based as well as model-based methods [21].

Call centers recordings and telephone survey corpora contain a large variety of speakers with bad audio quality due to cell phones and surrounding noises, unconstrained speech, variable utterance length and numerous disfluences like hesitations, reptitions and corrections[22].

The system can work on very noisy automatic transcriptions of spoken messages. There is a need to quantify the extent of similarity between any two (sub) sequences. An application was developed to extract the distribution of user's opinions from telephone surveys. The system works on very noisy automatic transcriptions of spoken messages [23].

The application presents a data mining system designed to be associated with TERAPERS system in order to provide information based on which one could improve the process of personalized therapy of speech disorders [25].

The tools increase the productivity of the analyst who seeks relationships among the contents of multiple utterances. It is shown how data mining techniques that are typically applied to text should be modified to enable an analyst to do effective semantic data mining on a large collection of short speech utterances [28].

A speech data mining system is used in generating a rich transcription having utility in call center management. One ot the systems is speech differentiation module which differentiating between speech of interacting speakers. Another is a speech recognition module improving automatic recognition of speech of one speaker based on interaction with another speaker employed as a reference speaker [29].

Recent Scenario

According to Silk Smita , Sharmila Biswas , Sandeep Singh Solanki [31] states that musical signal of harmonic structure is stable in nature, Different instruments are used to classify different signals for harmonic data, this concept is not valid for non- structure harmonic like drums, instead of it other algorithm's are used to classify drum sounds. Different classifiers are used to classify data into different classes but SVM classifier is used to prove the best classifier as its accuracy level is more in terms of elapsed time.

According to Varsha gupta, Anuj Sharma [34] defines a SHPRW by using a method of feature extraction MFCC and they used artificial neural network for classification.

According to Neeru Rathee [33] defines a lip reading using back propagation network for word recognition and used dynamic features in three orthogonal planes instead of landmark localization approach which makes computation exhaustive.

According to Frederic Le Bel [32] states that for a classification large amount of corpuses sound is used for processing computerized classifier or simply it develops a database under some constraints to organize or prepare a melody sound.

Challenges

When used in real time environment audio mining search and process data as fast as compared to human analyzing power due to this the accuracy level decreases .Audio mining is relatively depend on various factors as background noise and cross talk. Audio mining has number of domain specific knowledge which lies between noise and cross talk data which is not accurate for timestamps and speakers. All these issues makes audio mining task difficult to process audio data.

Conclusion

Various data mining techniques are available to mine huge amount of audio data in the field of prediction analysis, searching and information retrieval, learning new language concepts. New learning methods are developed to extract automatic data. For many applications like business and government new techniques are developed for audio, speech data, to extract useful information like documents, text summarized data, important news. Data mining techniques is the best emerging field in the research area where different classifier is used to classify the data by using classification algorithms with minimum feature set which varies from 8 to 9 factors. With the training set we can increase and decrease FP rate by doing classification as well as clustering. Naïve Bayes classifier, SVM, PNN are used and results are compared, SVM gives the best performance with respect to TP rate and lower FP rate.

References

[1] Neal Leavitt, "Let's Hear It for Audio Mining", IEEE Computer Magazine on Technology News, ISSN: 00189162, Volume 35, Issue 10, October 2002, pages 23-25

[2] Jonathan Foote, "An Overview of audio information retrieval", ACM Multimedia Systems, Volume 7, Issue 1, Jan 1999, pp. 2–10.

[3] L. Lu, S. Li, L. Wenyin, H. J. Zhang, Y. Mao, "Audio Textures". Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, ISSN:1520-6149, Volume 2, 13-17 May, 2002 pp. 1761–1764.

[4] E.Wold, T. Blum, D. Keislar, J. Wheaton, "Content- based classification, search, and retrieval of audio," IEEE Transaction on Multimedia, Volume 3, Issue 3, 1996 pp.

[5] Huang T.M., Kecman V., Kopriva I., "Kernel Based Algorithms for Mining Huge Data Sets, Supervised, Semi- supervised, and Unsupervised Learning". Springer-Verlag, ISBN 3-540-31681-7, 2006 pp 260.

[6] Chang, C. Lin, (2001). "LIBSVM: A Library for Support Vector Machines". [Online] Available :http://www.csie. ntu.edu.tw/cjlin/libsvm

[7] [Online] Available : http://www.cs.waikato.ac.nz/ml/weka

[8] [Online] Available : http://www.cutter.com/research/2002/edge020312.html

[9] [Online] Available : http://www.wordspy.com/words/audiomining.asp

[10] [Online] Available : http://www.voicerecognition.com/news/2_899.html

[11] Giampiero Salvi , "Mining speech sound", Machine Learning Methods for Automatic Speech Recognition and Analysis.

[12] "Speech Data Mining & Document Retrieval", publication of the IEEE signal processing.

[13] Surendra Shetty, K.K Achar, "Audio Data Mining Using Multi-perceptron ANN" International Journal of Computer Science and Network Security, Vol. 8, No.10 October 2008.

[14] M.A. Anusuya, S.K. Katti, "Speech Recognition by Machine: A Review", International Journal of Computer Science and Information Security,Vol. 6, No.3,2009.

[15] Santosh K. Gaikwad, Dr. Babasaheb Ambedkar Marathwada, Bharti W.Gawali, "A Review on Speech Recognition Technique", International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2010.

[16] National Science and Technology Council(NSTC),"Speaker Recognition" last updated Aug 2006.

[17] Dai Sheng-Hui, Lin Gang-Yong, Zhou Hua-Qing," A Novel Method for Speech Data Mining", Journal of Software, Vol. 6, No. 1, January 2011.

[18] A star exploit technologies," Voice Data Mining– Index & Retrieve Spoken Documents".

[19] Neal Leavitt**,"** Let's Hear It for Audio Mining",http://www.leavcom.com/pdf/Audio.pdf
[20] Ajay Divakaran, Koji Miyahara, Kadir A. Peker, Regunathan Radhakrishnan," Video Mining using Combinations of Unsupervised and Supervised Learning Techniques", Merl – A Mitsubishi Electric Research Laboratory.
[21] Srivatsn Laxman and P S Sastry ," A survey of temporal Data Mining", Sadhana Vol 2, April 2006.
[22] Andrzej CZYZEWSKI ," Mining Knowledge in Noisy Audio Data", Proceeding in KDD-96.
[23] Nathalie Camelin, Frederic Bechet, Geraldine Damnati, "Speech Mining in Noisy Audit Message Corpus" Antwerp Belgium, August 27-31
[24] Wei Zha , Wai – Yip Chan , " A Data Mining Approach to Objective Speech Quality measurement", ICASSP 2004
[25] Mirela Danubuanu, Stefan Gheorghe Pentiuc,"Model of a Data mining System for Personalized Therapy of Speech Disorders".
[26] Ying Shi; Weihua Song, "Speech emotion recognition based on data mining technology", Sixth International Conference on Natural Computation- Aug 2010.
[27] Park, A.S. Glass, J.R.," Unsupervised Pattern Discovery in Speech", IEEE transaction on Audio, Speech, and Language processing- Jan. 2008
[28] Lee Begeja, Harris Drucker, David Gibbon1, Patrick Haffner, Zhu Liu, Bernard Renger, Behzad Shahraray, "Semantic Data Mining of Short Utterances**",** IEEE Trans. On Speech & Audio Proc.: Special Issue on Data Mining of Speech, Audio and Dialog July 1, 2005.
[29] Marcin Paprzycki, Ajith Abraham, Ruiyuan Guo and Srinivas Mukkamala," Data Mining Approach for Analyzing Call Center Performance**".**
[30] Napoletani D.Struppa DC,Sauer T, Berenstein CA, Walnut D, " Delay-coordinates embeddings as a data mining tool for denoising speech signals"**.**
[31] Silk Smita , Sharmila Biswas , Sandeep Singh Solanki- Audio Signal Separation and Classification: A Review Paper (2014).
[32] Frédéric LE BEL – Agglomerative Clustering for Audio Classification using Low-level Descriptors(2017).
[33] Neeru Rathee- Investigating Back Propagation Neural Network for Lip Reading(2016).
[34] Varsha Gupta, Anuj Sharma- Classification of the spoken Hindi Partially reduplicated Words using Artifical Neural Network (2014).