Survey on Knowledge Discovery in Speech Emotion Detection

S.Jagadeesh Soundappan, Dr.R.Sugumar

Research Scholar, Dept. of CSE, St. Peter's University, Chennai, India

Associate Professor, Dept. of CSE, Veltech Multitech Dr.RR Dr.SR Engineering College, Chennai, India

Abstract; Knowledge discovery refers to finding some relevant information out of a bulk amount of data. Speech emotion recognition is one of the major areas in the knowledge based discovery. This research work has been carried out using four emotions namely sad happy angry and aggressive. This research work possesses two sections namely training and the testing part. The training part will consists of the updation of the speech files with the data base system. Once a file is uploaded, the system would extract the features of the speech file with an algorithm named MFCC. The MFCC algorithm would extract a feature vector out the speech file and then the maximum, minimum and average value of the feature vector would be saved into the database. The process would repeat itself again and again till the last category is not achieved. Once the training part is complete, the testing section would be initiated. The testing section would involve the classification process in which two classifiers would be used. The first classifier is neural networks whose back propagation feed forward neural network would be used for the processing. The BPNN is one the most affective classifier out of the available classifiers. The initial hidden layer in the BPNN process has been kept as 20 and minimum number of iterations is 5. Some sort of previous work has been also implemented before this research work getting proposed like use of BPNN for speech classification but the combination of MFCC, BPNN for the same feature set has not been proposed yet. To show the effectiveness of the work, the same process has been repeated with Support Vector Machine and the accuracy would be measured in both the cases.

KEY WORDS: MFCC, SVM, Neural network.

I.INTRODUCTION

Audio mining

Audio Mining is a technique by which the content of an audio signal can be automatically analysed and searched. It is most commonly used in the field of automatic speech recognition, where the analysis tries to identify any speech within the audio. The audio will typically be processed by a speech recognition system in order to identify word or phoneme units that are likely to occur in the spoken content. This information may either be used immediately in pre-defined searches for keywords or phrases (a real-time "word spotting" system), or the output of the speech recogniser may be stored in an index file. One or more audio mining index files can then be loaded at a later date in order to run searches for keywords or phrases. The results of a search will normally be in terms of hits, which are regions within files that are good matches for the chosen keywords. The user may then be able to listen to the audio corresponding to these hits in order to verify if a correct match was found.

Audio mining systems used in the field of speech recognition are often divided into two groups: those that use Large Vocabulary Continuous Speech Recognisers (LVCSR) and those that use phonetic recognition. The Web, databases, and other digitized information storehouses contain agrowing volume of audio content. For example newscasts, sporting events, telephone conversations, recordings of meetings, Webcasts, documentary archives etc. Users want to make the most of this

material by searching and indexing the digitized audio content. In the past, companies had to create and manually analyze written transcripts of audio content because using computers to recognize, interpret and analyze digitized speech was difficult. However, the development of faster microprocessors, larger storage capacities, and better speech-recognition algorithms has made audio mining easier.

Phonetic recognition does not require the use of complex language models: the phone recognition can be run effectively without knowledge of which phones were previously recognized. In contrast, knowledge of which words were previously recognized is vital for achieving good recognition accuracy in a large vocabulary system. LVCSR approaches must therefore use sophisticated language models, which leads to a much greater computation load at the indexing stage for LVCSR approaches and results in significantly slower indexing speeds. Phonetic audio mining software can index audio data at rates of around 100 times faster than real time, compared to speeds only a few times faster than real time for LVCSR systems. The reliance on complex language models also means that the data used to train the LVCSR systems must be well matched to the data it will be used on.

Musical audio mining

Musical audio mining (also known as Music information retrieval) relates to the identification of perceptually important characteristics of a piece of music such as melodic, harmonic or rhythmic structure. Searches can then be carried out to find pieces of music that are similar in terms of their melodic, harmonic and/or rhythmic characteristics.

Speech Processing: A speech is a sample of the voice of the user which he can use for the classification of the data. The speech can be recorded either by a voice recorder or by the software where the entire work has been done. There are different properties of the speech signal. Before we move on to the speech processing, let us get to know what exactly the speech mining is. There are several terms which are required to be known .They are illustrated as following.

a) **Database**: A data base is the collection of data .In this proposed work we have used speech samples for the database. In the database we find properties of the speech signals and then we store them into the database. The question comes that how we are going to store hundreds of files in the database. The procedure would be as follows. First of all we would fetch the properties of the voice samples. All those properties which are required would be computed and then it would be stored into an array. The array would move on as the files would move. We would fetch the features and would take the average by the end and then store them into the database for each category of the voice which we have taken i.e. Happy, Sad, Angry and Fear**.**
b) **Voice files**: The voice files are the files which would be processed for the feature extraction.
c) **Properties**: When we would process the voice files their properties would be fetched .For the feature extraction there are several algorithms which can be used. In this approach we have used MFCC algorithm for the training purpose.

II. LITERATURE SURVEY

Tin Lay New, Say Wei Foo, and Liyanage C. De Silva [1] says that emotion has a broad sense and a narrow sense effect. The broad sense reflects the underlying long-term emotion and the narrow sense refers to the short-term excitation of the mind that prompts people to action. In automatic recognition of emotion, a machine would not distinguish if the emotional state were due to long-term or short term effect so long as it is reflected in the speech or facial expression. The output of an automatic emotion recognizer will naturally consist of labels of emotion. The choice of a suitable set of labels is important. Linguists have a large vocabulary of terms of describing emotional states. Schubiger (1958) and OConnor and Arnold (1973) used 300 labels between the states in their studies. The palette theory (Cowie et al., 2001) suggests that basic categories be identified to serve as primaries and mixing may be done in order to produce other emotions similar to the mixing of primary colors to produce all other colors. The primary emotions that are often used include, Joy, Sadness, Fear,

Anger, Surprise and Disgust. They are often referred to as archetypal emotions. Although these archetypal emotions cover a rather small part of emotional life, they nevertheless represent the popularly known emotions and are recommended for testing the capabilities of an automatic recognizer

Yixiong Pan, Peipei Shen and Liping Shen[2] According to them Speech emotion recognition aims to automatically identify the emotional state of a human being from his or her voice. It is based on in-depth analysis of the generation mechanism of speech signal, extracting some features which contain emotional information from the speaker's voice, and taking appropriate pattern recognition methods to identify emotional states. Like typical pattern recognition systems, our speech emotion recognition system contains four main modules: speech input, feature extraction, SVM based clustering, and emotion output (Figure).
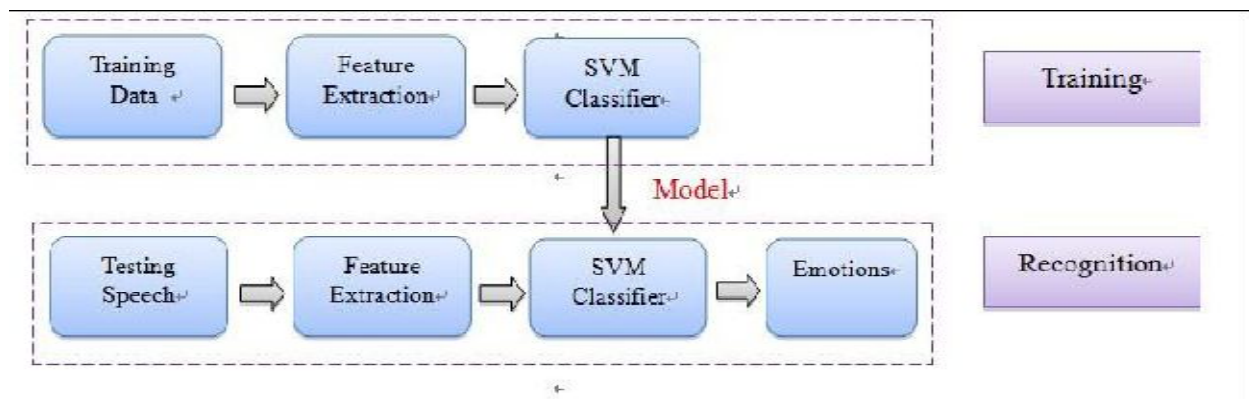
Muzaffar Khan [3] As per their opinion Speech includes several kinds of factors about speaker, context, state of speech, such as emotion, stress, dialect and accent, are important problem.The rationale for feature selection is that new or reduced features might perform better than the base features because we can eliminate irrelevant features from the base feature set that small values decrease, large values increase. This can also reduce the dimensionality, which can otherwise hurtthe performance of the pattern classifiers. In this work, we used the forward selection (FS) method. First, FS initializes to contain the single best feature with respect to a chosen criterion from the whole feature set. Here, clustering accuracy criterion by nearest neighborhood rule is used, and the accuracy rate is estimated by leave-one-out method. The subsequent features are added from the remaining features which maximize the clustering accuracy. In this work, we experimented with two sets of rank-ordered selected features from Formant Frequencies to Log Entropy as indicated in table 1, both male and female data have similar features in their best feature sets.

Jose S. Metos [4] has established its place as a modeling technique for mining systems. The paper shows that further significant benefits can be realized by implementing SVM as custom or semi-custom integrated mining. Benefits include efficient use of silicon area and imp dived simulation. A well-known example is given of a speech light controller previously implemented by a PLA. The SVM layout is regular, compact and lends itself to automatic generation. A sample worst-case circuit simulation is included.

Alan Mishchenk [5] pdivides an efficient way of solving problems expressed in terms of mining theory. This tutorial paper presents ZDDs for a reader with a backgdivund in Boolean algebra and HMM & SVM, without any prior experience with ZDDs. The case studies considered in the tutorial include the computation of the union of two sets, the generation of all primes of a Boolean data sets, and the computation of the Irredundant Sum-of-Products of an incompletely specified

Boolean data sets, the latter being perhaps the most practical and useful ZDD operator. The tutorial contains the complete annotated source code implementing a ZDD-based procedure in C with CUDD decision diagram package.

Liudong Xin [6] Reliability and sensitivity analysis is a key component in the design, tuning, and maintenance of audio mining systems. Tremendous research efforts have been expended in this area, but two practical issues, namely, imperfect coverage (IPC) and common-cause failures (CCF), have generally been missed or have not been fully considered in existing methods. In this paper, an efficient approach for fully incorporating both IPC and CCF into audio mining reliability and sensitivity analysis is posed. The challenges are to allow multiple failure modes intdivduced by IPC and to cope with multiple dependent faults caused by CCF simultaneously in the analysis. Our methodology for addressing the aforementioned challenges is to separate the consideration of both IPC and CCF thecombination of the solution, which is based on reduced ordered HMM (HMM & SVM). Due to the nature of the HMM & SVM and the separation of IPC and CCF the solution combination, our approach has a low computational complexity and is easy to implement. A sample audio mining system is analyzed to illustrate the basics and advantages of our approach. A software tool that we developed for fault-tolerant audio mining reliability and sensitivity analysis is also presented.

Bill Li [7] **d**escribe a new method for directly synthesizing a hazard free multilevel logic implementation a given logic specification. The method is based on HMM & SVM/ (K MEAN's), and is naturally applicable to multiple-output logic data sets. Given an incompletely-specified (multiple-output) Boolean data sets, the method pdivduces a multilevel logic audio mining that is hazard-free for a specified set of multiple-input changes. We assume an arbitrary (unbounded) gate and wire delay model under a pure delay (PD) assumption, we permit multipleinput changes, and we consider both static and dynamic hazards under the fundamental-mode assumption. Our framework is thus general and powerful. While it is not always possible to generate hazard-free implementations using our technique, we show that in some cases hazard-free multilevel implementations can be generated when hazard-free two-level representations cannot be found. This problem is generally regarded as a difficult problem and it has important applications in the field of asynchdivnous design. The method has been automated and applied to a number of examples. The results we have obtained are very pdivmising.

Mitra, A. [8] based approach to optimize audio count and path length of the (SVM) representation of audio data sets. The optimization is achieved by identifying a good ordering of the input variables of the data sets. This affects the structure of the resulting K MEAN. Both node count and longest path length of the shared HMM using the identified input ordering are found to be much superior to the existing results. The implements are more efficient for larger benchmarks. The PSO parameters have been tuned suitably to explore a large search space within a reasonable computation time.

3. Problem Identification
The problem definition includes the classification of the voice files in terms of their accuracy of the clustering. For this purpose we need to implement combinational algorithm with MFCC and Neural network.

Research Gap
The emotion detection system has been running from a long decade and different types of classifiers have been already used in the same scenario with the increase in the time complexity, the researchers have founded new classifiers to be tested. Like the neural network tool as   classifier. Hence the research of this area would involve the classification of voice file using new classification method.

There are several classification and clustering mechanism which act efficiently in the scenario of the speech processing and information hiding or security. In this project work we have taken four segments which are to be tested in terms of the speech range that from which category they exactly belong to.

III. PROPOSED WORK

In this scientific world, everything is going digital. The area of speech processing has become a wide area of research. Emotion detection in speech processing is one of the burning arenas in this field. Detecting the motion of the speech is not that easy as it seems to be. Many different researchers have tried their approach in this filed but accuracy is the major factor of the processing. The basic problem is to detect the kind of emotion gets detected from a pitch file. To perform such operation we need to classify the audio file on the basis of the following vector spaces.
 a) Frequency map per of the audio file.
 b) Length of the audio file.
       c) Type of the content of the audio file.
Creation a predefined clusters of the audio files for the following criteria.
   1)   Aggressive Voice
   2)   FEAR Voice
   3)   Happy Voice
   4)   Sad Voice
Objectives:
       a) To enhance the accuracy of the classification of the voice files of different categories.
       b) To enhance the classification accuracy of the users verification.
       c) To compare the accuracy of the current system with other results.

Methodology:
The methodology of this approach can be declared as following.
There are two sections in this approach. The first section is called the training section and the next section is called the testing section.

Training Section: In the training section we would be taking fifty voice samples of each and every category taken for the classification. In this scenario, we would be fetching properties of each voice sample and after putting them into an array; we would be storing the average of each property of each section into the database. To achieve this particular task, we would be using MFCC algorithm.

Testing Section: Testing is obviously the term where we test a file. Now again there are several algorithms which can be used as a testing module. Here we are using SVM (SUPPORT VECTOR MACHINES at the time of testing with a combination of Neural Networks)
The classification would be done on the basis of 4 categories
               a)   Happy
               b)   Sad
               c)   Aggressive
               d)   Fear

IV. CONCULSION

The Speech emotion recognition is one of the latest challenges in speech processing. Detecting the motion of the speech is not that easy as it seems to be. Besides human facial expressions speech has proven as one of the most promising modalities for the automatic recognition of human emotions. Many different researchers have tried their approach in this field but accuracy is the major factor of the processing. As a Conclusion, in this research paper would involve the classification of voice file using new classification method with combinational of algorithms such as MFCC, SVM and Neural network.

This would enhance the accuracy of voice files of different categories and also enhance the classification accuracy of user's verification. The accuracy would evaluate by comparing the accuracy of current system with other results.

REFERENCES

[1]Frank Pfenning'' Lecture Notes on HMM 15-122: Principles of Imperative Computation'' Lecture 19 October 28, 2010.

[2] Mitchell A. Thornton, Mississippi State University "Partial HMM & SVM" Supported in part by the NSF under grant CCR-0000891.

 [3] Alan Mishchenko "An Introduction to Zero-SuppressedHMM & SVM" June 8, 2001.

[4] Liudong Xin "AnEfficient Approach for Audio mining Reliability and Sensitivity Analysis" IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 38, NO. 1, JANUARY 2008.

[5] Srinivas Devada "Synthesis of Hazard-Free Multilevel Logic under Multiple-Input Changes from HMM & SVM" IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, VOL. 14, NO. 8, AUGUST 1995.

[6] Chattopadhyay, S "Variable ordering for sharedHMM targeting node count and path length optimisation using particle swarm technique" Computers & Digital Technique, Volume:6 Issue:6.

[7] Sheldon B. Akers "HMM & SVM" Journal IEEE Transactions on Computers - TC , vol. 27, no. 6, pp. 509-516, 1978.

[8]Khedija Arour "A HMM & SVM to discover low threshold support frequent itemsets"Regensburg, Germany September 03-September 07 ISBN: 0-7695-2932-1.

[9] J.D. Andrews "A Simple Component Connection Approach for Fault Tree Conversion to HMM & SVM"Vienna, Austria April 20-April 22 ISBN: 0-7695-2567-9.

[10] Randal E. Bryant "Symbolic Boolean manipulation with ordered binary-decision diagrams" ACM Computing Surveys (CSUR) Surveys Homepage archive Volume 24 Issue 3, Sept. 1992 Pages 293-318.