Comparative Analysis of Machine Learning Algorithms for Audio Signal Classification

Poonam Mahana[1], Gurbhej Singh[2]

[1, 2]*Department of Computer Science & Information Technology, Rayat Bahra institute of Engineering & Bio-Technology, Mohali Campus*

Abstract---Research in the area of Audio Classification and retrieval, in comparison with closely related areas, such as speech recognition and speaker identification is relatively new. Audio Classification is an important issue in current audio processing and content analysis researches.. Generally speaking, audio classification is a pattern recognition problem. Pattern recognition is the scientific discipline whose goal is the classification of objects into a number of categories or classes. Depending on the application, these objects can be images or signal waveform or any type of measurements that need to be classified. The goal of pattern recognition is to classify objects into a number of categories. The word pattern refers to the type of measurements that need to be categorized or classified. The measurement can be just about anything but typical examples are images and acoustic signals. The ongoing advancements in multimedia technologies drive the need for efficient classification of audio signals. This paper provides an improved audio classification and categorization technique using two ML algorithm.

Keywords- *Pattern Recognition, Audio Classification, Support Vector Machine and k-NN, Zero Crossing Rate, Short Time Energy, Spectral Flux and Spectral Centroid., Audio Signal Classification, SVM, Pre- processing, Pattern Recognition, Feature Extraction, Feature Selection, Sampling frequency, Frame forming and Pre-emphasized filter.*

I. INTRODUCTION

A. Motivation

The ongoing advancements in the multimedia technologies drive the need for efficient classification of the audio signals to make the content-based retrieval process more accurate and much easier from huge databases. Audio information often plays an important role in understanding the semantic content of multimedia. With this rapidly increasing amount of data, users require automatic methods to filter process and store incoming data. Audio data can be used as an important source of information that yields effective results for video indexing and content analysis. A human listener can easily distinguish between different audio types by just listening to a short segment of an audio signal but problem does arise when the sound is weak or there is noise. However, solving this problem using computers has proven to be very difficult. A major challenge in this field is the automatic classification of audio signals. Audio signal classification is a growing area of interest applicable to media services, search engines and intelligent human-computer systems.

Audio classification is important due to the following reasons:

a. Different audio types should be processed differently.

b. The searching space after classification is significantly reduced to a particular subclass during the retrieval process. Each classified audio piece will be individually processed and indexed to be suitable for efficient comparison and retrieval.

B. Audio Signals Classification

The process of Audio Signal Classification involves the extraction of features from sound and the use of these features to identify the class it belongs to. Generally speaking, audio classification is a pattern recognition problem. Pattern recognition is the scientific discipline whose goal is the classification of objects into a number of categories or classes. Depending on the application, these objects can be images or signal waveform or any type of measurements that need to be classified. The goal of pattern recognition is to classify objects into a number of categories. Pattern recognition develops and applies algorithms that recognize patterns in data. These techniques have important applications in character recognition, speech analysis, image analysis, clinical diagnostics, person identification, machine diagnostics, and industrial process supervision. Hence the fundamental objective for pattern recognition is classification. A pattern recognition system can be considered as a two stage device. The first is feature extraction and second is classification.

C. Applications of Audio Signals Classification

The main applications of audio signals classification include the following:

(i) *Computing Tools:* Audio signal classification can be used as a front-end for a number of currently existing computer applications to audio. Speech recognition is one of the more obvious applications of ASC, where signals are classified into phonemes and then assembled into words. ASC can be used to improve on current speech recognition technology in many ways.

(ii) *Automatic Bandwidth Allocation:* A telephone network with audio classification capabilities could dynamically allocate bandwidth for the signal being transmitted. More bandwidth would be allocated for music than for speech transmissions and no bandwidth at all if only background noise is detected. This would help multiplexing systems to work more efficiently. The same applies to audio streams in data networks such as the Internet.

(iii) *Audio Database Indexing:* This is especially useful for large audio and music collections, such as the audio archives in broadcasting facilities, sound track studios for the movie industry or in music content providers on the Internet. Currently, this classification is done manually, which is an extremely time and human resource consuming task. Perhaps the most common example illustrating this case are the large collections of downloaded MP3 files stored on hard disks, very often in a chaotic manner.

II. MACHINE LEARNING ALGORITHMS

A. *Machine Learning Algorithms*

Machine Learning is a technology for mining knowledge from data. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on that data. It is a scientific discipline that is concerned with the design and development of to evolve behaviors based on data.

Supervised Learning: Supervised Learning is the type of learning that takes place when the training instances are labeled 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.

Unsupervised Learning: In unsupervised learning, there is no any desired output, so no error signal is generated. It refers to the problem of trying to find hidden structure in unlabeled data. Here, the input vectors of similar types are grouped together during training phase.

Reinforcement Learning: Reinforcement learning allows the machine to learn its behavior based on feedback from the environment. This behavior can be learnt once and for all, or keep on adapting as time goes by. This automated learning scheme implies that there is little need for a supervisor who knows about the domain of application.

The ML algorithms to be used in this research work are SVM and k-Nearest Neighbor which are explained in the following section. These ML algorithms were chosen because they were found to perform well for pattern classification tasks as compared to the other classifiers.

1. *Support Vector Machine (SVM)*

Support Vector Machine is a powerful classifier arising from Structural Risk Minimization theory that has proven to be efficient for various classification tasks including speaker identification, text categorization and musical instrument recognition. The Support Vector Machine (SVM) is a binary classifier, separating two classes by an optimal hyper-plane of equation $w \cdot x + b = 0 , \omega \in R^N$ and $b \in R$ with the largest margin.Each optimal hyper-plane is obtained from positive and negative examples in the training set. There may be so many hyper-planes that might classify the data, but the one that maximizes the margin is the optimal one, due to the larger margin the lower the generalization error of the classifier. Consider the example in Fig.2.1.

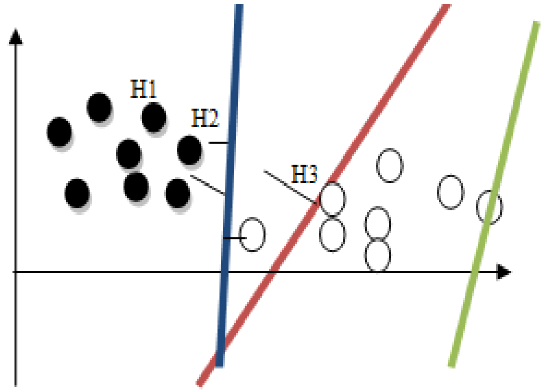

Fig.2.1: Optimal separating hyper plane

2. *The k nearest Neighbor Classifier*

In the k nearest neighbor algorithm (k-NN) is a method for objects based on closest training examples in the. k-NN is a type of where the function is only approximated locally and all computation is deferred until classification. The k nearest neighbor algorithm is amongst the simplest of all algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples.

Fig.2.2 shows the architecture of k-nearest neighbor in which each of the samples have been labeled either A, B or C except for the sample x. This needs to be labeled, the k nearest neighbor classifier takes the k nearest samples, i.e. the closest neighbors around the sample x and uses them to assign a label. This is usually done by a majority-voting rule, which states that the label assigned should be the one, which occurs most among the neighbors. The aim is to use the k-NN classifier for finding the class of an unknown feature **x**. As it can be seen in the figure, of the closest neighbors (k = 5 neighbors) four belong to *class a* and only one belongs to *class b* and hence **x** is assigned to *class a*.

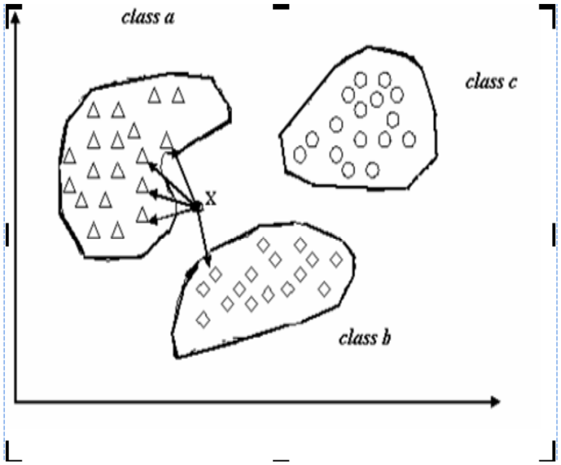

Fig.2.2: The k nearest neighborhood rule (k = 5)

B. Classification

Classification is a data mining function that assigns items in a collection to a target categories or classes. Classification is a supervised learning in which individual item of data set is categorized to different groups based on prior knowledge. The characteristics of data plays the important role in the performance of classifier depends. Classification is one of the most frequently studied problems by Data Mining and machine learning (ML) researchers. Classification derives a function or model which determines the class of an object based on its attributes. A set of objects is given as the training set. A classification function or model is constructed by analyzing the relationship between the attributes and the classes of the objects in the training set. This function or model can then classify future objects. This helps us develop a better understanding of the classes of the objects in the database.

1. Estimation of Classifiers Performance

Analysis of how well a specific classification scheme work is important to get some expectation of the classifiers accuracy. When a classifier is designed, there are a variety of methods to estimate how high accuracy it will have on future data. These methods allow comparison to other classifiers performance. A method described most often is cross-validation. Cross-validation is used to maximize the generality of estimated error rates. The error rates are estimated by dividing a labelled data set into two parts. One part is used as training set and the other as a validation set or testing set. The training set is used to train the classifier and the testing set is used to evaluate the classifiers performance.

2. The Design Process of a Pattern Classifier

The aim of pattern recognition is the design of a classifier, a mechanism which takes features of objects as its input and which results in a classification, a label or value indicating to which class the object belongs.

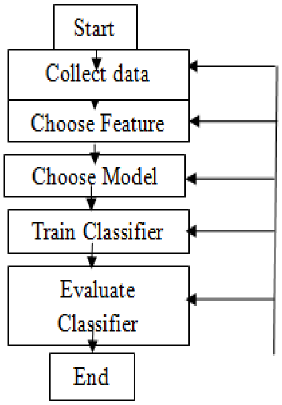

Fig.2.3: Design cycle of a classifier

III. LITERATURE SURVEY

[1] This paper describes a real time speech / music discriminator to be used in radio receivers for the automatic monitoring of the audio content of FM radio channels. The computational simplicity of the approach could lend itself to wide application including the ability to automatically change channels when commercials appear. In automatic speech recognition of broadcast news, it is important to disable the speech recognizer during the non speech portion of the audio stream. Automatic classification would remove the subjectivity inherent in the classification process and ultimately speed up the retrieval process. The algorithm provides the capability to robustly distinguish the two classes and runs easily in real time. A major contribution of this work is the introduction of a new feature extraction method and feature representation for discriminating speech from music on broadcast FM radio.

[2] In this paper they we present a new video shot classification and clustering technique to support content-based indexing, browsing and retrieval in video databases for speaker identification and video analysis for hierarchical video shot classification. The proposed method is based on the analysis of both the audio and visual data tracks. The clues obtained from the video and speech data are combined to classify and group the isolated video shots. This integrated approach also allows effective indexing of the audio-visual objects in multimedia databases.

[3] This paper describes audio and visual features for multimedia content analysis that can effectively characterize scene content for video archiving and retrieval. Multimedia content analysis refers to the computerized understanding of the semantic meanings of a multimedia document, such as a video sequence with an accompanying audio track. For this purpose an appropriate feature name ZCR (Zero Crossing

Rate) is used. The experiments show that promising results by using temporal features.

[4] In this paper the performance of the support vector machine (SVM) for a speaker verification task is assessed. Since speaker verification requires binary decisions, support vector machines seem to be a promising candidate to perform the task. A new technique for normalizing the polynomial kernel is developed and used to achieve performance comparable to other classifier on the YOHO database, which provides high quality recordings of speech from co-operative speakers. The normalization scheme is applicable to the polynomial kernel and imposes upon it some properties of the RBF kernel. With the new normalization technique, we were able to achieve excellent performance from SVMs on this database.

[5] This paper describes an approach to automatic segmentation and classification of audiovisual data based on audio content analysis is proposed. The audio signal from movies or TV programs is segmented and classified into basic types. Four kinds of audio features including the energy function, the average zero-crossing rate, the fundamental frequency, and the spectral peak tracks are extracted to ensure the feasibility of real-time processing. Based on audio feature analysis, a procedure for online segmentation and classification of the accompanying audio signal in audiovisual data into twelve basic audio types is accomplished. Then, analysis results of the audio information will be integrated into those of the visual information and the caption in video programs so that a fully functional system for video content parsing can be achieved.

[6] This paper presents a work on classifying the sound track of instructional videos into seven distinct audio classes using the Support Vector Machine (SVM) technology. The classification results are then used to partition a video into homogeneous audio segments, which forms the fundamental basis for its higher-level content analysis and exploration. This classifier is well suitable due to its capability in handling complicated feature space. The method proposed in this paper leads to better results than existing audio classification methods.

IV. PRESENT WORK

A. Problem Statement

From the literature survey it was found that prosodic features are efficient to characterize audio signal's, but none of the paper was found till date, using these prosodic features as a whole, for audio signal's classification task. Further it was found that ML algorithms overcome the drawbacks of traditional classifiers. An improvement in classification accuracy is expected by using these features along with ML algorithms. Thus the main task to be considered in this thesis is the performance analysis of two ML algorithms: SVM and k nearest neighbor using prosodic feature set in the classification task for determining the following:

i. Classification Accuracy
ii. Time taken by the ML algorithms in audio signal's classification task

Performance of these algorithms will be evaluated based on accuracy rate. To the best of our knowledge, we have not found any literature which compares these ML algorithms against these parameters.

B. Methodology

Audio signal classification system comprises of two distinct blocks, a feature extraction and a classification. Extracting relevant features from a sound, and using these features to identify which set of classes the sound is most likely to fit. Feature design is said to be the domain dependent part of classifier design, which means it requires an in-depth knowledge of the specific field of application. On the other hand, classification algorithms regard input data as a set of numbers, without any knowledge about their true nature. The feature extraction and grouping algorithms used can be quite diverse depending on the classification domain of the application. After that, feature vector is input to the ML algorithm for training and classification. The process to be followed is shown in Figure 4.1.

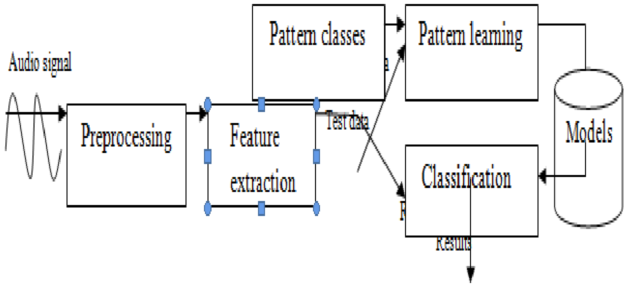

Figure 4.1 Block diagram of proposed audio classification system

1. Description of the audio data

The audio files used in the experiment were randomly collected from the internet and from the audio database. These files were in different formats (MP3) and in order to have a common format for all the audio files and to be able to use them in Matlab programs, it is necessary to convert these files to a wav format with a common sampling frequency. The recorded audio files were further partitioned into two parts: the training set and the test set. This was important since each audio file was intended to be used only once, either for training or for testing a classifier.

2. *Audio Preprocessing*

In MATLAB the following steps are implemented.

3. *Frame forming*

For feature extraction it is assumed that only short-term audio streams are present. For classifying longer audio streams, segmentation is necessary. Each of these short clips is divided into frames. Each frame contains the same number of samples.

4. *Pre emphasized filter*

Due to radiation effects of the sound from lips, high frequency components have relatively low amplitude, which will influence the capture of the features at the high end of the spectrum. One simple solution is to augment the energy of the high-frequency spectrum. This procedure is implemented via a pre-emphasizing filter that is defined as:

$$s'_n = (s_n) - (0.96 \times s_{n-1}) \quad for\ n=0,....,511 \quad (4.1)$$

where s_n is the nth sample of the frame S. Then the frame is hamming windowed by:

$$s_i^h = s'_i * h_i \qquad for\ n= 0,....,511 \quad (4.2)$$

where h_i is Hamming window given by:

$$h_i=(0.54)-\left(0.46\times cos\left(\frac{2\pi i}{511}\right)\right) \quad for\ i= 0,.,511\ (4.3)$$

5. *Feature extraction*

A Feature is any extractable measurement taken on the input pattern that is to classified. The aim of feature extraction is to represent audio data in a compact and descriptive manner such that it is efficient to deal with when applying classification algorithms. The first step in any classification problem is to identify the features that are to be used for classification. Independent of which classifier is used, the choice of feature set plays a key role in classification performance. Audio signal's features such as zero crossing rate, their short time energy, root mean square, spectral flux and spectral centroid were extracted from the audio samples. Feature extraction is the process of converting an audio signal into a sequence of feature vectors carrying characteristic information about the signal. These vectors are used as basis for various types of audio analysis algorithms. The design of descriptive features for a specific application is hence the main challenge in building audio classification systems.

6. *Performance Analysis*

Performance analysis of ML algorithms will be done on the basis of the performance metrics: accuracy rate and time taken for classification which is explained in the Section 4.3.

C. *Performance Metrics*

The following performance metrics are considered in analyzing the performance of ML algorithms.

Classification Accuracy Rate: The classification accuracy, which is defined as the ratio of correctly classified audio samples over the total number of audio files. This performance metric gives a measure of the overall accuracy rate of the classifier.

$$Accuracy\ Rate = \frac{number\ of\ correctly\ classified\ instances}{total\ number\ of\ instances} \times 100$$

Time Taken for Classification: It is another important performance parameter which is taken for the evaluation of classification using two different ML algorithms. Results and discussion on the basis of these metrics have been done in the next chapter.

V. RESULT AND DISCUSSION

A. *Work Done*

The aim of feature extraction is to represent audio data in a compact and descriptive manner such that it is efficient to deal with when applying classification algorithms. The first step in any classification problem is to identify the features that are to be used for classification. Independent of which classifier is used, the choice of feature set play key role in classification performance. The feature extraction begins with frame blocking followed by windowing to minimize edge-effects (spectral leakage) problem. Feature extraction is the process of converting an audio signal into a sequence of feature vectors carrying characteristic information about the signal. These vectors are used as basis for various types of audio analysis algorithms.

1. *Classification Algorithms*

After the feature extraction process, it is important to classify the signals. Classification is the process by which a particular label is assigned to a particular audio format. A classifier defines decision boundaries in the feature space, which separate different sample classes from each other. In this study, we make use of two classification methods, namely k-NN and SVM. SVM technique is chosen as employed pattern classifier due to better performance it showed in sound classification over other classification methods. The Support Vector Machine is a classifier that finds a maximal margin separating hyper plane between different classes of data. In order to compare the performance of both classifiers fairly, the same training set and the same test set were used for both the SVM and the k-NN classifiers on audio classification task. Classification of audio signals performed by two ML algorithms on audio dataset are shown in figures 5.1.1(a) and 5.1.1(b).

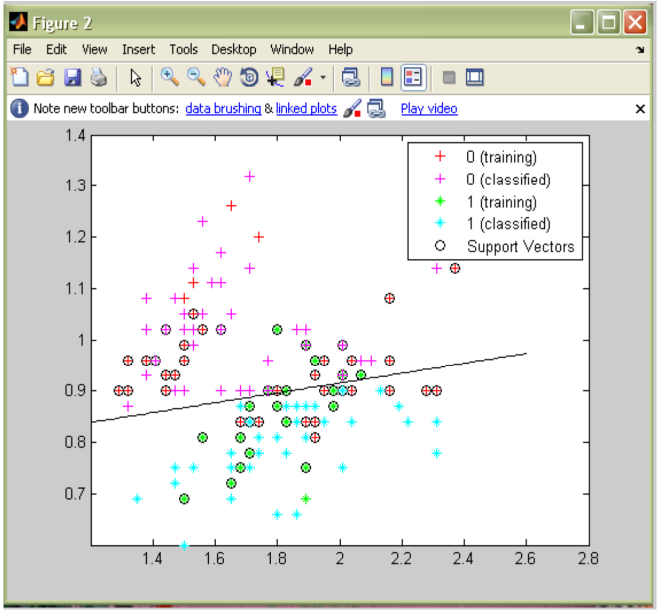

Fig.5.1.1 (a): Scatter Plot of SVM for classifying speech, music and noise

Fig.5.1.1 (b) contains the details of the performance of k nearest neighbor algorithm for audio signal's classification task.

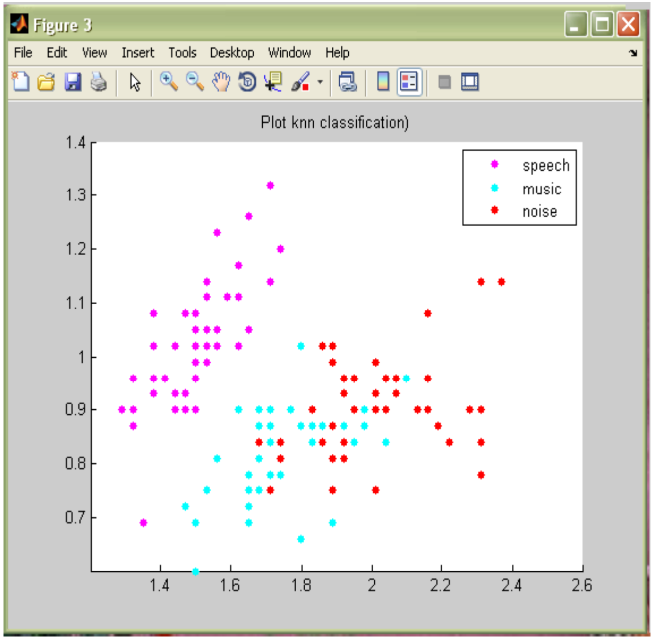

Fig.5.1.1 (b): Scatter Plot of k-NN for classifying speech, music and noise

2. *Classification Performance*

For audio signal's classification task, SVM is the best as it gives highest accuracy rate. Best performance of SVM can be attributed to the inherent nature of the algorithm. The SVM is a supervised classification system that minimizes an upper bound on its expected error. SVM attempts to find the hyper plane separating different classes of data that will generalize best to future data. Such a hyper plane is the so called maximum margin hyper plane, which maximizes the distance to the closest point from each class. k-NN performed least in comparison to the other ML algorithm. The results of the experiments performed on audio signals classification task using Support Vector Machine (SVM) and k Nearest Neighbor (k-NN) classifiers are shown in following table. Complexity wise, k-NN takes extremely long time to classify audio signal's whereas SVM takes only few seconds.

TABLE 5.1 COMPARATIVE ACCURACY OF MACHINE LEARNING ALGORITHMS

CLASSIFIERS	ACCURACY (%)
k-NN	74.6%
SVM	90%

TABLE 5.2 SPEED PERFORMANCE OF MACHINE LEARNING ALGORITHMS

CLASSIFIERS	TIME (s)
k-NN	5
SVM	2

Fig.5.1.2 (a) contains the details of performance of Machine Learning Algorithms for Audio Signal's Classification task in terms of Classification Accuracy Rate.

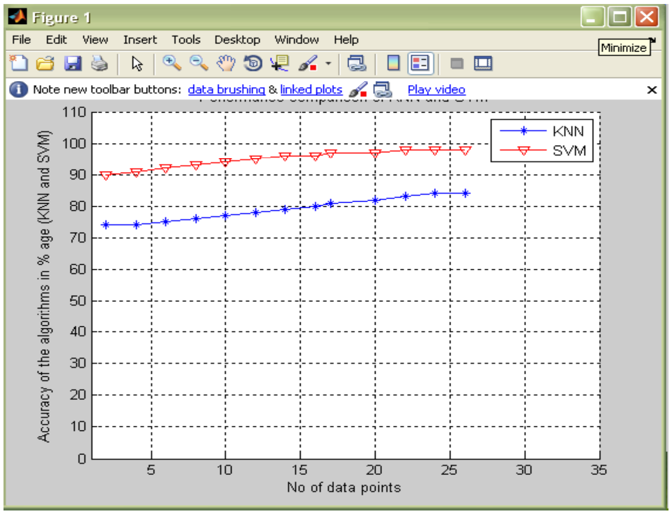

Fig.5.1.2(a): Performance of ML algorithms in terms of Accuracy rate

VI. CONCLUSION & FUTURE SCOPE

A. Conclusions

In this research, we have evaluated and compared the performance of two ML algorithms: SVM and k-NN using prosodic features. SVM with radial basis function kernel performed well, while requiring significantly less training data to achieve such an outcome in comparison to other ML algorithms. In future, the ML algorithm can be trained and tested on samples with degraded acoustical conditions. This is due to the growing demands of the tremendous distribution of telecommunication network, it becomes an important prerequisite for efficient coding and audio signal's enhancement.

B. Future Scope

The motivation for the future work in audio signal's classification can come from the fact that if the different feature sets provide uncorrelated information, they can be combined in order to give a joint decision. Further, the ML algorithm can be trained and tested on samples with degraded and mismatched acoustical conditions. This is due to the growing demands of the tremendous distribution of telecommunication network, it becomes an important prerequisite for efficient coding and audio signal enhancement. It would be interesting to see what features are most robust. To achieve goal, we need to explore more audio features that can be used to characterize the audio system and our audio classification scheme will upgraded to discriminate more audio classes, which contain important semantic information and therefore increase the accuracy and reliability of the audio classification system.

VII. REFERENCES

[1] J. Saunders, "Real-Time Discrimination of Broadcast Speech/Music", IEEE International Conference Proceedings on Acoustics, Speech and Signal Processing, vol.2, 1996, pp. 993 – 996.

[2] Jeho Nam, A. Enis Cetini and Ahmed H. Tewjik, "Speaker identification and video analysis for hierarchical video shot classification", IEEE International Conference on Image Processing, vol.2, 1997, pp. 550 – 553.

[3] Yao Wang, Zhu Liu and Jin-Cheng Huang, "Multimedia Content Analysis using Both Audio and Visual Clues", IEEE Transactions on Signal Processing, vol.17, 2000, pp. 12 – 36.

[4] V. Wan and W. M. Campbell, "Support Vector Machines for Speaker Verification and Identification", IEEE International Conference on Neural Networks, vol.2, 2000, pp. 775 – 784.

[5] T. Zhang and C. J. Kuo, "Audio Content Analysis for Online Audio-Visual data Segmentation and Classification", IEEE Transactions on Speech Audio Processing, vol.9, May 2001, pp. 441 – 457.

[6] Ying Li and Chitra Dorai, "SVM-Based Audio Classification for Instructional Video Analysis", IEEE International Conference

[7] Bai Liang, Hu Yanli, Lao Songyang, Chen Jianyun, Wu Lingda, "Feature Analysis and Extraction for Audio Automatic Classification", IEEE International Conference on Systems, Man and Cyber metrics, vol.1, 2005, pp. 767 – 772.

[8] Naoki Nitanda, Miki Haseyama and Hideo Kitajima, "Accurate Audio Segment Classification using Feature Extraction Matrix", IEEE International Conference on Acoustics, Speech and Signal Processing, vol.3, 2005, pp. 261 – 264.

[9] Ahmad R. Abu-El-Quran, Student Member, Rafik A. Goubran and Adrian D. C. Chan, "Security Monitoring Using Microphone Arrays and Audio Classification", IEEE Transaction on Instrumentation and Measurement,vol.55, 2006, pp. 1025 – 1032.

[10] S. Ravindran and D. V. Anderson, "Audio classification and scene recognition and for hearing aids", IEEE International Conference on Circuits and Systems, vol.2, May 2005, pp. 860 – 86.

[11] Tin Lay and LI Haizhou, "Broadcast News Segmentation by Audio Type Analysis", IEEE International Conference on Acoustics, Speech and Signal Processing, vol.2, 2005, pp. 1065 – 1068.

[12] Jia Ching Wang, Fa Jhing Wang, He Wai Kuok, and Hsu Shu Cheng, "Environmental Sound Classification Using Hybrid SVM/k-NN Classifier and MPEG-7 Audio Low-Level Descriptor", IEEE International Conference on Neural Networks, 2006, pp. 1731 – 1735.

[13] Jia Ching Wang, Fa Jhing Wang, Cai Bei, Kun-Ting Jian and Wai-He Kuok, "Content-Based Audio Classification Using Support Vector Machines and Independent Component Analysis", Proceedings of the 18th IEEE International Conference on Pattern Recognition, vol.4, 2006, pp. 157 – 160.

[14] Ying Li and Chitra Dora, "Instructional Video Content Analysis Using Audio Information", IEEE International Transaction on Audio, Speech and Language Processing, vol.14, 2006, pp. 2264 – 2274.

[15] Yingying Zhu, Zhong Ming, and Qiang Huang, "Automatic Audio Genre Classification Based on Support Vector Machine", IEEE International Conference on Natural Computation, vol.1, 2007, pp. 517 – 521.