Efficient Audio Retrieval using Supervised Learning Techniques

Kesavan Namboothiri T.
Master of Engineering in Applied Electronics
Sri Venkateswara College of Engineering
Sriperumbudur-602105

Anju L.
Assistant Professor, Department of ECE
Sri Venkateswara College of Engineering
Sriperumbudur-602105

Abstract: **Audio retrieval has been an effortful task for a long time. There are a lot of techniques available for content based audio retrieval. Here, audio retrieval is done in a supervised learning medium. Audio features are extracted using midterm and short term feature extraction method. Extracted features are classified using k-Nearest Neighbor (kNN) algorithms. Audio indexing is done using Support Vector Machine (SVM), a supervised learning technique. Finally audio matching is done using Dynamic Time Warping (DTW) technique.**

Keywords: Audio retrieval, Supervised learning, audio matching

I. INTRODUCTION

In this fast growing computer era, human beings are very much fascinated about their perception on hearing and how to inspire those skills in to machines using artificial intelligence, machine learning and neural networks. There are so many existing techniques available in audio retrieval, such as

1. Fuzzy similarity calculation based on spectrum histograms and fluctuation patterns.
2. Divide the audio into several homogenous segments, and train an ensemble classifier to provide the audio annotation.
3. Two stage speeds up retrieval process, which consisted of coarse search based on the histogram pruning algorithm and retrieval based on time information.
4. Chroma-based features that strongly correlated to the harmonic progression of the audio.
5. Method of structured representation of audio features that are helpful for content-based audio retrieval.

Here, audio retrieval is done based on a supervised learning technique. A short query of audio clip is given and the objective is to retrieve all similar audio clips from the database. The major problem here is to measure the similarity between audio clips. This is specifically difficult for waveform inputs since the computers can't process the audio inputs directly. Thus for processing audio data in the database, it is converted in to useful audio features. Another problem is features extracted from audio input are not that much user friendly and this problem is usually addressed as "Bridging the Semantic Gap". Finally retrieval process is time consuming and it also affects the performance of the system.

In this paper first audio features are extracted using short-term and mid-term feature extraction. Audio feature classification is done using kNN classification algorithm. Finally audio indexing and matching is introduced to speed up the retrieval process. The classified features are indexed using Support Vector Machine (SVM), a supervised learning algorithm. Audio matching is done using Dynamic Time Warping (DTW) algorithm.

This paper is organized as follows: Section II deals with feature extraction using short term and mid-term feature extraction procedure. Section III illuminates about audio classification. Section IV tells about audio indexing and its implementation methods. Section V introduces audio matching. Section VI conducts the experiment and evaluates the performance of the proposed system. Section VII concludes the scope and future work.

II. FEATURE EXTRACTION

In order to extract the features, mainly two feature extraction methods are used- Short term and Mid-term feature extraction.

A. Short-term feature extraction

The audio input is divided in to possibly overlapping short term frames, and audio features are extracted from each frame. This type of processing generates a sequence, F, of feature vectors per audio signal. The feature vector dimensionality depends on the nature of used feature vector. Single dimensional and multidimensional features are used for sophisticated audio analysis. Here 23 audio features are extracted for audio analysis. Thus the audio input is broken into short term frames and 23 features are extracted.

B. Mid-term feature extraction

In this method audio features are extracted on a mid-term basis. First audio data is divided into mid-term windows (frames), then for each window, short-term features are extracted. The feature sequence F, which has been extracted using mid-term window, is used for computing the feature statistics e.g. the average value of energy. In the end, each mid-term segment is represented by a set of statistics which corresponds to the respective short-term feature sequences. At the time of mid-term feature extraction, it is assumed that mid-term windows are homogenous with respect to audio type and thus can proceed the extraction of statistics on the mid-term frame basis.

The output of the mid-term feature extraction function will give the 23 features with two mid-term feature statistics. That is 23 features are selected on short-term basis and 2 feature statistics are calculated per feature (e.g. mean and standard deviation). Then the output of mid-term feature

extraction is a 46-dimensional vector. The order of the vector is as follows, element 1 and 24 is mean and standard deviation of the first short-term feature sequence, elements 2 and 25 are mean and standard deviation of second feature sequence and so on. The main features here used are Energy, Entropy, Zero Crossing Rate, Spectral Flux, Spread, MFCCs, and Spectral Roll off, etc.

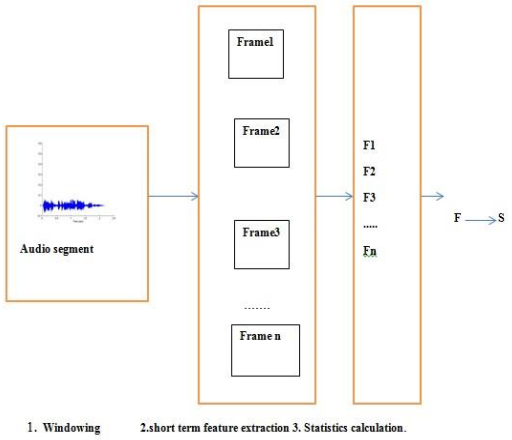

Fig 1: Mid-term feature extraction

III. AUDIO CLASSIFICATION

Classification is done using k-Nearest Neighbor algorithm. It is the one of the simplest classification algorithm in machine learning and data mining. It's well suited for both binary class and multi-class problems. Its best characteristics are that it does not require a training stage in the strict sense. The main idea behind this classifier is that if given a test pattern (unknown feature vector) , x ,it first detects k-nearest neighbors in the training set and count how many of those belong to each class. In the end, the feature vector is assigned to the class which has accumulated the highest number of neighbors. Two types of audio samples are used here. Speech samples are taken from carnage Mellon University. Music samples are taken from MARSYAS (Music Analysis Retrieval and Synthesis for Audio Signals). Here classification is done in supervised learning medium. So these samples are put in specific folders and classification is carried out.

IV. AUDIO INDEXING

Audio indexing is done using a support vector machine (SVM), a supervised learning algorithm. SVM does classification on labelled data. So data are already labelled using kNN algorithm. In SVM with the existing optimal decision hyperplane, parallel supporting hyperplanes are used in order redefine the margin that will reduce the classification error. The main SVM parameters are Type of kernel, Kernel properties, Constraint parameter, The support vector machine adopts the following steps:

- Loads the dataset of feature vectors
- Randomly splits the dataset into two equally sized subsets, one for testing and one for training .Validation techniques are then employed in order to achieve more reliable performance measurements. The random sub-sampling step has to be repeated several times.
- Finally, it computes the classification accuracy, i.e. the fraction of the correctly classified samples

A. Performance measures:

Confusion matrix is an important tool for analyzing the performance of binary class methods. This provides the means to group the classification results into a single matrix and helps to understand the types of errors that occur during the testing and training stage. The confusion matrix, CM, is a Nc x Nc matrix, whose rows and columns refer to the true and predicted class labels of the dataset, respectively. Each element, CM (i, j), stands for the number of samples of class i that were assigned to class j by the adopted classification method. It follows that the diagonal of the confusion matrix captures the correct classification decisions (i=j).

It is useful to normalize the confusion matrix so that its elements become probabilities and not simple counts of events. This can be basically achieved in two ways, the first of which is to divide each element of CM by the total number of samples in the dataset, i.e. by the sum of the elements of the confusion matrix as is given by the equation

$$CMn(i,j) = \frac{CM(i,j)}{\sum_{m=1}^{Nc}\sum_{n=1}^{Nc} CM(m,n)} \ldots (4.1)$$

Based on the standard version of the confusion matrix (before normalization takes place), it is possible to extract three useful performance measures, the first of which is the overall accuracy, Acc, of the classifier, which is defined as the fraction of samples of the dataset that have been correctly classified. It can easily be seen that the overall accuracy can be computed by dividing the sum of the diagonal elements (number of correctly classified samples) by the total sum of the elements of the matrix (total number of samples in the dataset). The quantity 1-Acc is the overall classification error. Acc is given by the equation

$$Acc = \frac{\sum_{m=1}^{Nc} CM(m,m)}{\sum_{m=1}^{Nc}\sum_{n=1}^{Nc} CM(m,n)} \ldots (4.2)$$

Apart from the overall accuracy, which characterizes the classifier as a whole, there also exist two class-specific measures that describe how well the classification algorithm performs on each class. The first of these measures is the class recall, Re (i) This is defined as the proportion of data with true class label I that were correctly assigned to class i. Re(i) is given by the equation

$$Re(i) = \frac{CM(i,i)}{\sum_{m=1}^{Nc} CM(i,m)} \ldots (4.3)$$

$where \sum_{m=1}^{Nc} CM\ (i,m)$ is the total number of samples that are known to belong to class i. If the confusion matrix has been row wise normalized, then $\sum_{m=1}^{Nc} CM\ (i,m)$=1, and as a result,Re(i) =CM (i , i), which means that the diagonal elements of the matrix already contain the respective recall values.

The second class-specific performance measure is class precision, Pr(i), which is defined as the fraction of samples that were correctly classified to class i if the total number of samples that were classified to that class. Class precision is, therefore, a measure of accuracy on a class basis and is defined according to the equation

$$Pr(i) = \frac{CM(i,i)}{\sum_{m=1}^{Nc} CM(m,i)} \ldots. (4.4)$$

Where $\sum_{m=1}^{Nc} CM\,(m,i)$ stands for the total number of samples that were classified to class i.
Finally, a widely used performance measure that combines the values of precision and recall is the F1-measure, which is computed as the harmonic mean of the precision and recall values.F1 is given by the equation

$$F1(i) = \frac{2Re(i)\mathrm{Pr}(i)}{\mathrm{Pr}(i)+Re(i)} \ldots. (4.5)$$

B. Cross-validation
Here mainly two cross-validation methods are used:

1. Repeated hold-out validation: to avoid overfitting, the hold-out method partitions the dataset into two non-overlapping subsets: one for training and the other for testing.

2. Leave –one-out validation: The leave-one-out method is actually a variation of the k-fold cross validation approach, where k = M, i.e. the number of folds is equal to the total number of samples available in the set. In other words, each fold consists of a single sample. Therefore, during each iteration, all the samples, apart from one, are used for training the classifier and the remaining sample is used in the testing stage. The leave-one-out method is an exhaustive validation technique that can produce very reliable validation results.

V. AUDIO MATCHING

Audio matching is a method used to find the similarity between audio samples. DTW method is used for audio matching

Dynamic Time Warping (DTW)
In time series analysis dynamic time warping (DTW) is an algorithm for measuring similarity between two temporal sequences which may vary in time or speed. This algorithm is mainly used in automatic speech recognition and online signature recognition. In general, DTW is a method that calculates an optimal match between two given sequences e.g. time series with certain restrictions. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This sequence alignment method is often used in time series classification.

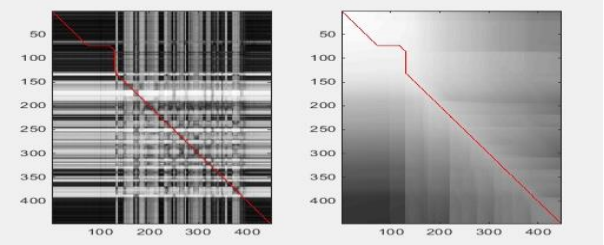

Fig 2: Audio matching using DTW

VI. EXPERIMENT RESULTS

A. Dataset
A Personal Digital Assistant (PDA) based speech database with 940 audio samples from Carnegie Mellon University (CMU) dedicated for speech research technology, is used for audio analysis.
Music audio samples collected from Music Analysis Retrieval and Synthesis for Audio Signals (MARSYAS), an authorized music speech database website.

B .Feature extraction
Features are extracted from audio files using short-term and mid-term feature extraction methods. After calculating mid-term feature extraction, statistical measurement of features is done and plotted.

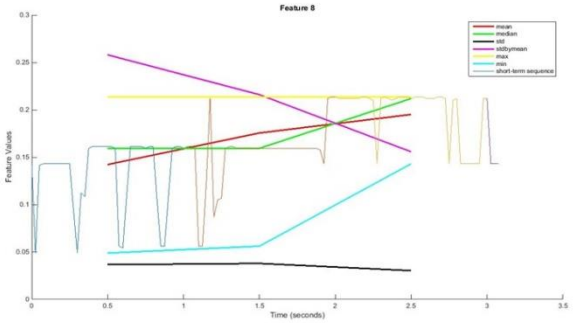

Fig 3: Mid-Term Feature extraction with order statistic

C. Audio classification
Using the extracted audio features, they are classified in to classes. Here for simplicity two classes are chosen for speech and music which is shown in table

Name	Description	Classes	Samples per classes
Modelsm	Speech Vs music	2	201

Table 1: speech vs music class

D. Audio Labelling
Audio labelling is done using support vector machine (SVM) classifier. Support Vector Machine algorithm is applied to the classified features and the labeled features are obtained. The labelling is done with different values of cost function C to understand how the cost function will affect the classification.

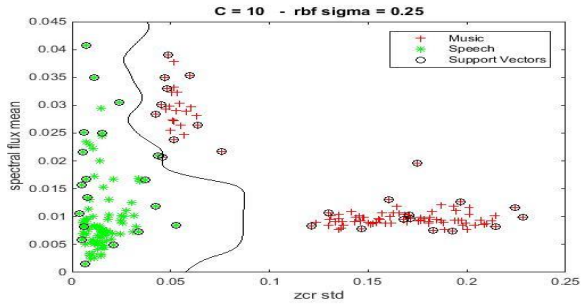

Fig 4: SVM labelling with cost function =10

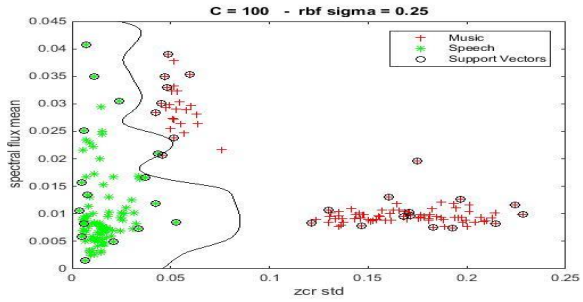

Fig 5 : SVM labelling with cost function C=100

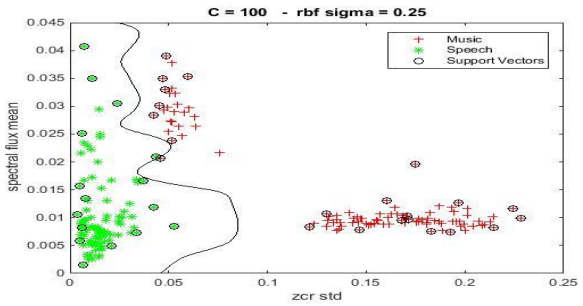

Fig 6: SVM labelling with cost function C=100000

From the plot, it is clear that as C is increased, overfitting problem appears. It can be seen that, for very high C values, the resulting classification scheme manages to classify correctly almost every sample of the training set. However, the classification accuracy on the test set decreases considerably.

Fig7 shows the classification accuracy of the training dataset on different values of C. Overfitting occurs at high values. In those cases, the resulting SVM classifies correctly almost all training samples but fails on a large percentage of the samples of the testing set.

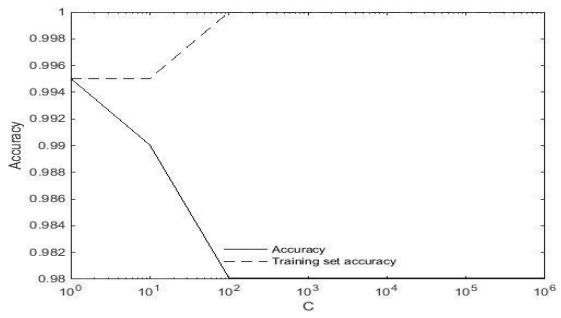

Fig7: Classification accuracy on the training dataset on different values of C

E. Cross validation

In this paper, two methods of cross-validation are used.

1. Repeated hold-out cross-validation
2. Leave-one-out validation method

These methods are introduced to eliminate the validation problems in the previous validation techniques.

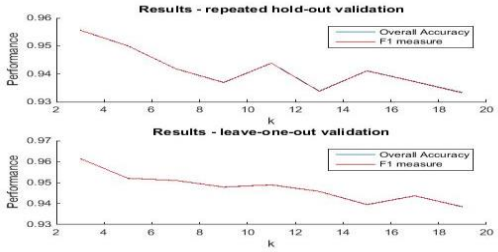

Fig7:Cross-validation of SVM labelling

The performance measurement of the cross-validated classified features.

True	Amount of Speech (%)	Amount of Music (%)
Speech	98.1	1.9
Music	5,8	94.2

Table2: The confusion matrix values

Performance measurement (per class)

Performance measure	Speech	Music
Precision (%)	94.4	98.0
Recall (%)	98.1	94.2
F1	96.2	96.1

Table 3: Performance measure Speech vs Music

F. Audio matching

Audio matching is done using Dynamic Time Warping (DTW) algorithm. Different samples are taken from the database and DTW algorithm is applied with cosine distance measurement for audio matching.

The minimum cost value shows the best matching. For positive values, minimum positive value gives the best match and for negative values, most negative value gives the best match.

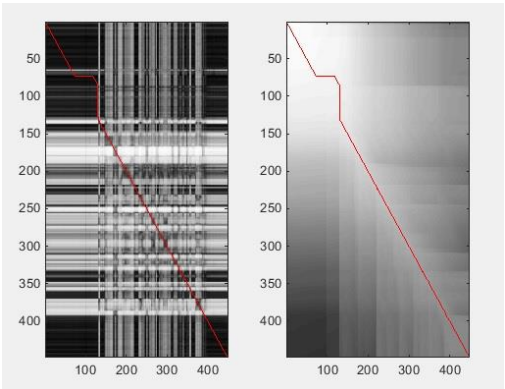

Fig 8: Audio matching using DTW

Sample	Minimum cost value	
PDAm01_001_1 &PDAm01_001_1	-7.9936e-15	Select from speech class
PDAm01_001_1 & PDAm01_001_5	105.045	Retrieve from speech class
clarinet10o & clarinet10e	203.459	select from music class

Table 4: DTW minimum cost values for audio matching from samples from database

Table 4 shows the minimum cost values of similarity check. From Fig 8, in the positive slope minimum positive value is the best match. For negative slope, the most negative value give the best match

VII. CONCLUSION AND FUTURE WORK

In this paper an efficient audio retrieval method is proposed which is done in a supervised learning medium. The simple and one of the best classifier, kNN classifier is used for audio classification. An accuracy of 96.2 % is obtained in audio labelling using Support Vector Machine (SVM) which is a supervised learning classifier.
Dynamic Time Warping (DTW) with cosine distance measurement is used for audio matching In future, this can be implemented in unsupervised learning medium and the performance of audio retrieval in both supervised and unsupervised learning techniques can be compared.

REFERENCES

[1] Teng Zhang, Ji Wu, Dingding Wang, Tao Li (2014) "Audio Retrieval Based on Perceptual Similarity". In IEEE international conference on collaborative computing.
[2] Qinghua Wu, Xiaolei Zhang, Ping Lv, Ji Wu (2012)."Perceptual Similarity between audio clips and feature selection for its measurement". In International symposium on Chinese language processing, IEEE.
[3] Kathy Melih, Ruben Gonzales (1998). "Audio Retrieval Using Perceptually Based Structures", In Proceedings of the IEEE Conference on Protocols for Multimedia Systems and Multimedia Networking, PROMS-Mm Net, pp338-347.
[4] Hung-Yi Lo, Ju-Chiang Wang, Hsin-Min Wang (2010). "Homogeneous segmentation and classifier ensemble for audio tag annotation and retrieval". In Proceedings of IEEE International Conference on Multimedia and Expo, pp 304-309.
[5] E. Wold, T. Blum, D. Keislar, J. Wheaten (1996). "Content-based classification, search and retrieval of audio". In IEEE Multimedia, vol.3 (3): 27-36.
[6] Theodoros Giannakopoulos, Aggelos Pikrakis (2014). "Introduction to audio analysis: A matlab approach. Elsevier publications limited.
[7] Frank Kurth, Meinard Müller (2008), "Efficient Index-Based Audio Matching"IEEE Transactions On Audio, Speech, And Language Processing, Vol. 16, No. 2.
[8] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi (2010), "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques". Journal Of Computing, Volume 2, Issue 3, March