Application of Vector Quantization for Audio Retrieval

Shruti Vaidya
ME.IT-Student
TCET, Mumbai

Kamal Shah, PhD
ME.IT-Professor
TCET, Mumbai

ABSTRACT
Due to the progress of the unlimited data storage capabilities and the proliferation use of the Internet, information retrieval systems encountered a large interest. Much of this data is in different forms from various sources. So, it becomes important to develop the necessary technologies for indexing and browsing such audio data. The consideration will be from audio information retrieval domain. In this paper; the study of audio content analysis for classification is presented, in which an audio signal is classified according to audio type. An approach that is capable of classifying an audio signal into speech, music, environment sound is used. Audio classification is processed in two steps as follows:- The first step of the classification is speech and nonspeech discrimination. The second step further divides nonspeech class into music, background sounds. Algorithms based on K-nearest-neighbor (KNN) and Fast Fourier Transform (FFT) are used. Experimental results indicate that satisfactory results are produced.

General Terms
Audio, music retrieval, precision recall, vector quantization

Keywords
Vector quantization, audio retrieval, classification of audio signals

1. INTRODUCTION
Vector quantization (VQ) is a classical quantization technique, where it allows the modeling of probability density functions by the distribution of prototype vectors, from the signal processing. It is generally used for data compression. It proceeds by dividing a large set of points (vectors) into groups having approximately the same number of points closest to them. It represents each group by its centroid point [1].

Vector Quantization has powerful density matching property, specially for identifying the density of large and high-dimensioned data. As data points are represented by the index of their closest centroid, therefore recurring data have low error, and rare data high error. Hence, Vector Quantization is more suited for lossy data compression. Its use can be extended to lossy data correction and density estimation. This paper considers the audio vector quantization domain.

Data compression using vector quantization has received great attention because of its promising compression ratio and simple implemented structure. For the simplest VQ implementation, it separates the signal into several sections and compresses each section into one vector. Each vector of the signal to be compressed is compared to the codevectors of a code-book. The address of the codevector most similar to the signal vector is sent to the receiver. At the receiver, the decoder accesses a codevector from an identical codebook, thus an approximation of the original signal is reconstructed. Compression is obtained by sending the index of the particular codevector thereby requiring fewer bits than sending the signal vector. The key to VQ data compression is a good codebook design.

2. SYSTEM OVERVIEW
Audio signal classification finds its utility in many research fields such as audio content analysis, information retrieval, etc. An audio signal classification system should be able to categorize different audio input formats. Particularly, detecting the audio type of a signal (speech, music, and background noise) allows such new applications as automatic organization of audio databases, segmentation of audio streams, intelligent signal analysis, etc. All classification systems employ the extraction of a set of features from the input signal [2].

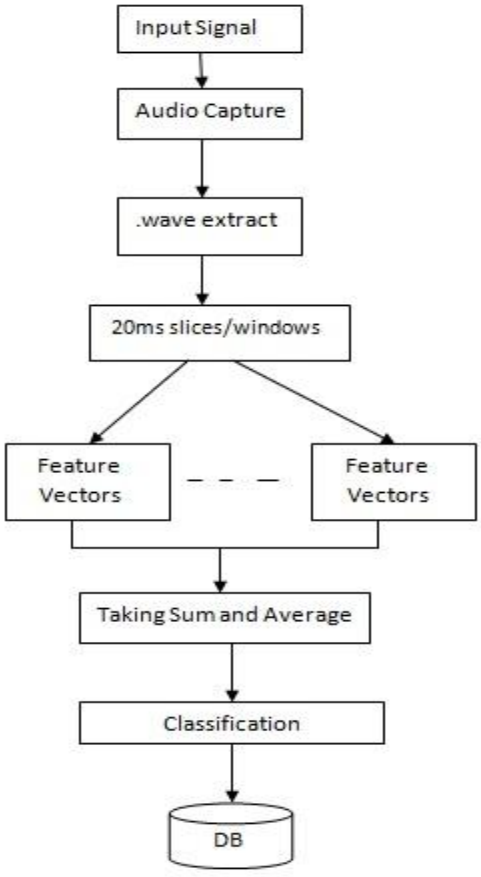

Figure 1: System Overview

Figure 1 represents the system architecture. An audio signal is received by any input means such as mike. The signal received passes through the audio capture where basic filtering occurs and then the signal is received by our system. The signal received is in the wave format, which is sliced into window size of 20 ms. After this the feature vectors are extracted. Later a sum and average is performed to get the average value which is then used to classify the signal into its appropriate class namely speech, music or background.

The next section discusses in detail the audio features used in audio segmentation and speaker segmentation. Further it presents the audio classification and segmentation scheme.

3. FEATURE ANALYSIS

In order to improve the accuracy of classification and segmentation for audio sequence, it is critical to select good features that can capture the temporal and spectral characteristics of audio signal or the characteristics of speaker vocal tract. The following features selected to classify or segment audio stream, high zero-crossing rate ratio (HZCRR), low short-time energy ratio (LSTER), spectrum flux (SF), spectrum roll off (SR) and spectrum centroid (SC). These features will be described in detail in this section [3].

3.1 High zero-crossing rate ratio

Zero-crossing rate (ZCR) is proved to be useful in characterizing different audio signals. It has been popularly used in speech/music classification algorithms. In the experiments, it is seen that the variation of ZCR is more discriminative than the exact value of ZCR. Therefore, high zero-crossing rate ratio (HZCRR) is used as one feature in our approach. HZCRR is defined as the ratio of the number of frames whose ZCR are not equal or below than 1.5-fold average zero-crossing rate in an 1-s window, as

$$HZCRR = \frac{1}{2N}\sum_{n=0}^{N-1}[sgn(ZCR(n) - 1.5\ avZCR) + 1] \quad (1)$$

Where is the frame index, ZCR is the zero-crossing rate at the n^{th} frame, is the total number of frames, avZCR is the average ZCR in a 1-s window; and sgn[.] is a sign function, respectively.

In general, speech signals are composed of alternating voiced sounds and unvoiced sounds in the syllable rate, while music signals do not have this kind of structure. Hence, for speech signal, its variation of zero-crossing rates (or *HZCRR*) will be in general greater than that of music, as shown in Figure 2.

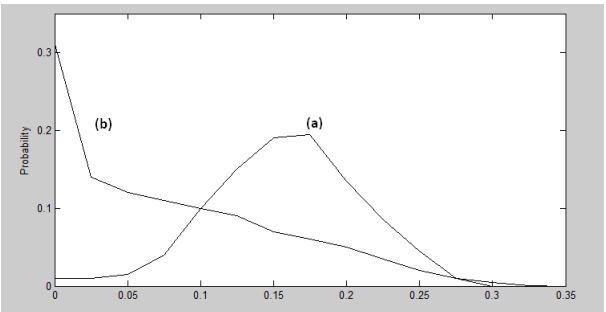

Figure 2: Probability Distribution Curves of HZCRR
a) Speech b) Music
b)

3.2 Low Short Time Energy Ratio

Here, low short-time energy ratio (LSTER) is used to represent the variation of short-time energy (STE). LSTER is defined as the ratio of the number of frames whose STE are less than 0.5 time of the average short-time energy in a 1-s window, as the following:

$$LSTER = \frac{1}{2N}\sum_{n=0}^{N-1}[sgn(0.5\ avSTE - STE(n)) + 1] \quad (2)$$

Where the total number of frames, STE is the short-time Energy at the nth frame and avSTE is the average STE in a 1-s window.

LSTER is an effective feature, especially for discriminating speech and music signals. Most commonly, there are more silent frames in speech than in music; as a result, the LSTER measure of speech will be much higher than that of music. This can be explained by the probability distribution curves of LSTER for speech and music signals.

3.3 Spectrum Flux

Spectrum flux (SF) is defined as the average variation value of spectrum between the adjoining two frames in a 1-s window

$$SF = \frac{1}{(N-1)(K-1)}\sum_{n=1}^{N-1}\sum_{k=1}^{K-1}.[\log(A(n,k) + \delta - \log(A(n-1,k) + \delta)]^2 \quad (3.1)$$

Where A(n,k) is the discrete Fourier transform of the nth frame of input signal

$$A(n,k) = \sum_{m=-\infty}^{\infty} x(m)w(nL-m)e^{-j\left(\frac{2n}{L}\right)km} \quad (3.2)$$

and x(m) is the original audio data, w(m) is the window function, L is the window length, K is the order of DFT, N is the total number of frames and δ is a very small value to avoid calculation overflow.

In the experiments, it is found that, in general, the SF values of speech are higher than those of music. The environment sound is among the highest and changes more dramatically than the other two types of signals. Figure 3 shows an example of spectrum flux of speech, music and environment sound. SF is a good feature to discriminate speech, environmental sound and music.

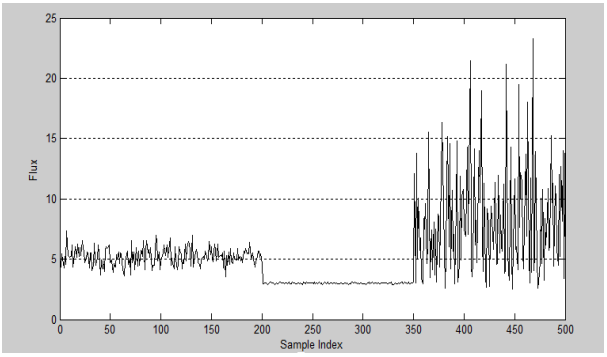

Figure 3: Spectrum Flux Curve (0-200 s is speech, 201-350 s is music, and 351-450 s is environment sound)

3.4 Spectrum Roll Off

The spectral roll off can be said as the frequency Rt, where 85% of the magnitude distribution is concentrated below [4].

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n=1}^{N} M_t[n] \quad (4)$$

3.5 Spectrum Centroid

The gravity's centre of the magnitude spectrum of the STFT is defined as the spectral centroid [4].

$$C_t = \frac{\sum_{n=1}^{N} n \cdot |M_t[n]|^2}{\sum_{n=1}^{N} |M_t[n]|^2} \quad (5)$$

Where Mt [n] is the magnitude of the Fourier transform at frame t and frequency bin n. Measure of spectral shape and higher centroid values correspond to "brighter" textures with more high frequencies is the centroid.

4. ALGORITHMS USED FOR CLASSIFICATION

4.1 (kNN) k- Nearest Neighbour Classifier

The nearest neighbor method consists of assigning to the unlabelled feature vector the label of the training vector that is nearest to it in the feature space. Here, a training set T is considered to determine the class of a previously unseen sample X. Primarily, we determine the mean and maximum values in T, and also, for the unseen sample X. Then an appropriate distance measure in the feature space is used to determine k elements in T closest to X. If most of these, k nearest neighbors contain similar values, then X is classified accordingly. This classification scheme clearly defines nonlinear decision boundaries and thus improves the performance. Furthermore, the feature distribution suggests that the number of data-points used in the example set T can be considerably reduced for faster processing; only those examples that are close to the decision boundary are actually required [2,5].

This can be explained by the following example [7]. Referring to Figure.4 each of the samples (marked by stars) have been labeled either A or B, except for the sample x. This needs to be labeled, the kNN classifier takes the k nearest, that is, the closest neighbors around the sample x and uses them to assign a label. This is generally done by a majority-voting rule, which says that the label assigned should be, which occurs the most among the neighbors.

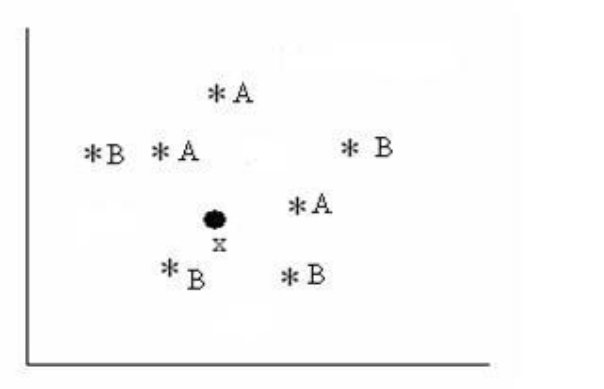

Figure 4: Example of k-nearest neighbour rule [2]

For example if k=1, the sample that is nearest to the sample x is sample B. Hence the unknown sample x is referred as B. But if k=7, then there are four samples of A and three samples of B that are closer to the sample x. The sample x gets assigned as A by virtue of majority. Hence it can be inferred that the value of k is critical, to assign an unknown sample by its nearest neighbors.

As it can be seen that for k=7 sample x gets assigned to A on the basis of majority polling. It is also important to note that the k neighbors have been assumed to have equal influence on predictions irrespective of their relative distance from the query point. Since the three samples of B are closer than the four samples of A on the basis of distance, the former shows to have a greater influence on sample x even though the samples A are in majority. Hence, to increase the efficiency, there is an equal need to pay attention to the relative distance of the k nearest samples to the query point in order that the unknown sample gets assigned to the sample that has greater influence on it.

4.2 (FFT) Fast Fourier Transform

A Fast Fourier Transform (FFT) is an algorithm to compute the discrete Fourier transform (DFT) and its inverse. A Fourier transform converts time (or space) to frequency and vice versa; an FFT rapidly computes such transformations. The FFT is obtained by decomposing a sequence of values into components of different frequencies more quickly. This operation is useful in many fields [6]. After the conversion into the frequency domain feature vectors can be applied for further processing.

5. DISTANCE CALCULATION

The use of the above described feature vectors such as high zero-crossing rate ratio (HZCRR), low short-time energy ratio (LSTER), spectrum flux (SF), spectrum roll off(SR) and spectrum centroid (SC), allows the input signal to be classified accordingly into its appropriate domain.
Once the appropriate domain is obtained, the different signals belonging to the domain can be retrieved by the use of any of the following distance calculation methods [8]:-

5.1 Manhattan Distance

The Manhattan distance function computes the distance that would be traveled to get from one data point to the other if a grid-like path is lead. The Manhattan distance between two entities is the sum of the differences of their corresponding components. The formula for this distance between a point X=(X1, X2, etc.) and a point Y=(Y1, Y2, etc.) is:

$$d = \sum_{i=1}^{n} |X_i - Y_i| \quad (6)$$

Where n belongs to the number of variables, and Xi and Yi are the values of the i^{th} variable, at points X and Y respectively.

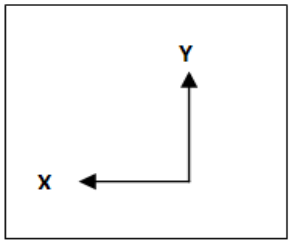

Figure 5: Manhattan Distance

5.2 Euclidean Distance

The formula for Euclidean distance between a point X (X1, X2, etc.) and a point Y (Y1, Y2, etc.) is:

$$d = \sqrt{\sum_{j=1}^{n}(x_j - y_j)^2} \quad (7)$$

Deriving the distance between two data points involves computing the square root of the sum of the squares of the differences between corresponding values.

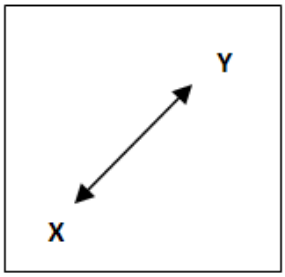

Figure 6: Euclidean Distance

6. PRECISION AND RECALL

Precision and recall are greatly used parameters in evaluating the correctness of a pattern recognition algorithm [9].

Precision is a measure of fidelity whereas recall is a measure of completeness. Precision basically is a measure of the number of retrieved documents that are relevant to the search. Precision can also be evaluated at a given cut off n.
Recall as mentioned earlier is a measure of completeness i.e. it is basically the probability of a relevant document being returned by the query. In binary classification, recall is also referred as sensitivity. The mathematical computation of precision and recall is as follows:

Precision = |{Relevant Documents}∩{Retrieved Documents}| / | {Retrieved Documents} |

Recall = | {Relevant Documents} ∩ {Retrieved Documents} | / | {Relevant Documents} | (8)

Crossover point in precision recall is the point on the graph where both the precision and recall curves meet.
Crossover point in itself can be used a way to measure the correctness of an algorithm. A higher crossover point indicates a better performance for a particular method.

7. RESULT

The study is presented on the basis of different audio input signals taken in wave format. The obtained audio signals are in double channel and need to be converted into single channel. The audio input should have: 44100 sampling frequency, 16 bit single channel.

The received incoming signals are analysed by using the various feature vectors as studied above, to determine the class it belongs to. The graphical representation supports the result.

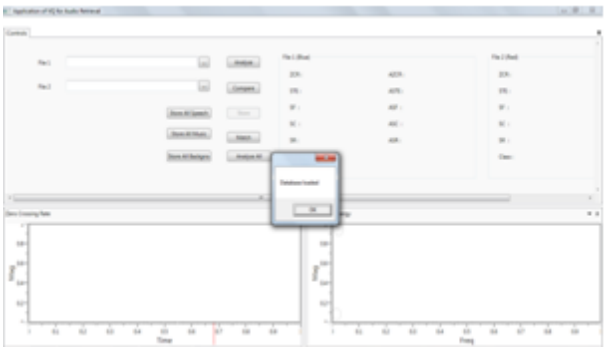

Figure 7: Loading the database

Figure 7 represents the result. When the database for each class speech, music, background is loaded, a popup message is displayed saying "Database Loaded".

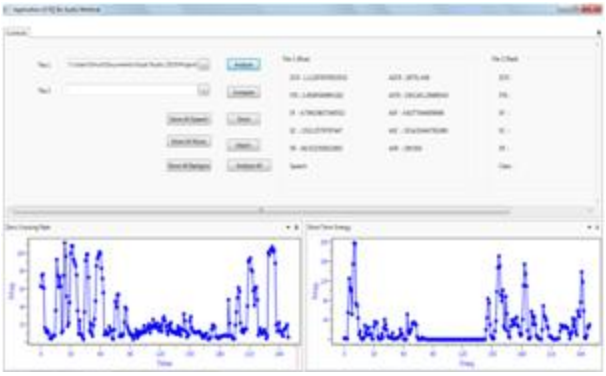

Figure 8: Analyzing signal

Figure 8 shows the signal analysis. When an unknown signal is analyzed, it gives the desired outcome, to which class it belongs. Graphical representation for zero crossing rate and short time energy feature vectors is shown.

Figure 9: Retrieving similar signals

Once the incoming signal is classified into its appropriate domain such as music, background, speech, the other signals belonging to that particular domain are retrieved by distance calculation, and selected signal can be heard. Figure 9 represents this.

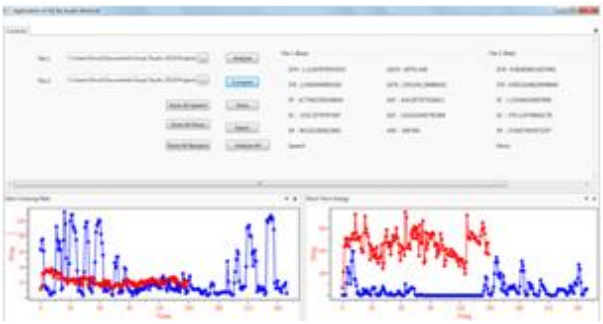

Figure 10: Distinguishing incoming signals

Figure 10 represents the analysis of two incoming signals and distinguishes them into their appropriate domain.

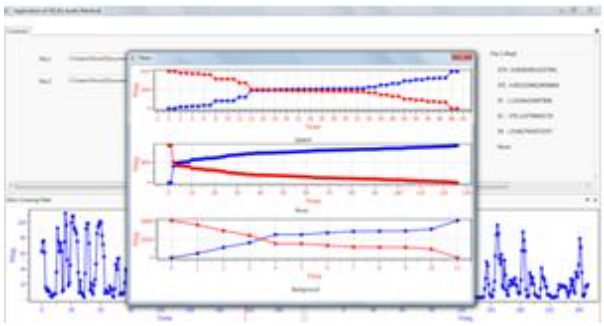

Figure 11: Precision Recall Graph

Figure 11 displays a Precision Recall Graph for all the three classes namely, Speech, Music, Background.

8. CONCLUSION

It can be inferred that the incoming unknown audio signal is classified as speech or music signal as per its nature. This is done by analyzing its properties which means that its features that define its nature have to be extracted. The different features have been studied. Care is taken to optimize the number of features selected because each feature represents a dimension in the feature space. Therefore reducing number of features reduces the computational costs and at the same time maintains the accuracy levels. The subsequent process after the feature extraction is the classification process. It is the responsibility of the classifier to accurately label the signal using the features selected so that the nature of the unknown audio taxonomy is known and it is classified under a known class of audio signals. The requirements for a classifier require that it must be computationally efficient with less complexity in its algorithm that economizes its cost.

9. REFERENCES

[1] T. Rammohan Associate Professor, Dept. of Electronics and communication Engineering, K. Sankaranarayanan PhD. Dean, Electrical sciences, EASA college of Engineering and Technology, Shalakha Rajan Dept. of Electrical and Electronics Engineering, " Image Compression using Fusion Technique and Quantization", International Journal of Computer Applications (0975 – 8887) Volume 63– No.22, February 2013

[2] Hariharan Subramanian, Prof. Preeti Rao, Dr. Sumantra. D. Roy" Audio Signal Classification" M.Tech. Credit Seminar Report, Electronic Systems Group, EE. Dept, IIT Bombay, Submitted November2004

[3] Lie Lu, Hong-Jiang Zhang, *Senior Member, IEEE*, and Hao Jiang. "Content Analysis for Audio Classification and Segmentation ", IEEE Transactions On Speech And Audio Processing, Vol. 10, No. 7, October 2002

[4] Cheng-ya Sha, "Time Frequency Analysis for Acoustics", ntu.edu.tw

[5] Finnish Meteorological Institute, www.geo.fmi.fi/~syrjasuo/Analysis/node6.html, October 2004.

[6] Gokul P, Karthikeyan T , KrishnaKumar R. Malini S., Final Year ECE students, Department of Electronics and Communication Engineering, Assistant Professor, Department of Electronics and Communication Engineering, " Computational Time Analysis of Signal Processing Algorithm-An Analysis" IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) e-ISSN: Gokul P, Karthikeyan T , KrishnaKumar R. Malini S., Final Year ECE students, Department of Electronics and Communication Engineering, Assistant Professor, Department of Electronics and Communication Engineering, " Computational Time Analysis of Signal Processing Algorithm-An Analysis" IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) e-ISSN:

[7] The FreeDictionary.Com by Farlex, encyclopedia.thedictionary.com, October 2004.

[8] T. Soni Madhulatha, Associate Professor, Alluri Institute of Management Sciences, Warangal, "An Overview On Clustering Methods", IOSR Journal of Engineering Apr. 2012, Vol. 2(4) pp: 719-72

[9] Dr. H. B. Kekre Sr. Prof, Department of Computer Science, NMIMS University, Mr. Dhirendra Mishra Associate Prof, Department of Computer Science, NMIMS University, Mr. Anirudh Kariwala Student, Department of Computer Science, NMIMS University," A Survey Of CBIR Techniques And Semantics" , Dr. H.B. Kekre et al. / International Journal of Engineering Science and Technology (IJEST)

[10]