Volume 4, No. 12, December 2013
Journal of Global Research in Computer Science
ISSN-2229-371X
REVIEW ARTICLE
Available Online at www.jgrcs.info

COMPARISON OF ENHANCED SCHEMES FOR AUDIO CLASSIFICATION

Dr. V. Radha[*1] and G.Anuradha[2]
[*1]Professor, [2]Research Scholars
Department of Computer and Applications
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore – 43.
radhaasrimail@gmail.com*, anu.anuradhasrinivas@gmail.com

Abstract: In the modern era of communication, audio plays an important role in understanding a digital media. Due to the rise of economical audio capturing devices, the amount of audio data available both online and offline is enormous and techniques that can automatically classify and retrieve these audio data is an immediate need. An automatic content based audio classification and retrieval system consists of three modules namely, feature extraction, classification and retrieval. This paper presents a comparative study of two algorithms that performs these three steps in different manners. The performance of the selected systems are analyzed while using four different features (acoustic, perceptual, mel-frequency cepstral coefficients (MFCC) and a combination of perceptual and MFCC) and four classifiers that enhanced Support Vector Machine (SVM) and Centroid Neural Network (CNN) along with its base versions, SVM and CNN. Experimental results showed that the enhanced SVM algorithm when using the combined feature vector produced improved accuracy and reduced error rate.

Keywords: Audio Classification, Audio Retrieval, Support Vector Machine, Centroid Neural Network, Audio Features, Modified Euclidean Distance.

INTRODUCTION

A Content-based Audio Classification and Retrieval (CACR) System automatically group audio data in large database into different audio types (categories) which can then be searched for a particular sound or a class of sound electronically based on the content analysis of audio signals [15]. The primary goal is to group audio files (based on their content) into one of a number of predefined categories. A general framework is shown in Figure 1.Audio data collection consists of audio instances or audio files. The first step of CACR extracts audio features that represent characteristic information about these audio instances. The extracted features are stored as feature vectors or spaces, which are used to train a classifier. When an input music is obtained, the same features are extracted. The machine learning algorithm associates the feature patterns of instances with their classes and maps it to a class. All the data in that class are then retrieved as matched audio. Thus, any CACR system, consist of two steps, namely, feature selection and classification based on the extracted features. The manner of handling these two steps is directly related to the efficiency of the audio classification and retrieval system.

The methods proposed for CACR can be grouped intofive categories, namely, query by example, query by humming, music information retrieval, similarity matching methods and machine learning methods. Query by Example systems aims at automatic retrieval of media samples from a database, which are similar to a user provided example [25, 10,11,3, 28]. Query by Humming techniquesare similar to Query by Example, but here the user provides the sample by humming the song [14, 8, 5, 21, 16] (iii) Music Information Retrieval systems are Google-like search enginesmainly used to retrieve similar audios[23, 4, 22, 2] iv) Similarity Matching based Methods use distance measures like

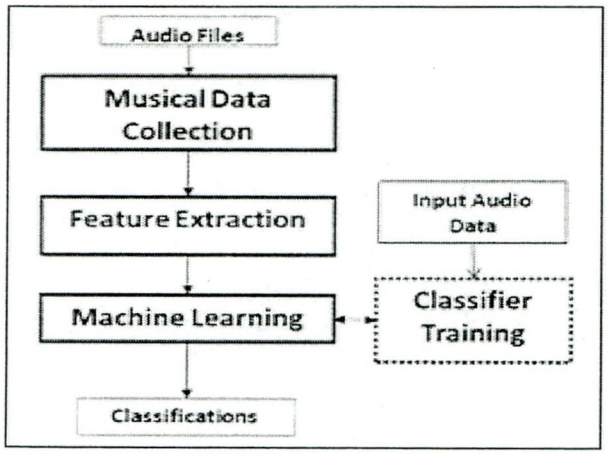

Figure 1 : CACR System

Euclidean distances to compare input audio data with database files and all the files that are close (similar) to each other are retrieved [1, 6, 8, 27] and (v) Machine Learning Methodsthatuse learning algorithms like artificial neural networks, k nearest neighbours, AdaBoost, to build models which play a vital role during classification and retrieval [13, 20, 24].

As it can be seen, several algorithms have been proposed under each of these categories, which either enhance an existing algorithm to improve classification accuracy or propose a new method that work better than existing algorithms. Two works that belong to the second category are the proposals of [17] and [7] uses a centroid neural network with a divergence measure to perform of Gaussian Probability Density Function (GPDF) data. In comparison with other conventional algorithms, the DCNN designed for probability data has the robustness advantages of utilizing a

audio data representation method in which each audio data is represented by a Gaussian distribution feature vector. The author used a total of 42 features covering timbral, rhythmic and pitch features during classification. This feature set is referred to as acoustic feature set in this paper. On the other hand, [7] used SVMs with a binary tree recognition strategy for classifying and retrieving audio data. For audio retrieval, the authors proposed a new metric, called Distance-From-Boundary (DFB). When a query audio is given, the system first finds a boundary inside which the query pattern is located. Then, all the audio patterns in the database are sorted by their distances to this boundary. All boundaries are learned by the SVMs and stored together with the audio database. This system used two types of feature sets, namely, perceptual features and MFCC (mel-frequency cepstral coefficients) features during classification.

The research problem of the present research work is to compare these two works on their ability to classify audio data to enhance the content-based audio retrieval process. For convenience, the [7] model is referred to as GL-AC and [17] model is referred to as P-AC in this dissertation. The rest of the paper is organized as follows. Section 2 presents the GL-AC system and Section 3 presents the P-AC system. Section 4 presents the results and compares both the algorithms in their efficiency in classifying audio data. Section 5 concludes the work with future research directions.

GL-AC SYSTEM

The GL-AC method consists of three main modules, namely, feature extraction module, classification module and retrieval module. The first module extracts two audio features namely, perceptual features and mel-cepstral features, which are then combined to form a third feature set. Perceptual features refer to the sensation of sound by humans. The perceptual features collected are total spectrum power (Equation 1), subband-power (4 subbands) (Equation 2), brightness (Equation 3), bandwidth (Equation 4) and pitch. Pitch is the fundamental period of a human speech waveform and is an important parameter in the analysis and synthesis of speech signals. In GL-AC algorithm, a simple pitch detection algorithm based on detecting the peak of the normalized autocorrelation function is used. The pitch frequency is returned if the peak value is above a threshold (T = 0.65, chosen empirically) or the frame is labeled as non-pitched. Apart from this, two more features, namely, silence ratio which is the ratio of number of silent frames to total number of frames and pitched ratio which is the ratio of number of pitched frames total number of frames are also calculated.

$$P = \log\left(\int_0^{\omega_0} |F(\omega)|^2 \, d\omega\right) \quad (1)$$

$$P_j = \log\left(\int_{L_j}^{H_j} |F(\omega)|^2 \, d\omega\right) \quad (2)$$

$$\omega_C = \frac{\int_0^{\omega_0} \omega |F(\omega)|^2 \, d\omega}{\int_0^{\omega_0} |F(\omega)|^2 \, d\omega} \quad (3)$$

$$B = \sqrt{\frac{\int_0^{\omega_0} (\omega - \omega_C) |F(\omega)|^2 \, d\omega}{\int_0^{\omega_0} |F(\omega)|^2 \, d\omega}} \quad (4)$$

Whereω and $\omega_{0\ 0}$ is the frequency and half sampling frequency, $|F(\omega)|^2$ is the power at the frequency ω, L_j and H_j are lower and upper bound of sub-band j.

The cepstrum can be illustrated by use of the Mel-frequency cepstra coefficients (MFCCs). These are computed from the FFT power coefficients. The power coefficients are filtered by a triangular bandpass filter bank. The filter bank consists of K=19 triangular filters. They have a constant mel-frequency interval and cover the frequency range of 0Hz – 4000Hz. Denoting the output of the filter bank by S_k (k = 1, 2, …, K), the MFCCs are calculated as :

$$C_n = \sqrt{\frac{2}{K}} \sum_{k=1}^{K} (\log S_k) \cos[n(k-0.5)\pi/K] \quad (5)$$

Where n = 1, 2, …, L and L is the order of the cepstrum.

To form the feature vector, the mean and standard deviation of the perceptual features along with silece and pitch ratio forms the 18-dimensional perceptual feature vector denoted as Percfeature Set in this paper. Similarly, CepsL feature set is obtained using cepstrumfaeturs. Both these features are combined to form PercCepsL feature set using Equation 6.

$$PercCeptL = (Perc/s_1) \oplus (CepsL/S_2) \quad (6)$$

Where $\oplus$ stands for the concatenation operation. The second module uses Support Vector Machine (SVM) for classification. During video retrieval, a new Distance From Boundary (DFM) distance measure is used instead of the traditional Euclidean distance measure.Given a set of training vectors belonging to two classes, SVM tries to separate the data into two hyperlanes. Several possible hyperplanes can be formed, but the algorithm should select one that maximizes the margin (the distance between the hyperplane and the nearest data point of each class).

For this purpose, kernel functions are used. The problem of audio classification is a multi-class problem and is solved by SVM by combining results of binary classifiers. The problem is now on the decision process used to combine the binary classification results to obtain the final decision. A common method used frequently is the voting strategy, which is computation expensive as it requires c(c-1)/2 comparisons. This problem is solved in GA-AC with the use of bottom-up binary trees. The formation of binary tree starts at the lowest level where a comparison is made between each pair and a winner is chosen. At the next stage, the winner will be moved one level up and the process is repeated. At the end of iteration, a unique class label will be at the top level. Usage of binary tree reduces the number of comparisons required from c(c-1)/2 times to (c-1) times.

During the retrieval stage, conventional methods use Euclidean distance to measure the similarity between audio patterns of the database and query. The traditional method

has disadvantages like being sensitive to the sample distribution, different query patterns of the same class produces different retrieval results and finally, the average retrieval accuracy is low.These problems are solved by the use of a new metric called Distance from Boundary metric and work on the principle that a boundary exists and separates the samples belonging to one class with the remaining. This nonlinear boundary encloses the similar patterns inside no matter what the distribution is. These boundaries can easily be combined with SVM training process and requires only simple operations and therefore, are computation inexpensive.

P-AC SYSTEM

The P-AC method consist of two modules, namely, feature extraction and classification. In the feature extraction module, three types of features, namely, timbral texture features, rhythmic content features and pitch content features are extracted from the audio data. The Timbral texture features should exhibit properties related to general ɔre of the sound. They are based on a Short Time Fourier Transform (STFT) and they are calculated on short-time frames of a sound (MFCC). The feature vector for describing timbral texture consists of the following features: means and variances of spectral centroid, rolloff, flux, zero crossings over texture window, low energy and mean and variances of the first five MFCC coefficients over the texture window. The rhythmic content features represent rhythmic structure of the music. The selected features are relative amplitude of the first and the second peaks, ratio of the second and the first peaks, period of the first and the second peaks, overall sum of the beat histogram. These features are based on detecting the most salient periodicity of the signal by using Discrete Wavelet Transform technique. The Pitch content features characterize audio signals in terms of energy of different frequency bands and are based on multiple pitch detection techniques [26].

The classification module uses a divergence based CNN classifier [26, 19, 9] which uses Bhattacharyya distance equation 7) instead of the traditional Euclidean distance

$$D(G_i,G_j)=\frac{1}{8}(\vec{\mu}_i,\vec{\mu}_j)^T\left[\frac{\vec{\Sigma}_i+\vec{\Sigma}_j}{2}\right]^{-1}(\vec{\mu}_i,\vec{\mu}_j)+\frac{1}{2}\ln\frac{\left|\frac{\vec{\Sigma}_i+\vec{\Sigma}_j}{2}\right|}{\sqrt{|\vec{\Sigma}_i|+|\vec{\Sigma}_j|}} \quad (7)$$

Where $\vec{\mu}_i$ and $\vec{\Sigma}_i$ denote the mean vector and covariance matrix of Gaussian distribution G_i, respectively. $\vec{\mu}_j$ and $\vec{\Sigma}_j$, denote the mean vector and covariance matrix of Gaussian distribution G_j, respectively and T denotes the transpose matrix. The concept of winner and loser in the CNN can be adopted for the D-CNN without any change except for application of the divergence measure for the distance calculation. In this case, however GPDFs have two parameters to consider: mean, μ, and diagonal covariance, Σ. The weight update for mean is the same as the CNN weight update. By using the divergence distance as its distance measure, the D-CNN have abilities in clustering the probability data while it still keep advantageous features of the CNN. Because the CNN have been proven to outperform other conventional clustering algorithms such as k-means and CNN, the D-CNN should show improvements over the k-means and CNN algorithms in probabilistic data. The pseudo code of D-CNN algorithm is given in Figure 2.

Algorithm D-CNN(C,D) [*C: number of clusters, D: number of data vectors*]
[*Initialize weights* $\mathbf{w}_1=(\mu_{w_1},\Sigma_{w_1})$ *and* $\mathbf{w}_2=(\mu_{w_2},\Sigma_{w_2})$]
Find the centroid, **c**, of all data vector
Initialize $\mathbf{w}_1$ and $\mathbf{w}_2$ around **c** with small **ε** :
$\mu_{w_1}:=\mu_c+\varepsilon$, $\mu_{w_2}:=\mu_c-\varepsilon$, $\Sigma_{w_1}:=\Sigma_{w_2}:=\Sigma_c$
$k:=2$, $epoch:=0$
for $(k\le C)$
do
$loser:=0$
for $(n=1)$ **to** D
Apply a data vector $\mathbf{x}(n)$ to the network
Find the winner neuron, j, using Divergence distance for $1\le j\le k$.
if $(epoch\neq 0)$ **then** Set i is winner neuron, i, for $\mathbf{x}(n)$ in previous epoch.
if $(i\neq j)$, **then** neuron, i, is loser neuron.
if $(epoch=0$ or $i\neq j)$
Run UpdateD-CNNWeightMean(μ_{w_j} , μ_{w_i} , $epoch$)
$loser:=loser+1$
endif
endfor [*check for all data*]
$\Sigma_{w_j}(epoch+1)=\sum_q^{N_j}(\Sigma_{xqj}+(\mu_{xqj}-\mu_{w_j}(epoch))^2)/N_j$
$j=1,2,\ldots,k$
$epoch:=epoch+1$
while $loser\neq 0$
if $k\neq M$
split the most erroneous group, j, by adding a small vector, ϵ, nearby group j
$\mu_{w_{k+1}}=\mu_{w_j}+\epsilon$, $\Sigma_{w_{k+1}}=\Sigma_{w_j}$
endif
$k:=k+1$
endfor
end

Procedure UpdateD-CNNWeightMean(μ_{w_j}, μ_{w_i}, $epoch$)
[*Update the winner neuron,* $\mathbf{w}_j$, *and loser neuron,* $\mathbf{w}_i$]
Update winner neuron : $\mu_{w_j}(n+1)=\mu_{w_j}(n)+(\mu_x(n)-\mu_{w_j}(n))/(N_j+1)$
if $epoch\neq 0$ [*loser neuron is occurred only when* $epoch\neq 0$]
Update losing neuron : $\mu_{w_i}(n+1)=\mu_{w_i}(n)-(\mu_x(n)-\mu_{w_i}(n))/(N_i-1)$
endif
end

Figure 2: Divergence-Based Centroid Neural Network Algorithm (Source: [18])

EXPERIMENTAL RESULTS

During experimentation, two datasets were used. The first is an audio dataset downloaded from [12]. This dataset has 409 sounds having 16 classes. The names of the audio classes are altotrombone, animals, bells, cellobowed, crowds, female, laughter, machines, male, oboe, percussion, telephone, tubularbells, violinbowed, violinpizz, water. This dataset is referred to as MuscleFish dataset. The second dataset was created with 2,663 audio signals having rock, pop, jazz, hiphop, folk, country, speech and natural sounds. This dataset is referred to as Web dataset. With both datasets, a 70%-30% hold-out method was used to separate the training and testing datasets. All the experiments were conducted using 10-fold method and the average results are projected.

To evaluate the classification performance, two metrics, namely, error rate and average retrieval accuracy are used. Error rate is defined as the ratio between the number of misclassified examples and the total number of testing examples. The average retrieval accuracy is defined as the

average percentage number of patterns belonging to the same class as the query in the top matches. Further, to analyze the efficiency gain obtained by D-CNN and SVM-BTS, the results are compared with the traditional counterparts, CNN and SVM, respectively.

The average accuracy obtained while using four different datasets, Acoustic, Perc, CepsL and PercCepsL for the four classifiers, CNN, SVM, D-CNN and SVM-BTS is shown in Figure 3.From the results, it is evident that the concatenated PERC and CEPSL feature sets while used with SVM-BTS algorithm produced better accuracy when compared with other classifiers and feature sets. The SVM-BTS classifier showed 2.33%, 1.37% and 0.56% accuracy efficiency gain when compared with CNN, D-CNN and SVM respectively. This shows that the usage of concatenated features with SVM-BTS is well suited for automatic audio classification and retrieval.

The average error rate of the four classifiers while using different features sets is shown in Figure 4.The trend ained while considering the error rate is similar to that of accuracy. The SVM-BTS classifier with the combined feature set produces the lowest error rate.

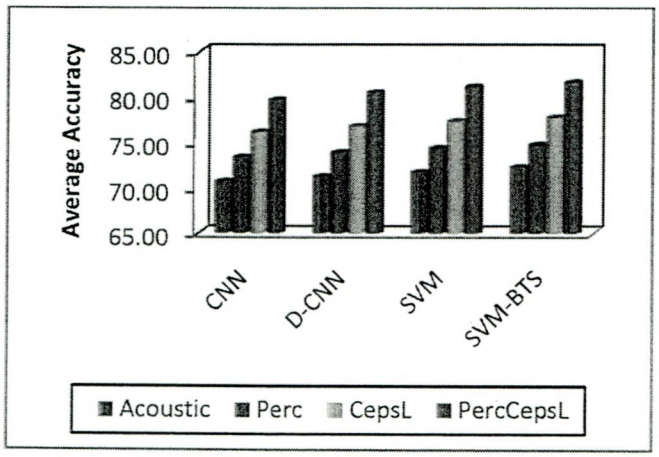

Figure 3: Average Accuracy

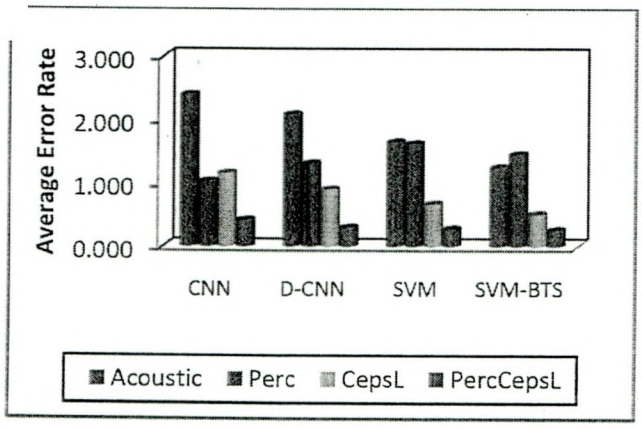

Figure 4: Average Error Rate

CONCLUSION

This paper presented a comparative study on two audio classification and retrieval systems, namely, GP-AC and P-AC, which enhanced the traditional classification algorithms. While both the algorithms follow the same steps, they different in the number of feature sets and classifiers used. The GP-AC used three features sets (perceptual, MFCC and a combined set) while P-AC used acoustic features composed on timbral texture, rhythmic content features and pitch content features. The GP-AC used an enhanced SVM classifier using bottom-up binary tree to reduce the computations while the P-AC enhanced Centroid Neural Network (CNN) to employ Bhattacharyya distance instead of Euclidean distance. Experimental results showed that SVM combined with DFB distance measure using the combined feature vector is more accurate and produced minimum error and hence is the best candidate for audio retrieval systems.

REFERENCES

[1]. Aronovich, L. and Spiegler, I. (2007) CM-tree: A dynamic clustered index for similarity search in metric databases, Data and Knowledge Engineering, Vol. 63, Pp. 919–946.

[2]. Aucouturier, J.J., Pachet, F. and Sandler, M. (2005) the way it sounds: Timbre models for analysis and retrieval of music signals, IEEE Transactions onMultimedia, Vol. 7, No. 6, Pp. 1028–1035.

[3]. Barrington, L., Chan, A., Turnbull, D. and Lanckriet1, G. (2007) Audio information retrieval using semantic similarity, IEEE International Conference on Acoustics, Speech and Signal Processing, Pp. 725–728.

[4]. Casey, M.A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C. and Slaney, M. (2008) Content-based music information retrieval: Current directions and future challenges, Proceedings of IEEE, Pp. 668–696.

[5]. Chen, L. and Hu, B.G. (2007) An implementation of web based query by humming system, ICME, Pp. 1467–1470.

[6]. Gartner, D., Kraft, F. and Schaaf, T. (2007) An adaptive distance measure for similarity based playlist generation, IEEE International Conference on Acoustics, Speech and Signal Processing, Pp. 229–232.

[7]. Guo, G. and Li, S.Z. (2003) Content-Based Audio Classification and Retrieval by Support Vector Machines, IEEE Transactions on Neural Networks, Vol. 14, No. 1, Pp. 209-215.

[8]. Guo, L., He, X., Zhang, Y., Lu, Y. and Peng, K. (2007) A noise robust content-based music retrieval for mobile devices, IEEE International Conference on Multimedia and Expo, Pp. 2222–2225.

[9]. Hartigan, J. (1975) Clustering Algorithms, New York, Wiley.

[10]. Helan, M. and Virtanen, T. (2007a) A similarity measure for audio query by example based on perceptual coding and compression, Proceedings of 10th International Conference on Digital Audio Effects, Bordeaux, France.

[11]. Helen, M. and Virtanen, T. (2007b) Query by example of audio signals using Euclidean distance between Gaussian mixture models, IEEE International Conference on Acoustics, Speech and Signal Processing, Pp. 225–228.

[12]. http://www.musclefish.com/cbrdemo.html, Last Access Date : 20-07-2012.

[13]. Koh, L.H., Ranganath, S. and Venkatesh, Y.V. (2002) An integrated automatic face detection and recognition system, Pattern Recognition, Vol. 35, Pp. 1259-1273.

[14]. Kohonen, T. (1990) The Self-Organizing Map. Proc. IEEE, Vol. 78, Pp. 1464-1480

[15]. Malik, H. (2012) Content-Based Audio Indexing and Retrieval: an overview, Pp. 1-10, www-personal.engin.umd.umich.edu/~hafiz/CBAIR_survey.pdf, Last Access Date : 01-07-2012.

[16]. Mulder, T.D., Martens, J., Pauws, S., Vignoli, F., Lesaffre, M., Leman, M., Baets, B. and Meyer, H. (2006) Factors affecting music retrieval in query-by-melody, IEEE Transactions on Multimedia, Vol. 8, No. 4, Pp. 728–739.

[17]. Park, D.C. (2010) Content-based Retrieval of Audio Data using a Centroid Neural Network, IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), South Korea, Pp. 394 – 398.

[18]. Park, D.C. and Kwon, O.H. (2008) Centroid Neural Network with the Divergence Measure for GPDF Data Clustering, IEEE Transactions on Neural Networks, **Vol.** 19 , Issue 6, **Pp.** 948 – 957.

[19]. Park, D.C. and Woo, Y. (2001) Weighted Centroid Neural Network for Edge Reserving Image Compression, IEEE Transactions on Neural Networks, Vol. 12, Pp. 1134-1146

[20]. Ravindran, S., Schlemmer, K. and Anderson, D.V. (2001) A physiologically inspired method for audio classification, Journal on Applied Signal Processing, Vol. 9, Pp. 1374-1381.

[21]. Rho, S. and Hwang, E. (2006) FMF query adaptive melody retrieval system, Journal of systems and software, Vol. 79, Pp. 43–56.

[22]. Rho, S., Han, B., Hwang, E. and Kim, M. (2007) Musemble: A music retrieval system based on learning environment, ICME, Pp. 1463 1466.

[23]. Ruxanda M.M., Chua, B.E., Nanopoulos, A, and Jensen, C.S. (2009) Emotion-based music retrieval on a well-reduced audio feature space, IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Pp. 181–184.

[24]. Song, Y. and Zhang, C. (2008) Content-based information fusion for semi-supervised music genre classification, IEEE Transactions onMultimedia, Vol. 10, No. 1, Pp.145–152.

[25]. Sundaram, S.S. and Narayanan, S. (2008) Audio retrieval by latent perceptual indexing, IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, Nevada, Pp. 49–52.

[26]. Tolonen,T. and Karjalainen, M. (2000) A computationally efficient multipitch analysis model, IEEE Trans. Speech Audio Processing, Vol. 8, Pp. 708-716

[27]. Virtanen, T. and Helen, M. (2007) Probabilistic model based similarity measures for audio query-by-example, Proceedings of IEEE Applications of Signal Processing to Audio and Acoustics, New Paltz, NY.

[28]. Wan, C. and Liu, M. (2006) Content-based audio retrieval with relevance feed-back, Pattern Recognition Letters, Vol. 27, No. 2, Pp. 85–92.