Music Genre Classification using Auto-Associative Neural Networks

Abhishek Ballaney, Suman Mitra, Anutosh Maitra
{abhishek_ballaney, suman_mitra, anutosh_maitra} @ daiict.ac.in

***Abstract*—Classification of musical genres gives a useful measure of similarity and is often the most useful descriptor of a musical piece. Principal Component Analysis (PCA) has been generally applied on raw music signals to capture the major components for each genre. As a large number of principal components are obtained for different genres, the purpose of applying PCA is not satisfied. This led to, in the proposed work, feature vector extraction directly from the music signal and building an alternative model to capture the feature vector distribution of a music genre. Timbre modeling is done using Mel Frequency Cepstral Coefficients (MFCCs). The modeling of the decision logic is based on Auto Associative Neural Network (AANN) where the models perform an identity mapping on the input space. The property of a five layer AANN model to capture the feature vector distribution is used to build a music genre classification system.**

***Index Terms*—AANN, MFCC, PCA**

I. Introduction

Music genre classification is crucial for the categorization of bulky amount of music content. Automatic music genre classification finds important applications in professional media production, radio stations, audio-visual archive management, entertainment and others. Although it is hard to precisely define the specific content of a music genre, it is generally agreed that audio signals of music belonging to the same genre contain certain common characteristics since they are composed of similar types of instruments and having similar rhythmic patterns. These common characteristics motivated recent research activities to improve automatic music genre classification [1, 2, 3, 4]. The problem is inherently challenging as the human identification rates after listening to 3 seconds samples are reported to be around 70% [5]. Genres of music are often determined by tradition and presentation as by the actual music. As the boundaries between genres remain fuzzy, it makes the problem of automatic classification a nontrivial task [1]. Both the tasks of feature extraction and classifier design of music genres are complicated, especially when the decision window spans over only a short duration, such as a couple of seconds. One can also expect to observe similarities of spectral content and rhythmic patterns across different music genre types, and with a short decision window misclassification and confusion rates increase.

II. Review of State of the Art

The focus of the survey is on the approaches followed at the feature, model and decision levels for music genre classification

A. *Features to represent genre information*

Music information can be described relatively accurately by higher level model based representations like MIDI and MusicXML [1]. Features for music signals are generally related to melody, harmony, rhythm or timbres. Timbre based features analyze the spectral distribution of the signal. These include temporal features like zero crossing rate (ZCR) and linear prediction coefficients (LPCs); energy features like root mean square energy of the signal frame and energy of the harmonic content of the power spectrum; spectral shape features like centroid, spread, skew ness, kurtosis, slope, roll-off frequency, variation and mel-frequency cepstral coefficients (MFCCs); and the perceptual features like loudness, sharpness and spread.

Harmony is the use and study of pitch simultaneity and chords in music. Melody is a series of linear events or a succession, such that it contains a change of some kind and perceived as a single entity. It includes patterns of changing pitches and durations. Cook et. al. [2] has explored such features for music genre classification.

Rhythm is the variation of the duration of sounds over time. It is inherent in any time dependent medium, but it is mostly associated with music, dance and poetry. It is clear that rhythmic content may be a dimension to consider when discriminating between music genres [1].

B. *Models for genre classification*

The state of the art provides three broad classification schemes. The first is the expert systems where a set of rule is used in the decision logic to define a particular genre. Pachet and Cazaly [3] analyzed existing music genre taxonomies and provided a few guiding principles for building such a taxonomy. They suggest some descriptors of genres like instruments, voice, rhythm or tempo.

There are other unsupervised classification techniques that cluster the data based on objective similarity measures. A music file is converted to a set of features and then comparisons are based on certain metrics. Shao et. al. [4] has used Hidden Markov Models (HMMs) to model the features over time. They cluster their music collection with agglomerative hierarchical clustering. Rauber et. al. [5] used growing hierarchical self-organizing map (GHSOM) to organize the sample music sugnals.

Supervised classification methods assume that the taxonomy of genres is known. Classifiers are trained by manually labeled data. Cook et. al. [2] has used K-nearest neighbor (KNN) and Gaussian Mixture Modeling (GMM) technique for music genre classification. Support vector machines (SVMs) have also been used to classify music genres in [6]. Soltau et al. [7] proposed explicit time modeling of temporal structure of music where a multi-layer perceptron is trained so that the activation of its hidden neurons becomes a representation of the input feature vector. Each hidden neuron is seen as an abstract musical event. A feature vector is formed by the sequence of such abstract events, which is then used in another neural network for making class decision.

C. Motivation for new models

It has been generally observed that the feature vector analysis for music genres are not suitable for obtaining a signature of a particular genre unless a large number of principal components are used at the time of classification. This defeats the very purpose of an automated classification system as identifying a reasonable number of major components to define a genre; especially when the signals are of short duration (not more than a few seconds), is rarely fullproof. The goal of any pattern classification is that, given the training data in terms of feature vectors of a class; a predefined model should be able to capture the feature characteristics for each class. Parametric model appears general enough to characterize the distribution of the given feature vectors, but the model is constrained by the fact that the shape of the components of the distribution is assumed to be specific, and that the number of mixtures are generally fixed a priori [8].

The logical reasoning by Ikbal et. al. [11] proves the ability of auto associative neural network (AANN) models to capture a nonlinear subspace. There is a proven relation between the feature vector distribution and the training error surface captured by an AANN model in the input feature space [8]. AANNs are nonlinear models, which are explored in the proposed work for music genre classification task.

III. Auto Associative Neural Network

An auto associative neural network (AANN) receives input from external sources, while its activation level is part of the final output produced by the network. External inputs arrive and activate some subset of nodes in the network while other nodes remain inactive. This pattern of active and inactive nodes represents knowledge to be stored by the network [12]. The degree of difference between the original input pattern and the output pattern produced by the network is a measure of the error in the reconstruction [12].

Kamp et. al. [11] showed that nonlinear hidden units in a three-layer AANN model do not provide better solution than the conventional principal component analysis (PCA). Addition of hidden layers before and after the compression layer projects the input data onto a nonlinear subspace [10]. Five layer models of AANN are used in the proposed model for dimension reduction by projection of input data onto nonlinear subspace captured by the network. The weights of the five-layer AANN model capture the distribution of the given data [8]. The second and fourth network layers have more neurons than the input layer. The third layer has less number of neurons than the first or fifth layer. The activation functions in the hidden layers are nonlinear, while those in the input and output layers are linear.

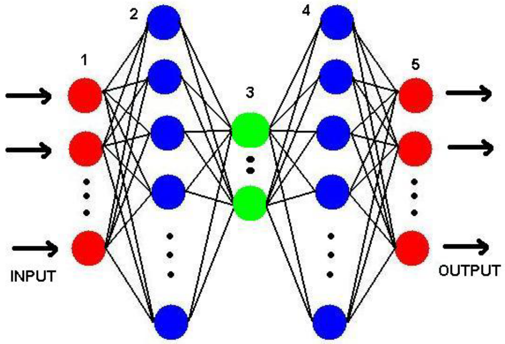

Fig. 1. 5 layer AANN model

IV. AANN Based Music Genre Classification System

In the present work, two separate AANN models are used to capture the feature vector distribution of each music genre. The database consists of 10 music genres, each having 100 pieces of music [2]. Development data consists of 900 music pieces (90 pieces from each of the 10 genres) and the evaluation data consists of 100 music pieces (10 pieces from each of the 10 genres). All the music pieces are sampled at 22.05 kHz. The training data consists of 900 songs of 30 seconds each, while the duration of the test piece varies between 5 to 30 seconds.

A. Feature Extraction

Genre information from music tracks can be extracted both at the higher and lower levels. The low-level features are the features extracted from short (10-30 ms) segments of the music signal. Most of the approaches focus on timbre modeling based on combinations of low-level descriptors [1]. In this work, spectral features represented by the MFCCs have been used.

The signal is now segmented into frames of 512 samples using a Hamming window in order to remove edge effects with a shift of 256 samples [2]. Then the Discrete Fourier Transform (DFT) of each frame is taken and only the logarithm of the amplitude spectrum is retained. The log magnitude is then weighted by a series of ‘Mel’ filter frequency responses whose center frequencies and bandwidths roughly match those of the auditory critical band filters. 30

filters are used to find the 5 MFCCs because they provide the best genre classification performance [2]. The energy in the STFT weighted by each mel-scale filter frequency response is calculated. Discrete cosine transform (DCT) is applied to de-correlate the original mel-scale filter log-energies.

B. *Generation of Genre Models*

The extracted features from the training data of a particular genre are used to train an AANN model using back propagation learning algorithm. Before training, mean and standard deviation of the training inputs and expected outputs are normalized to zero and unity respectively [13].

The structure of the AANN used here is 5L15N3N15N5L, where L denotes linear neurons and N denotes nonlinear neurons with tan-sigmoid as the activation function. The integer indicates the number of neurons in a layer.

The algorithm used to train the AANN models is batch gradient descent with momentum. The weights are iteratively adjusted by the following function:

$$\Delta w_{k+1} = \alpha w_k + (1-\alpha)\eta_k g W$$

It is known that the performance of the steepest descent algorithm can be improved if the learning rate changes during the training process [13]. So, the learning rate is made responsive to the complexity of the local error surface and it minimizes the training time [13]. Back propagation is used to calculate the derivatives of performance function Mean Square Error (MSE) with respect to the weight and bias variables X. Each variable is adjusted according to gradient descent with momentum,

$$d\mathbf{X} = \alpha d\mathbf{X}_k + \eta\alpha \frac{dMSE}{d\mathbf{X}}$$

The 10 AANN models are trained for 2000 epochs. MSE was below a predefined threshold and was steady at the end of the training stage.

C. *Classification*

During testing, the feature vectors in the form of five MFCCs are extracted from the test music piece and given to the all genre model to obtain a confidence score. The score of each model is defined as

$$C_k = \frac{1}{l}\sum_{i=1}^{l} e^{-D_i/\beta}$$

where k = 1, 2, …, 10 and l is the number of feature vectors of the test piece and

$$D_i = \frac{\left\|\mathbf{x}_i - \mathbf{y}_i\right\|^2}{\left\|\mathbf{x}_i\right\|^2}$$

where xi is the input feature vector of the model, yi is the output feature vector generated by the model, β is the temperature parameter for the training of the AANN model.

For classification, after extracting the features of the test music piece, the confidence score for each of the genre models are computed. The maximum confidence score identifies the genre of the song.

$$C = \arg\max_i[C_i]$$

where C is the class where the test piece belongs. As mentioned already, the test data consist of 500 songs over 10 genres with the test piece length varying from 5 to 30 seconds of duration.

D. *Results*

Table I shows the overall performance efficiency of the proposed model. The performance of the training and test datasets are averaged to obtain the overall percent efficiency of the system. The percentage efficiency is defined as the ratio of correct classifications to the total test cases. The size of the feature vector is 5 MFCCs. The β (Beta) value was varied from 0.2 to 5.0 to check if it has any appreciable impact on training accuracy. The test piece for the reported duration is selected from the central part of each music piece [9].

TABLE I
OVERALL PERFORMANCE

Overall % Efficiency				
Beta	5 sec	10 sec	20 sec	30 sec
0.2	30.89	33.30	35.11	35.11
0.4	30.99	33.40	35.21	35.10
0.6	31.30	33.20	34.91	35.20
0.8	31.60	32.90	35.10	35.29
1.0	31.50	32.60	34.90	35.30
2.0	31.90	32.60	34.70	35.49
5.0	31.39	32.50	34.20	34.99

It is observed that though the AANN apparently got trained with the weights adjusted to a steady value, the classification efficiency is poor. The major issue here was to identify the feature vectors correctly. In the next experiment, the size of the feature vector is increased to 15 MFCCs. Yet another set of neural networks of size 15L30N5N30N15L is trained for 2000 epochs and tested with above specified database. The efficiency of this set of networks is shown in Table II.

TABLE II
EFFICIENCY OF 15 INPUT AANN

% Efficiency		
Beta	5 second	
	Training	Testing
0.2	44.22	34.44
0.4	43.22	33.33
0.6	43.00	33.33
0.8	42.33	33.33
1.0	42.44	33.33
2.0	41.22	35.56
5.0	40.56	35.56

Though the performance of this set is improved, still it was considered inadequate.

Similar experiments were performed using high frequency MFCCs and STFTs. on only three classes of music, viz. Blues, Classical and Country. A set of three AANNs with 5-input and 10-input configuration each were trained. The results obtained in these experiments are shown in Table III.

TABLE III
EFFICIENCY OF 5 & 10 INPUT AANNS USING HF COEFFICIENTS

% Efficiency				
	10 coefficients		5 coefficients	
	Training	Testing	Training	Testing
MFCC	61.85	41.67	57.41	33.33
STFT	54.81	45.83	48.15	37.50

It is clear that as the number of inputs or features is increased, the performance of the system increases. The performance is also enhanced when the duration of the test piece is increased. Interestingly, even though music signals have significant high frequency components, incorporating high frequency features could not make appreciable improvement in the performance.

V. CONCLUSION

The work reported here explores an alternative model for music genre classification. Success of AANNs in certain domains was the primary driving force for this alternative modeling. However, the experimental results obtained thus far do not advocate the usage of the AANN model for the kind of problem investigated. The difficulties in identifying a proper signature vector for each genre mostly remain even in the alternative model specified. It is fairly certain that for the success of AANN models in music genre classification, determination of the input vector of the AANN is still an open research issue. The results clearly indicate that the size of the feature vector is a major parameter for proper training of the model.

ACKNOWLEDGMENT

We are obliged to Prof. George Tzanetakis, University of Victoria, Canada for sharing his music genre database.

REFERENCES

[1] Scaringella N., Zoia G., Mlynek D., “Automatic Genre Classification of Music Content”, IEEE Signal Processing Magazine, pp.133-141, Mar. 2006.

[2] Tzanetakis G., Cook P., “Musical genre classification of audio signals”, IEEE Trans. on Speech and Audio Processing, Vol.10, Issue 5, pp.293-302, Jul. 2002.

[3] Pachet F., Cazaly D., “A taxonomy of musical genres”, in Proc. Content Based Multimedia Information Access (RIAO), 2000.

[4] Shao X., Xu C., Kankanhalli M.S., “Unsupervised classification of music genre using hidden Markov model”, Proc. of ICME'04, Vol.3, pp.2023-2026, Jun. 2004.

[5] Rauber A., Pampalk E., and Merkl D., “Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarity”, Proc. 3rd ICMIR, 2002.

[6] Xu C., Maddage N.C., Shao X., Cao F., Tian Q., “Musical genre classification using support vector machines”, Proc. of ICASSP'03, Vol.5, pp.429-432.

[7] Soltau H., Schultz T., Westphal M., and Waibel A., “Recognition of music types”, Proc. IEEE ICASSP'98, vol. II, pp. 11371140.

[8] Yegnanarayana B., Kishore S.P., “AANN: an alternative to GMM for pattern recognition”, Neural Networks, Vol.15, pp.459-469, Apr. 2002.

[9] Lippens S., Martens J.P., De Mulder T., Tzanetakis G., “A comparison of human and automatic musical genre classification”, Proc. of IEEE ICASSP'04, Vol.4, pp.233-236, May 2004.

[10] Kramer M. A., “Nonlinear principal component analysis using auto-associative neural networks”, AIChE, Vol.37, pp. 233-243, Feb. 1991.

[11] Ikbal M. S., Misra H., Yegnanarayana B., “Analysis of Auto-associative Mapping Neural Networks”, in IJCNN 1999.

[12] Gluck M. A., Gateway to Memory, The MIT Press, 2001.

[13] Demuth H., Beale M., Neural Network Toolbox for Use with MATLAB, version 4.