Building a Generative Model for Affective Content of Music

Dr.B.Vinayagasundaram,
Department of Information Technology, MIT, Anna University, Chennai,India.
Rahul Mallik, Aravind M,
Department of Information Technology, MIT, Anna University, Chennai,India.
R.J.Aarthi,
Department of Information Technology, MIT, Anna University, Chennai,India.
Senthilrhaj S.
Department of Information Technology, MIT, Anna University, Chennai,India.

***Abstract*—In the Music Information Retrieval (MIR), music emotion recognition is one of the active research topics. Many methods have been proposed for modeling the emotions from the music excerpts. Most of the methods model the emotions as a deterministic hard label which fails to handle the issue of subjectivity. Some methods which model the emotions as a distribution lack the classification [1] (Recognition in this case) accuracy. In this paper, a novel method has been proposed which models the affective content as a Gaussian probability distribution model called AEG (Acoustic Emotion Gaussian) which is generative, flexible, interpretable and transparent. AEG model incorporates two GMM's: Affective GMM used for model emotion and Acoustic GMM for modeling the audio data. In addition to the acoustic feature (MFCC) and annotations from the listeners (Valence and Arousal),additional acoustic features, Raagas (melodic modes used in Indian Classical music) and emotions obtained from the sentiment analysis of lyrics helps to optimize and increase the classification accuracy and robustness of the AEG model.**

Keywords: Acoustic Emotion Gaussian, Raagas, Sentiment Analysis, Music Information Retrieval.

I.INTRODUCTION

MUSIC is used in our daily life for entertainment, mood management (happy, sad and agony etc.) and therapy relaxation. Personalized music emotion recognition[8] is another important aspect to be considered. Instead of assigning a hard label to an emotion, modeling the emotion as probability distribution (soft labels) in the V-A space will provide a interpretable, flexible, transparent and generative training model. Valence represents the different types of emotion and Arousal represents degree of intensity of the particular emotion. To build the novel generative model, Acoustic Emotion Gaussian(AEG) model has been proposed which uses two separate GMM's(Gaussian Mixture Model). Acoustic GMM for modeling the audio data and Affective GMM for modeling the emotional data. To optimize the performance of the training model, two major enhancements in the existing learning algorithm have been proposed. Predicting the emotion from the Raaga[5](melodic modes used in Indian Classical Music) of a music excerpt is one of the proposed enhancements. A Raaga is nothing but a melody constructed from a set of five or more notes from the fixed scale of seven notes. Each Raaga has separate emotion like joy, sad and anger that helps to analyze the sentiment of the music. Another one is to predict the emotion using Sentiment Analysis of lyrics[9]. Including more acoustic features other than the most widely used acoustic feature Mel Frequency Cepstral Coefficient (MFCC) [2] helps to make the AEG model robust and effective.

II. RELATED WORK

The relationship between the music and emotion is designed as the mathematical model. This is one of the active research topics in Music Information Retrieval (MIR)[11]. Lot of existing models dealt with the automatic annotation of music excerpts to the emotion. The existing models are discriminative and highly subjective and assigned a hard label for the excerpts. So, the proposed model will overcome these issues by soft distribution. Valence and Arousal system is one of the essential system that is used to model emotions. The existing models used emotional content of music fragment as a single point in Valence and Arousal space and used average (mean) from certain number of users without taking covariance into account. These methods decrease the accuracy of the model. V-A employs two models generally (Heatmap and Gaussian parameter approach). Heatmap uses r x r dimension to predict the emotion. Each cell in the r dimension represents emotion with intensity. The drawback of this approach is because of no relationship between the nearby cells. In this approach, emotion distribution is represented as Gaussian, mean and covariance is obtained by using Support Vector Regression (SVR). But this approach is also discriminative and does not utilize probability distribution. The base rate predictor trains the model using a prior Gaussian distribution as fixed for the predicted results for every test fragment. The prior Gaussian mean and covariance parameters are obtained using the joint probability distribution of valence arousal ratings (Annotation) of the music fragment in the training set. Thus, the base-rate predictor does not consider the audio features of the music fragment.

978-1-4673-9802-2/.16/$31.00©2016 IEEE

III. PROPOSED WORK

The proposed work is to model the music's affective content

(i) performs better for the subjectivity
(ii) probability distribution representation and
(iii) graphical model to extract acoustic features
(iv) represents emotion in Valence-Arousal space

The proposed model is AEG deals with subjectivity issue and it is generative. AEG overcomes the existing method drawbacks by the consideration of the acoustic features and the annotations of the music excerpt. AEG parameterizes the acoustic features as acoustic GMM which employs an unsupervised based (clustering) learning technique and parameterizes the annotation of the listeners in Valence Arousal space as affective GMM which employs a supervised based learning technique by considering acoustic Features (Acoustic GMM) and annotation from the listeners as the feature set. Raaga is the melodic modes used Indian Classical music. Each Raaga is a permutation of Swaras (music notes,7 in total). Each Raaga has an affective content associated with it.

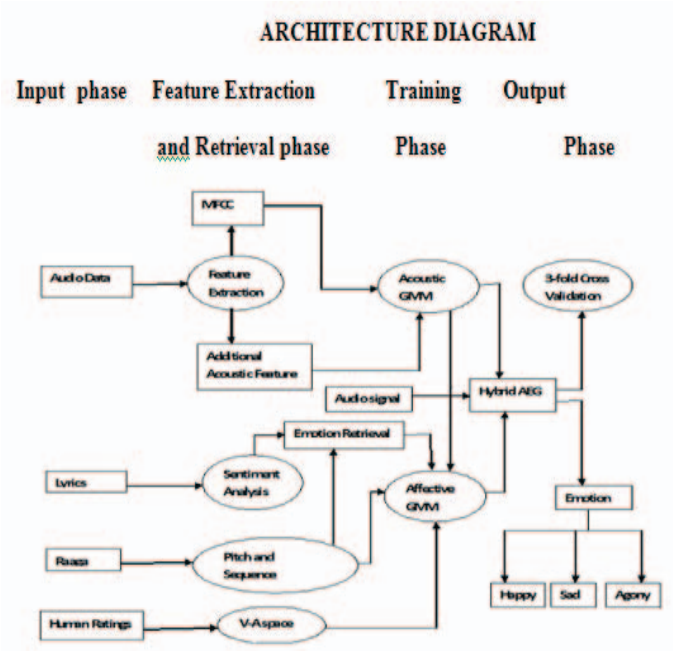

Fig. 1. Architecture of Raaga, Sentiment Human Annotation

Many methods have been proposed to identify the Raaga of music excerpts. One of the identification methods to recognize Carnatic Raaga from instrumental music piece is K-NN approach applied on TPMs. This system is the educational tool to music lovers. It also has various applications such as indexing song databases with Raagas which can be used as tools for music therapy/healing and song suggestion for radio programs. Sentiment analysis is the process of text analysis using various computational methods. There are several supervised machine based learning classifiers for the sentiment classification task. The best performance has been obtained with the emerging technology known as extreme learning Algorithm for the solutions to the feed-forward networks[6], not restricted to single or multi- hidden-layer neural networks, radial basis function networks, and kernel learning.

This model is non-specific and basis for emotion and Raaga, which is the melodic modes used Indian Classical music. Each Raaga is a permutation of Swaras (music notes,7 in total).Each Raaga has an affective content associated with it. Many methods have been proposed to identify the Raaga of music excerpts. One such method is an identification method K-NN approach applied on TPMs to recognize the Carnatic Raaga on which an instrumental music piece is based on. This system will serve as an educational tool to music enthusiasts.

It also has various applications such as indexing song databases with ragas which can be used as tools for music therapy/healing and song suggestion for radio programs. Sentiment analysis is the process of

(i) computational identification
(ii) categorizing judgment expressed in text
(iii) view can be positive, negative or neutral

There are several supervised machine based learning classifiers for the sentiment classification task. The best performance has been obtained with the Extreme Learning Machine (ELM), an emerging learning technique. Fig.1 represents the architecture of the entire system with Raaga, sentiment analysis, human annotations and song as input phase, modules for feature extraction phase, Acoustic and Affective GMM to get a hybrid model in the training phase and 3-fold cross validation in the testing phase which predicts the final emotion of the music excerpt.

The input phase shows the various types of data to the acoustic and affective model. The audio data has been analyzed by applying feature extraction tool like MFCC(Mel Filter Cephstral Coefficient). Various attributes are made to undergo the training phase, so as to predict the oncoming various acoustic features.

By means of EM (Expectation Maximization) algorithm the Affective Gaussian Model is built with its own mean vector and variance matrix.

The emotion of human is predicted by means of Valence and Arousal space.

The following models are implemented as part of proposed work.

1. AEG training model

The training module for AEG using the DEAP(Dataset for Emotion Analysis using EEG and Physiological signals).This module involves building two models namely Acoustic GMM and Affective GMM. Refer Fig.2 for the breakdown of the modules. Initially, the features known as MFCC (Mel Frequency Cepstral Co-efficient) are extracted

from the music excerpt. The excerpt is segmented into short frames.

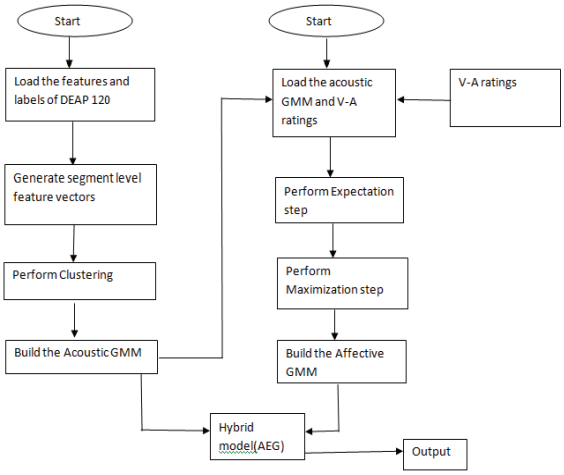

Fig. 2. Models of Acoustic GMM and Affective GMM

Using the mean and covariance parameters the Acoustic GMM is built first. Using the acoustic prior and labels provided by humans the Affective GMM is built using Expectation Maximization Algorithm. The Stopping criteria for training are either threshold computation or fixing a number of iterations for convergence of mean, covariance parameters and maximization of the log-likelihood function. Fig. 3 shows the valence and arousal of various music excerpts.

This shows the valence and arousal rating by various people for different Music excerpt. Based on the valence and arousal index the mood of the listener can be identified.

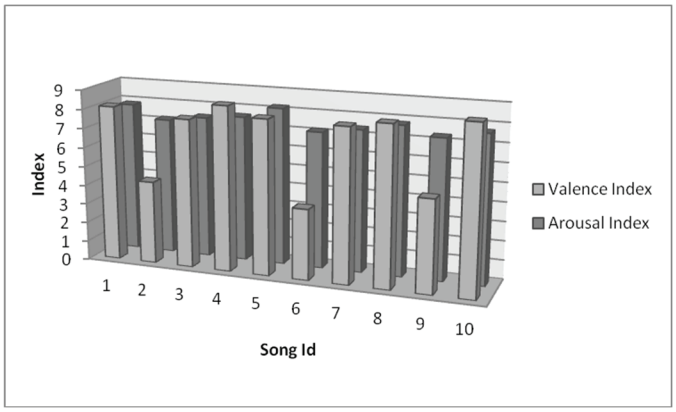

Fig. 3. Valence Arousal output

The analogy between discriminative and generative model is to determine the language that is spoken by someone, generative approach is to learn each language and determine which language the speech belongs to.

To model the random variables, X - a smaller time frame of acoustic features of the audio content $\{x_1, x_2, \ldots x_T\}$, $x_t \in R^M$. The position of the excerpt on the continuous valence-arousal emotion space $y \in R^2$, and the associated discrete dormant topic is $z \in \{1,2,\ldots K\}$.

The graph X→z→y, shows that emotion y is independent of audio features X for a particular latent z.

The Gaussian distribution of a small time frame x is

$$p(x \mid z = k) \sim \mathcal{N}(m_k, S_k) \tag{1}$$

The corresponding probability density function for x,

$$p(x) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}\left(x \middle| m_k, S_k\right) \tag{2}$$

where π_k, m_k, S_k are linked with the k^{th} dormant topic.

The posterior probability for an observed frame x is computed by ,

$$p(z = k|x_t) = \frac{\pi_k \mathcal{N}(x_t \mid m_k, S_k)}{\sum_{h=1}^{K} \pi_h \mathcal{N}(x_t \mid m_h, S_h)} \tag{3}$$

The fragment based posterior probability of z=k for a given X, can be approximately equated to the average of the time-frame based posterior probabilities,

$$p(z = k|X) \approx \frac{1}{T} \sum_{t=1}^{T} p(z = k|x_t) \tag{4}$$

The distribution of y given z=k is Gaussian

$$p(y|z = k) \sim \mathcal{N}(\mu_k, \sum_k) \tag{5}$$

where μ_k and $\sum_k$ are the corresponding parameters of k^{th} latent topic.

The marginal distribution of y is given by

$$p(y|X) = \sum_k p(y|z = k) p(z = k|X)$$

$$= \sum_k \mathcal{N}(y \mid \mu_k, \sum_k) \, p(z = k|X) \tag{6}$$

Where $p(z|X^{(i)})$ is the acoustic prior is computed for each training fragment.

The parameters $\Theta \equiv \{\pi_k, m_k, S_k, \mu_k, \sum_k\}_{k=1}^{K}$ can be estimated using maximum likelihood function, given by the following expression,

$$\hat{\theta} = \operatorname{argmax} \sum_{i=1}^{N} \log p\left(Y^{(i)} \middle| X^{(i)}, \theta\right) \tag{7}$$

where N is the number of training audio fragment and i is the i^{th} fragment in the total training set.

The log likelihood function is given by

$$L = \log \prod_{i=1}^{N} \prod_{j=1}^{U^{(i)}} p(y_j^{(i)} | X^{(i)}) \tag{8}$$

$$\hat{L} = \sum_{i,j} log \sum_k \mathcal{N}(y_j^{(i)}|\mu_k, \sum_k)\ p\left(z = k\middle|x_j^{(i)}\right) \quad (9)$$

is the annotation prior of the model.

At E-step the posterior probability for z=k given annotation prior $y_j^{(i)}$ is

$$p(z = k|y_j^{(i)}) = \frac{p(z=k|X^{(i)})\mathcal{N}(y_j^{(i)}|\mu_k, \sum_k)}{\sum_h p(z=k|X^{(i)})\mathcal{N}(y_j^{(i)}|\mu_h, \sum_h)} \quad (10)$$

At M-step we update mean vector and covariance matrix

$$\mu_k' \leftarrow \frac{\sum_{i,j} y_j^{(i)} p(z=k||y_j^{(i)}) y_j^{(i)}}{\sum_{i,j} y_j^{(i)} p(z=k||y_j^{(i)})} \quad (11)$$

$$\Sigma_j^{(i)'} \leftarrow \frac{\sum_{i,j} y_j^{(i)} p(z=k||y_j^{(i)}) \left(y_j^{(i)} - \mu_k'\right)(y_j^{(i)} - \mu_k')^T}{\sum_{i,j} y_j^{(i)} p(z=k||y_j^{(i)})} \quad (12)$$

The EM algorithm maximizes the Log likelihood until the convergence. Algorithm 1 explains the initial stage of the moods of GMM.

ALGORITHM 1: To fit the Affective GMM

INPUT: Acoustic prior: $\{(p(z|X^{(i)})\}_{i=1}^{N}$

Annotation prior: $\{\gamma_j^{(i)}\}_{i=1,j=1}^{N,U^{(i)}}$

Initial model : $\{\mu_k^0 = \mu_L, \Sigma_k^0 = \Sigma_L\}_{k-1}^{K}$

Repeat the steps from 1 through 4 till it reaches the maximum value R or threshold of stopping ratio ┌

Output parameters : $\{\mu_k^r, \Sigma_k^r\}_{k-1}^{k}$

1. Initialize r to 0 and evaluate L_0 using equation (9);
2. Do until
   2.1. Calculate the posterior probability using equation(10) with $\{\mu_k^r, \Sigma_k^r\}_{k-1}^{k}$;
   2.2. Increment r by 1
   2.3. Update $\{\mu_k^r, \Sigma_k^r\}_{k-1}^{k}$; using equation (11) and (12)
   2.4. Compute L_r
3. Until r=R or $(L_r - L_{r-1})/|L_{r-1}| <$ ┌
4. Let $\mu_k' \leftarrow \mu_k^r$ and $\Sigma_k' \leftarrow \Sigma_k^r$

2. *Performing 3-fold cross validation of the AEG model*

Cross validation is a model evaluation method compared to residuals. The drawback of residual evaluations is that for the new data set, test set will give the output as the prediction based on how the data set has been trained. In order to overcome this problem only partial data set is used when training a learner. The whole data set is divided into chunks, some of them are used as training data set and others are kept as test set. Predictions based on training data set will be tested against test set. The above depicts the idea for a whole class of model evaluation methods.

(i) AKL Divergence-Acoustic Diversity

The divergence of emotions of the various excerpts measured by AKL (Average Kullback Leibler) and also used to evaluate the moods of the corpus. This is measured using AKL(Average Kullback Leibler) AED-Annotation Diversity divergence. AKL divergence for our AEG model is 0.453 ± 0.456.

(ii) AED(Average Euclidean Distance)

AED is the measure of the average Euclidean distance between the mean vectors of different music excerpts. AED for our training model was found out to be around 1.145±0.516.Thus, smaller and average of values of divergence indicate better performance of the training model.

Gaussian $g^{(i)}$ labels the notes Y(i) for every individual song of an emotion collection of records L as,

$$D_{KL}\langle g_A|g_B\rangle = \frac{1}{2}\left(tr\left(\sum A\,\Sigma_B^{-1}\right) - \log\left|\left(\sum A\Sigma_B^{-1}\right|\right.\right.$$
$$\left.\left. + (\mu_A - \mu_B))^T \Sigma_B^{-1}(\mu_A - \mu_B) - d\right) \quad (13)$$

Where $g_A \cong N(\mu_{,A}, \sum_A)$, $g_B \cong N(\mu_{,A}, \sum_B)$, in a 2 dimensional valence-arousal space. Accordingly, the pair wise KL divergence (PWKL) of L can be defined as

$$PWKL(L) = \frac{1}{N_{PW}} \sum_{i \neq j} D_{KL}(g^{(i)}||\ g^{(j)}) \quad (14)$$

Where $N_{PW} = \frac{N(N-1)}{2}$ denotes the number of pairs in L.

ALGORITHM 2: Regularization of Covariance parameters

INPUT: Covariance parameters $\{\sigma_{ij}\}^{d,d}_{i,j=1}$
1. For each diagonal element σ_{ij} do
2. If σ_{ii} lies between -0.01 and 0
3. End For
4. For each element, other than diagonal elements $\sigma_{ij(i \neq j)}$ do
   4.1. if $\sigma_{ij} > \sqrt{\sigma_{ii}}\sqrt{\sigma_{ij}}$ then $\sigma_{ij} < -\sqrt{\sigma_{ii}}\sqrt{\sigma_{ij}}$;
   4.2. if $\sigma_{ij} < -\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}$ then $\sigma_{ij} < --\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}$;
5.End

Algorithm 2 makes the predicted parameters as valid one.

3. *Comparison with existing methods*

There are two major existing methods for music emotion recognition,

- **(i)** Base-rate prediction
- **(ii)** Support-Vector Regression method

The above methods are compared with Gaussian parameter approach, hybrid of affective and acoustic GMM.

Table 1. Performance metrics for DEAP dataset using AKL AED methods

Dataset	PWKL	Method	AKL	AED
DEAP 120	1.194	Base-rate(E)	0.759±0.744	1.405±0.612
		SVR(E)	0.530±0.502	1.212±0.587
		AEG(P)	0.453±0.456	1.145±0.516

The PWKL of DEAP120 dataset is evaluated by Base-rate method, which is probabilistic, uses the fixed prior Gaussian for every music excerpt. The SVR approach is used to optimize the Gaussian parameters. Algorithm 2 explains how to validate the covariance parameters. The Table 1 shows the Euclidean distance measurement and KL divergence of the music excerpts using various methods. The data for AEG shows smaller deviations compared to the two other methods

4. AEG personalization

This module is used to personalize the already built AEG training model based on the topic posteriors of 40 music clips and the ratings provided by the listeners. This module adjusts the Gaussian distribution parameters to customize the AEG model according to every listener. The Fig. 5 depicts the output of personalized AEG model.

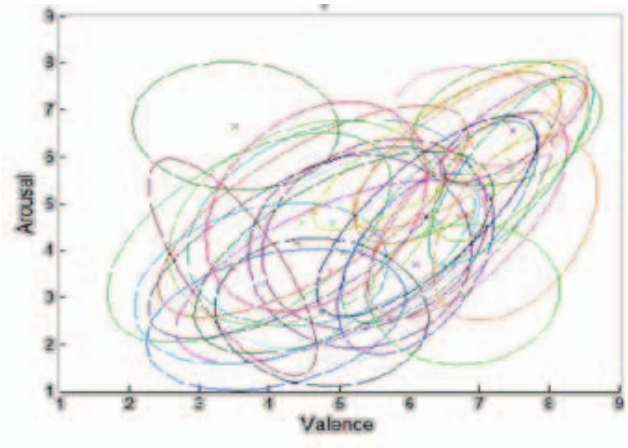

Fig. 4. Valence arousal for personalised affective GMM for number of iterations

IV. FUTURE RESEARCH

I. SENTIMENT ANALYSIS

Sentiment Analysis[7] is the process of categorizing opinions. It outputs the attitude in a piece of text represents positive or negative or neutral emotion. Sentiment Analysis[9] can be used as an additional feature for Affective GMM. The Best performance has been obtained using Extreme Learning Machine-ELM algorithm.

Extreme learning Machines are

(i) Feed forward neural network
(ii) Input node to hidden node connected by weights(randomly assigned)
(iii) Single step learning from hidden node to output
(iv) Linear model performance faster than back propagation.

The Fig.5 is the breakdown of sentiment analysis model. Using any of the classifiers like Naïve Bayes classifier, decision tree classification to identify the emotion of the lyrics. Later that can be combined with music excerpt model.

It is implemented using NLTK (Natural Language ToolKit). Million Song Dataset has been used as a benchmark dataset.

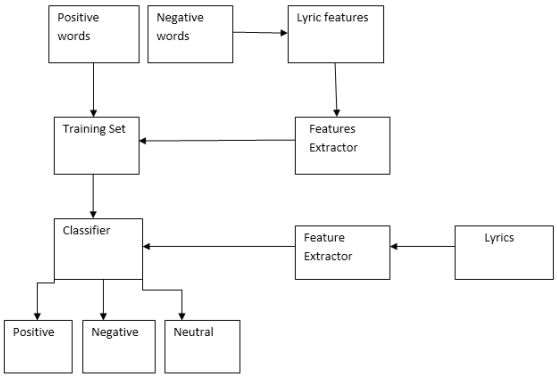

Fig. 5. Architecture of Sentiment Analysis

The ELM training model is given by,

$$\dot{Y} = W_1 \sigma W_2 X \quad (15)$$

where W_1 is the weight matrix from input to the hidde layer, σ refers activation function, and W_2 is the weight matrix from hidden to output layer.

(i) Fill W_1 can be noise signal which is of Gaussian,
(ii) Estimate W_2 by least-squares fit of response variables Y, calculated using the pseudo inverse, given a matrix X:

$$W_2 = \sigma(W_1 X) + Y \quad (16)$$

II. RAAGA IDENTIFICATION

Each music excerpt is based on a Raaga (Indian Classical Music) which has affective content associated with it. There are 72 base Raagas (Melakartha) and many derived (Janya) Raagas from the base Raagas. Each Raaga contains an ascent (Arohana) and a descent (Avarohana)[5].Each Raaga is represented as a Transition Probability Matrix (TPM). First step is to extract the melody using a tool called Audacity and convert the polyphonic signal to homophonic signal using the plug-in

Melodia. A Knowledge-based containing TPM's of Melakartha Raagas, Janya Raagas and relationship between Melakartha and Janya Raagas (one to many or one to one). Algorithm have been proposed to identify the pitch (Scale) and Swara sequence of the melody, compute the TPM distance between input Swara sequence and the knowledge based and classify it as either a base Raaga (Melakartha Raaga) or derived Raaga (Janya Raaga). Fig. 6 shows the flow diagram of feature extraction of Raaga from the given audio data and comparing with the Raaga database.

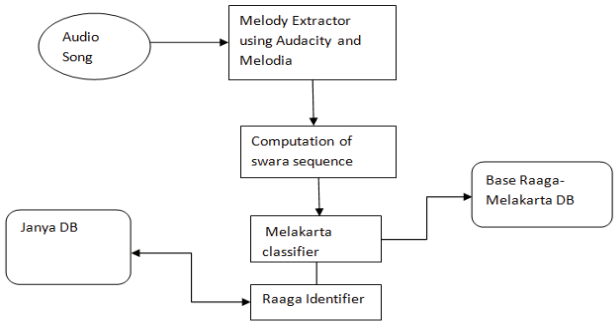

Fig. 6. Flow Diagram of Raaga Identification

V. CONCLUSION

Raaga can be added as future work. Raaga helps in to optimize the performance of the AEG model as it is one of the key features that carry the affective content of the music excerpt. The affective emotional model has been analyzed and can be extended to the stream of data from the online audio data as a future work by applying the machine learning algorithms like Hidden Markov Model (HMM). The interdisciplinary multimodal module can be developed combining the sentiment analysis, Raaga from the lyrics of the songs.

REFERENCES

[1] Konstantin Markov, Tomoko Matsui, "Music Genre and Emotion Recognition Using Gaussian Processes", IEEE,Vol. 2, July 2014.
[2] Norbert Braunschweiler and Mark J. F. Gales, Fellow, IEEE, Langzhou Chen, Member, IEEE,"Speaker and Expression Factorization for Audiobook Data: Expressiveness and Transplantation".
[3] A. J. Lonsdale and A. C. North, "Why do we listen to music? Uses and gratifications analysis," Brit. J. Psychol., vol. 102, pp. 108–134, 2011.
[4] J.Kleinberg and B.Scholkopf,"Information Science and Statistics by M Jordan".
[5] B.Tarakeswara Rao, S. Chinnam, M.Gargi, "Automatic Melakarta Raaga Identification System in Carnatic Music", International Journal of Advanced Research in Artificial Intelligence, Vol.1, No.4, 2012.
[6] C. M. Bishop. Pattern Recognition and Machine Learning. Springer-Verlag New York, Inc., 2006.
[7] W Medhat, A Hassan, H Korashy,"Sentiment analysis algorithms and applications: A survey" - Ain Shams Engineering Journal, 2014.
[8] J.C. Wang ; Inst. of Inf. Sci., Taipei, Taiwan ; Y. H. Yang ; H. M. Wang ; S. K. Jeng,"Modeling the affective content of music with a Gaussian Mixture Model ", IEEE Transactions on Affective Computing (Volume:6 , Issue: 1).
[9] Soujanya Poria , Erik Cambria , Newton Howard, Guang-Bin Huang, Amir Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content", NeuroComputing Volume 174, Part A, 22 January 2016, Pages 50–59.
[10] C.C.Chang and C.J.Lin, LIBSVM:A library for support Vector Machines, ACM Trans. Intel Syst.Technol.,vol.2,pp.27:1-27:27,2011.
[11] E.M.Schmidt and Y.E.Kim, " Prediction of time varying musical mood distributions from audio", in Proc. Int. Soc. Music Inf. Retrieval Conference, 2010, pp.465–470.