

Project Report on

**""""""Music Emotion Recognition using Acoustic
""""""Gaussian Mixture Model**

Submitted in partial fulfillment of the requirements

of the degree of

BACHELOR OF ENGINEERING

in

ELECTRONICS AND TELECOMMUNICATION

by

Rutuja Girmal (Roll No.36)

Jitali Kamat (Roll No. 46)

Sayali Martal (Roll No. 56)

Pooja Nair (Roll No. 68)

Under the guidance of

Mr. Santosh Chapaneri



Department of Electronics and Telecommunication Engineering

St. Francis Institute of Technology, Mumbai
University of Mumbai
(2017-2018)

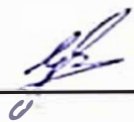
CERTIFICATE

This is to certify that the project entitled "Music Emotion Recognition Using Acoustic Gaussian Mixture Model" is a bonafide work of "Rutuja Girmal (36), Jitali Kamat (46), Sayali Martal (56) and Pooja Nair (68)" submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of Bachelor of Engineering in Electronics and Telecommunication Engineering.



(Mr. Santosh Chapaneri)

Project Guide



(Dr. Gautam Shah)

HOD (EXTC)



(Dr. Sincy George)

Principal

Declaration

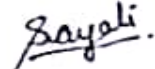
We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included; we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in this submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.



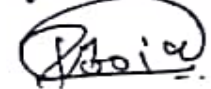
Rutuja Girmal

J. G. Kamat

Jitali Kamat



Sayali Martal




Pooja Nair

Date: 16/4/18

Project Report Approval for B.E.

This project entitled '**Music Emotion Recognition Using Acoustic Gaussian Mixture Model**' by Rutuja Girmal, Jitali Kamat, Sayali Martal and Pooja Nair is approved for the degree of *Bachelor of Engineering in Electronics and Telecommunication* from University of Mumbai.

Examiners:

Sandesh Chavan  23/4/18

PROF. ASHISH VANMAU 

Date: 23/4/18

Place: Borivali

ABSTRACT

Music conveys emotions/moods and arouses them in the listener which is why it is fascinating for people across the world. The music libraries have grown immensely and are easily available due to their digitization in recent times. This has led to a rapid increase in the music information retrieval research for automating systems in order to search and organize music and the related details. Common ways for search and retrieval are using genre or artist for sorting of songs which is easier to quantify and hence received more recognition in music information retrieval. In this work, we describe a generative approach for modelling music emotion while concentrating specifically on the valence-arousal (VA) dimensions of emotion. This generative model, called Acoustic Emotion Gaussians (AEG), considers the subjectivity of interpretation of emotion in a better manner by using probability distributions in the valence-arousal space thus making it possible to customize an emotion-based music information retrieval system. It generalizes the mapping of music to emotion. The model can also incorporate additional details such as user feedback using adaptation techniques. It learns a collection of latent feature classes from the available data to connect the acoustic and the affective Gaussian Mixture Models that are associated with the acoustic and the emotion spaces respectively. The entire process is easy and straight forward. It involves light-weight computations for prediction of emotion and hence can be used in real time for analyzing the distributions of an audio. Thus, this framework can be used for various applications including both general and personalized recognition and music retrieval systems based on emotion.

TABLE OF CONTENTS

1 Introduction.....	1
1.1 Motivation.....	2
1.2 Scope of the Project	2
1.3 Organization of the Project	3
2 Preliminaries	4
2.1 Introduction to Probability Distribution.....	4
2.2 Gaussian Distribution.....	5
2.3 Gaussian Mixture Model.....	6
2.4 Expectation Maximization Algorithm	9
2.4.1 Applications	13
2.4.2 Limitations	13
2.5 Support Vector Regression (SVR).....	13
2.6 Vector Quantization	15
2.7 AKL and AED	17
2.7.1 Average KL Divergence (AKL)	17
2.7.2 Average Euclidean Distance (AED)	18
2.8 MIR ToolBox.....	19
3 Literature Survey.....	22
3.1 Classification of Emotions.....	22
3.1.1 Categorical Approach	22
3.1.2 Dimensional Approach	24
3.2 Valence and Arousal	25
3.3 Approaches for modeling Valence and Arousal	26
3.3.1 VA Point Approach.....	26
3.3.2 Heatmap Approach	27
3.3.3 Gaussian Parameter Approach.....	27
3.4 Related Work	28
4 Music Emotion Recognition using Acoustic Gaussian Mixture Model.....	46
4.1 System Block Diagram:	46
4.2 Fitting the Acoustic GMM.....	48
4.3 Model the Annotation Prior	49
4.4 Fitting the Affective GMM.....	49
4.4.1 Singularity Issue in Learning the Affective GMM.....	51
4.5 Predicting Emotions.....	52
5 Results and Discussion.....	53
5.1 Histogram.....	53

5.2 AMG1608 Dataset	54
5.3 Feature Extraction	55
5.4 Generation of Posterior probabilities	55
5.5 Affective GMMs	56
5.5.1 Affective GMM using AEG.....	56
5.5.2 Affective GMM using VQ	57
5.6 Cross Validation.....	59
5.7 Performance metrics	60
5.8 Predicted Gaussians Using AEG and VQ	63
6 Conclusion	64
6.1 Conclusion	64
6.2 Future Scope	64
Appendix	65
Timeline Chart of the Project.....	65
References	67

List of Figures

Figure No.	Title	Page No.
2.1	Histogram plots of the mean of N uniformly distributed numbers for various values of N . As N increases, the distribution tends towards a Gaussian.	5
2.2	Plot of the standard normal probability density function	6
2.3	(a) Unclustered Data, (b) Fit with one Gaussian distribution, (c) Fit with Gaussian mixture model with three components	8
2.4	Illustration of the EM algorithm	11
2.5	Clustering using a Gaussian mixture model. Each color represents a different cluster according to the model.	12
2.6	Support Vector Regression (prediction) with different thresholds ε .	15
2.7	Vector Quantization	16
2.8	Block diagram of generation of 13D MFCC	20
3.1	Categorical approach	22
3.2	Hevner's model	23
3.3	Thayer's model	24
3.4	Russell's model	25
3.5	VA space	26
3.6	VA point approach	27
3.7	Participants mean PAD ratings for the 10 songs	30
3.8	System block diagram	31
3.9	A schematic diagram of the personalized system	33

3.10	Time-varying emotion distribution regression results for three example 15-second music clips (markers become darker as time advances)	35
3.11	Emotion space heatmap prediction using conditional random fields	36
3.12	Schematic diagram of emotion distribution prediction	37
3.13	Contours representing the distribution of annotation in case of static (left) and dynamic (right) annotations.	38
3.14	Qualitative illustration of the performance of PMER	40
3.15	Block diagram of Deep GP	41
3.16	Fixed aggregation vs. proposed adaptive aggregation of Gaussian process regressors for music emotion recognition.	43
4.1	System block diagram	46
5.1	Histogram of experimental annotations	53
5.2	Histogram obtained from AMG1608 dataset	54
5.3	Acoustic Posterior Probabilities	56
5.4	Affective GMMs using AEG (a) $K=64$ (b) $K=128$ (c) $K=256$	57
5.5	Affective GMMs using VQ (a) $K=64$ (b) $K=128$ (c) $K=256$	58
5.6	Threefold cross validation	60
5.7	AKL	61
5.8	AED	62
5.9	Predicted Gaussians using AEG and VQ	63

List of Tables

Table No.	Title	Page No.
2.1	Pros and cons of GMM	9
3.1	Comparison of performance of ALL, Psy15, RRF	32
3.2	Map of the Emotion Labels	44
3.3	Experimental Result of Dataset-2	45
5.1	AKL	60
5.2	AED	61

List of Abbreviations

AED	Average Euclidean Distance
AEG	Acoustic Emotion Gaussian
AKL	Average Kullback-Leibler Divergence
EM	Expectation Maximization
GMM	Gaussian Mixture Model
MFCC	Mel Frequency Cepstral Coefficients
VA	Valence Arousal
VQ	Vector Quantization

Chapter 1

Introduction

Music plays a significant role in enhancing an individual's life. It sometimes even imparts a therapeutic approach and is an important medium of entertainment for music lovers and listeners. The escalation in the number of MP3 players and the sudden increase in the amount of digital music content calls for different approaches for the organization and retrieval of music to fulfill the ever-increasing demand for easy and beneficial information access. Almost every audio is created to convey emotion, music organization and retrieval by emotion is a sensible method of accessing music information. Efforts have been made worldwide in the field of music information retrieval to train a machine to automatically recognize the emotion of a music signal.

Music Emotion Recognition describes the consideration of subjective nature of emotion perception in the development of automatic music emotion recognition (MER) systems. Music emotion recognition (MER) identifies the presence of the inherently emotional expression of people for an audio clip. MER is beneficial in music retrieval, music understanding, and other music-related applications. As volume of online musical contents has expanded rapidly in recent years, demands for retrieval by emotion have come up lately. The process of deciding the emotional content in an audio computationally, involves research in various domains such as machine learning, signal processing, musicology, cognitive science, auditory perception etc. There are very few well developed emotion models for giving the details of music emotion. This poses a difficulty in assessing the automation of music emotion detection. Low clarity associated with the systems dedicated to emotion recognition based on acoustic features leads to difficulty in the interpretation of data generated using this process.

Due to the subjective nature of emotion interpretation, there is a need for customization of music retrieval or recommendation systems based on emotion. The concept of Acoustic Emotion Gaussians (AEG) for processing of emotion-based audio information and retrieval is to consider the subjective nature of emotion perception in a better way by representing the possible affective responses to an audio piece with probability distributions. Emotions are mapped using valence (pleasure or displeasure) and arousal (activation, energy level). Any point in the valence-arousal (VA) space can be viewed as a specific emotion state. AEG is a parametric, probabilistic model and can integrate personal information of a particular individual via model adaptation techniques to make custom predictions. The personal

information of a user may include the profile of the user, transaction records, personal emotion annotation, listening history, relevance feedback, etc.

1.1 Motivation

Music is used to modulate or convey emotion and is considered to be superior to language. Our day to day life consists of emotions at every stage. Now-a-days, a variety of music players are available for users that have advanced features such as fast forward, reverse, local playback, customizable playback speed, sorting according to genre and/or artist, modulation of volume etc. The basic requirements of the user are satisfied with these features, but the user still has to manually browse through the list of songs in his/her device and select music to create a playlist to suit the current mood and emotional experience of the user. This is the disadvantage of the traditional music players. There is a lack of instinctive playlist generation tools for music listeners. The available digital music libraries are constantly growing making it more difficult to recall a particular song in the music library or to create a playlist to suit a specific event. It is labour intensive and time consuming for listeners to manually select songs suiting a particular mood or occasion. In addition, there is a huge variety of our music ranging from various albums/artists/composers which is heavily influenced by mood. Therefore, there is a need for music recommendation based on the mood of the user.

1.2 Scope of the Project

We intend to create a system that allows user to simply select a mood from a VA space (so that a variety of moods can be selected) and create a playlist from the available song database in the user's system without requiring an internet connection. This system can be used by shop owners to attract certain clients and by ad films requiring a highly memorable and positive emotion invoking music for their products. It can also be applied in games to invoke moods such as excitement, danger, fear, victory and happiness.

1.3 Organization of the Project

Chapter 2 describes the preliminaries of our project that includes Gaussian Distribution, GMM, EM algorithm, MIR toolbox. It also describes two other methods SVR and VQ which can be used instead of AEG. The performance metrics-AKL & AED are also discussed.

Chapter 3 describes the categorical approach that includes the Hevner's model and dimensional approaches that includes Thayer's model and Russell's model which are used to classify emotions. It also specifies the various approaches for modelling valence and arousal which include VA point approach, Heatmap approach and Gaussian parameter approach. It also describes the literature review of various methods used for music emotion recognition and classification.

Chapter 4 explains the overview of the system and the project methodology which includes fitting of the Acoustic and Affective GMMs followed by emotion prediction.

Chapter 5 includes a description of the dataset AMG1608 and interprets the histogram obtained from it. It also interprets the results obtained for Acoustic and Affective GMMs and compares the performance metrics obtained by using AEG and VQ.

Chapter 6 concludes the project work and mentions the future advancements which can be incorporated.

Chapter 2

Preliminaries

2.1 Introduction to Probability Distribution

A probability distribution is a mathematical function and gives the probabilities of occurrence of different possible outcomes in an experiment. It gives the details of a random phenomenon in terms of the probabilities of events. The results of an experiment or survey can be included in the examples of random phenomena. A probability distribution is defined with the help of an underlying sample space, which is the set of all possible outcomes of the random phenomenon being observed. The sample space may be a collection of real numbers or a higher-dimensional vector space, or it may be a list of non-numerical values; for example, the sample space of a coin flip would be {heads, tails}.

There are two types of probability distributions. A discrete probability distribution is applicable to the scenarios where the set of possible outcomes is discrete, such as a coin toss or a roll of dice and can be encoded by a discrete list of the probabilities of the outcomes, known as a probability mass function. A continuous probability distribution is applicable to the scenarios where the set of possible outcomes can take on values in a continuous range (e.g. real numbers), such as the temperature on a given day) is typically described by probability density functions (with the probability of any individual outcome actually being 0). The normal distribution is a common example of continuous probability distribution.

A probability distribution whose sample space is the set of real numbers is called univariate, while a distribution whose sample space is a vector space is called multivariate. A univariate distribution gives the probabilities of a single random variable taking on various alternative values; a multivariate distribution (a joint probability distribution) gives the probabilities of a random vector—a list of two or more random variables—taking on various combinations of values. Important and commonly encountered univariate probability distributions include the binomial distribution and the normal distribution. The multivariate normal distribution is a commonly encountered multivariate distribution [1].

2.2 Gaussian Distribution

In probability theory, the normal (or Gaussian) distribution is a very common continuous probability distribution. Normal distributions are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known and are important in statistics. The normal distribution is also sometimes called the bell curve [2]. The normal distribution, also known as the Gaussian or standard normal distribution, plots all of its values in a symmetrical fashion, and most of the results surround the probability's mean. Values are equally likely to plot either above or below the mean. Grouping takes place at values close to the mean and then tails off symmetrically away from the mean [3].

A normal deviate is a random variable with a Gaussian distribution. If the number of events is very large, then the Gaussian distribution function may be used to describe physical events as indicated in Fig. 2.1. The Gaussian distribution is a continuous function which approximates the exact binomial distribution of events.

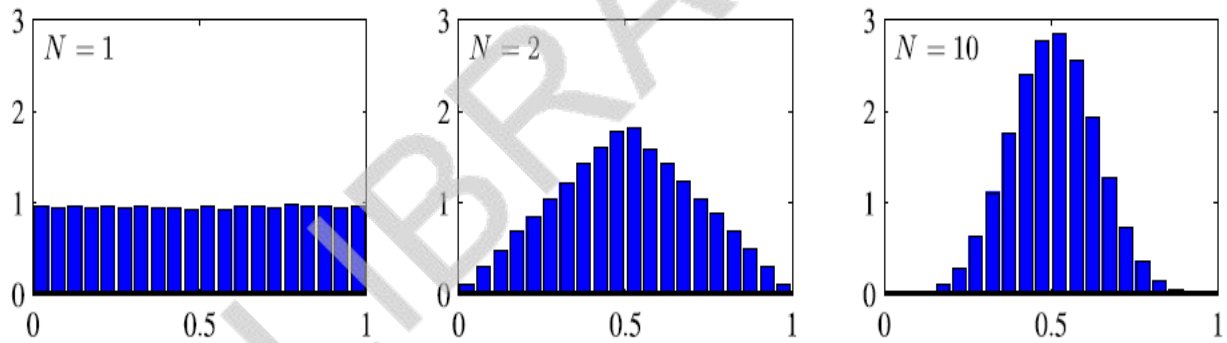


Fig. 2.1 Histogram plots of the mean of N uniformly distributed numbers for various values of N . As N increases, the distribution tends towards a Gaussian [4].

The probability density of the normal distribution is represented in Fig. 2.2 and is given by Eq. (2.1) where μ is the mean or expectation of the distribution (and also its median and mode), σ is the standard deviation, σ^2 is the variance.

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2.1)$$

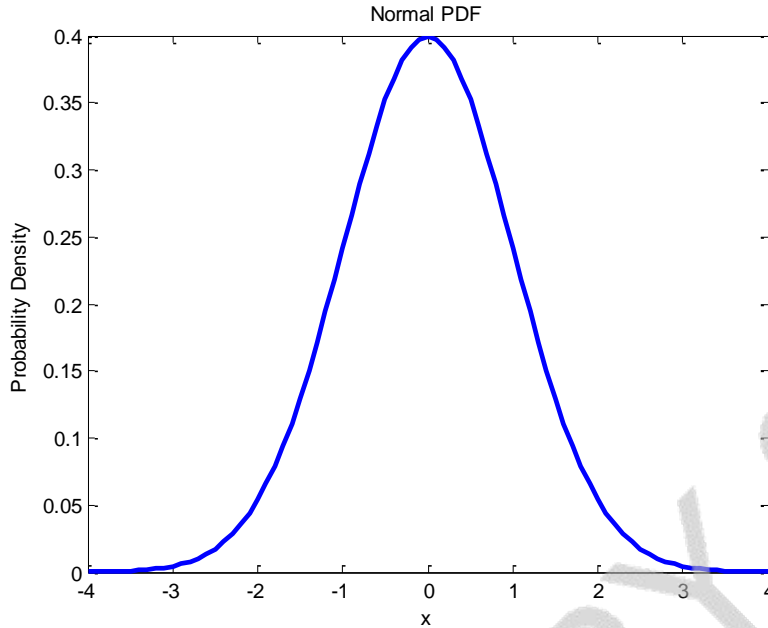


Fig. 2.2 Plot of the standard normal probability density function

For both theoretical and practical reasons, the normal distribution is probably the most important distribution in statistics. For example,

- Many classical statistical tests are based on the assumption that the data follow a normal distribution. This assumption should be tested before applying these tests.
- In modeling applications, such as linear and non-linear regression, the error term is often assumed to follow a normal distribution with fixed location and scale [5].

2.3 Gaussian Mixture Model

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. Training data is used to estimate the GMM parameters using the iterative Expectation-Maximization (EM) algorithm or Maximum *A Posteriori* (MAP) estimation from a well-trained prior model.

A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. Mixture models generalize k -means clustering to incorporate information about the covariance structure of the data as well as the centers of the latent Gaussians [6].

A Gaussian mixture model is parameterized by two types of values, the mixture component weights and the component means and variances/covariances. If the

component weights aren't learned, they can be viewed as an a-priori distribution over components. If they are instead learned, they are the a-posteriori estimates of the component probabilities given the data.

The Gaussian mixture distribution can be written as a linear superposition of Gaussians as shown in Eq. (2.2) where π_k is the mixing coefficient such that $\sum_{k=1}^K \pi_k = 1$, μ_k is the mean, Σ_k is the covariance.

$$p(x) = \sum_{k=1}^K \pi_k N(x|\mu_k, \Sigma_k) \quad (2.2)$$

Gaussian mixture model is a probabilistic model for representing normally distributed subpopulations within an overall population. Mixture models in general don't require knowing which subpopulation a data point belongs to, allowing the model to learn the subpopulations automatically. Since subpopulation assignment is not known, this constitutes a form of unsupervised learning.

A GMM may have more than two components. Estimating the parameters of the individual normal distribution components is a canonical problem in modeling data with GMMs. GMMs have been used for feature extraction from speech data and have also been used extensively in object tracking of multiple objects, where the number of mixture components and their means predict object locations at each frame in a video sequence.

The data might follow a mixture model is that the data looks multimodal, i.e. there is more than one "peak" in the distribution of data. Fig. 2.3 (a) shows a sample unclustered data, (b) shows the fitting of the data using a single cluster while (c) shows fitting the same data using three clusters of a Gaussian mixture. Since many simple distributions are unimodal, an obvious way to model a multimodal distribution would be to assume that it is generated by multiple unimodal distributions. For several theoretical reasons, the most commonly used distribution in modeling real world unimodal data is the Gaussian distribution. Thus, modeling multimodal data as a mixture of many unimodal Gaussian distributions makes intuitive sense. Furthermore, GMMs maintain many of the theoretical and computational benefits of Gaussian models, making them practical for efficiently modeling very large datasets [7].

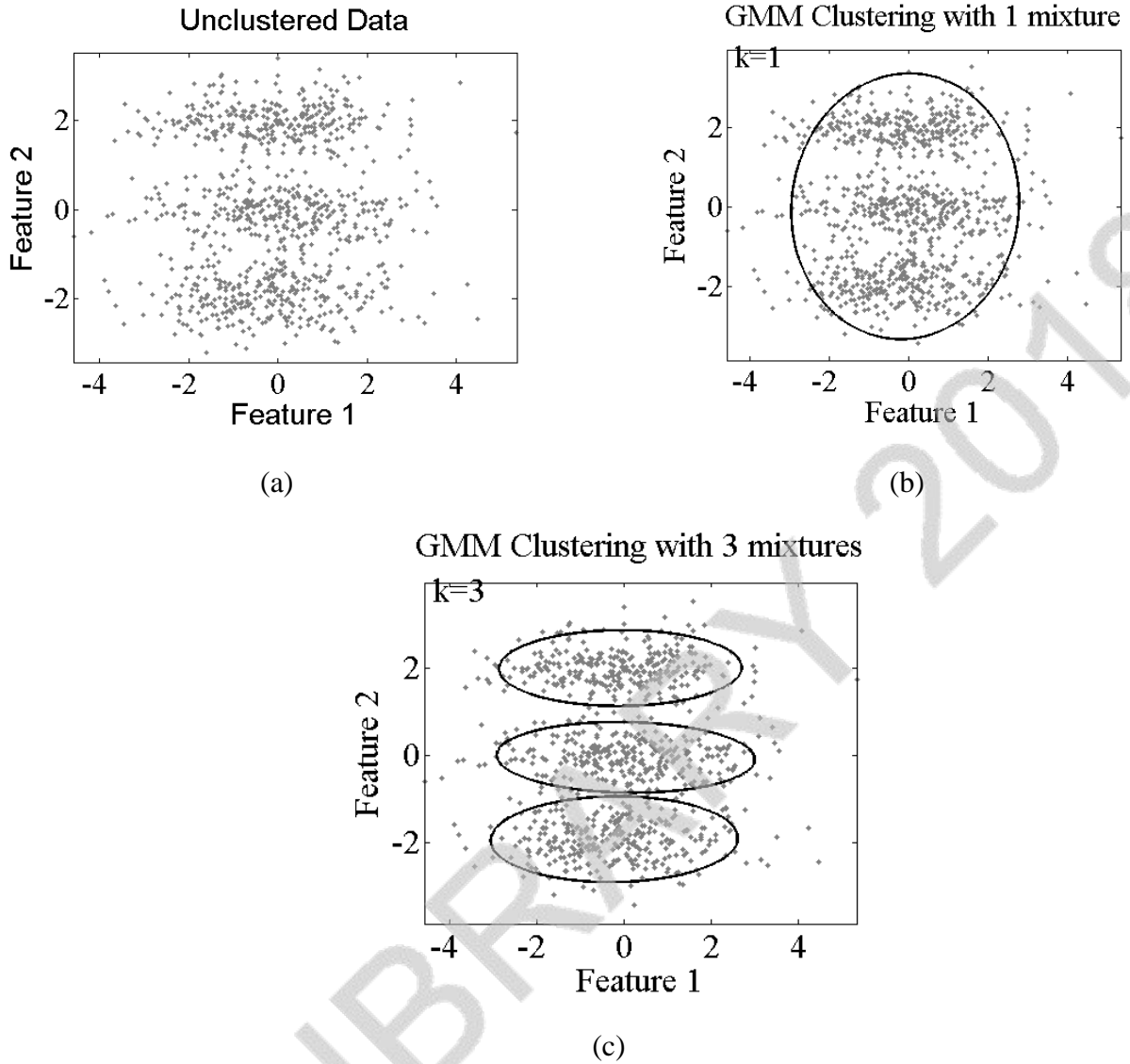


Fig. 2.3 (a) Unclustered Data, (b) Fit with one Gaussian distribution, (c) Fit with Gaussian mixture model with three components

GMM is a lot more flexible in terms of cluster covariance. k -means is actually a special case of GMM in which each cluster's covariance along all dimensions approaches 0. This implies that a point will get assigned only to the cluster closest to it. With GMM, each cluster can have unconstrained covariance structure. Hence, there is more flexibility for cluster assignment in GMM as compared to k -means. Another implication of its covariance structure is that GMM allows for mixed membership of points to clusters. In k -means, a point belongs to one and only one cluster, whereas in GMM a point belongs to each cluster to a different degree. The degree is based on the probability of the point being generated from each cluster's (multivariate) normal distribution, with cluster center as the distribution's mean and cluster covariance as its covariance. Depending on the task, mixed membership may be more

appropriate (e.g. news articles can belong to multiple topic clusters) or not (e.g. organisms can belong to only one species) [8]. The advantages and drawbacks of GMM are described in Table 2.1.

Table 2.1 Pros and cons of GMM [8]

Pros:

Speed	It is the fastest algorithm for learning mixture models
Agnostic	As this algorithm maximizes only the likelihood, it will not bias the means towards zero, or bias the cluster sizes to have specific structures that might or might not apply.

Cons:

Singularities	When one has insufficiently many points per mixture, estimating the covariance matrices becomes difficult, and the algorithm is known to diverge and find solutions with infinite likelihood unless one regularizes the covariances artificially.
Number of components	This algorithm will always use all the components it has access to, needing held-out data or information theoretical criteria to decide how many components to use in the absence of external cues.

2.4 Expectation Maximization Algorithm

The main difficulty in learning Gaussian mixture models from unlabeled data is that one usually doesn't know which points came from which latent component (if one has access to this information it gets very easy to fit a separate Gaussian distribution to each set of points). Expectation-maximization is a well-founded statistical algorithm to get around this problem by an iterative process. First one assumes random components (randomly centered on data points, learned from k -means, or even just normally distributed around the origin) and computes for each point a probability of being generated by each component of the model. Then, one tweaks the parameters to maximize the likelihood of the data given those assignments. Repeating this process is guaranteed to always converge to a local optimum [9]. If the number of components is known, expectation maximization is the technique most commonly used to estimate the mixture model's parameters. In frequentist probability theory, models are typically learned by using maximum likelihood estimation techniques, which seek to maximize the probability, or likelihood, of the observed data given the model parameters. Unfortunately, finding the maximum likelihood solution for mixture models by differentiating the log likelihood and solving for is usually analytically impossible.

Expectation maximization (EM) is a numerical technique for maximum likelihood estimation and is usually used when closed form expressions for updating the model parameters can be calculated (which will be shown below). Expectation maximization is an iterative algorithm and has the convenient property that the maximum likelihood of the data strictly increases with each subsequent iteration, meaning it is guaranteed to approach a local maximum.

Expectation maximization for mixture models consists of two steps:

The first step, known as the Expectation, or E, step, consists of calculating the expectation of the component assignments for each data point. The second step is known as the Maximization, or M, step, which consists of maximizing the expectations calculated in the E step with respect to the model parameters.

The entire iterative process repeats until the algorithm converges, giving a maximum likelihood estimate. Intuitively, the algorithm works because knowing the component assignment for each makes solving for and easy, while knowing and makes inferring easy. The expectation step corresponds to the latter case while the maximization step corresponds to the former. Thus, by alternating between which values are assumed fixed, or known, maximum likelihood estimates of the non-fixed values can be calculated in an efficient manner [8].

Given a Gaussian Mixture Model, the goal is to maximize the likelihood function with respect to the parameters (comprising the means and covariances of the components and the mixing coefficients) [4].

1. Initialize the means, covariances and mixing coefficients and evaluate the initial value of the log likelihood.
2. **E step:** Evaluate the responsibilities using the current parameter values using Eq. (2.3).

$$\gamma(z_{nk}) = \frac{\pi_k N(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j N(x_n | \mu_j, \Sigma_j)} \quad (2.3)$$

3. **M step:** Re-estimate the parameters using the current responsibilities using Eq. (2.4), (2.5) and (2.6).

$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) x_n \quad (2.4)$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (x_n - \mu_k^{new})(x_n - \mu_k^{new})^T \quad (2.5)$$

$$\pi_k^{new} = \frac{N_k}{N} ; N_k = \sum_{n=1}^N \gamma(z_{nk}) \quad (2.6)$$

4. Evaluate the log likelihood using Eq. (2.7) and check for convergence of either the parameters or the log likelihood. The convergence criterion involves comparison of log likelihoods of successive iterations and compare with the threshold value. If it is less than threshold, return to E step.

$$\ln p(X | \mu, \Sigma, \pi) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k N(x_n | \mu_k, \Sigma_k) \right\} \quad (2.7)$$

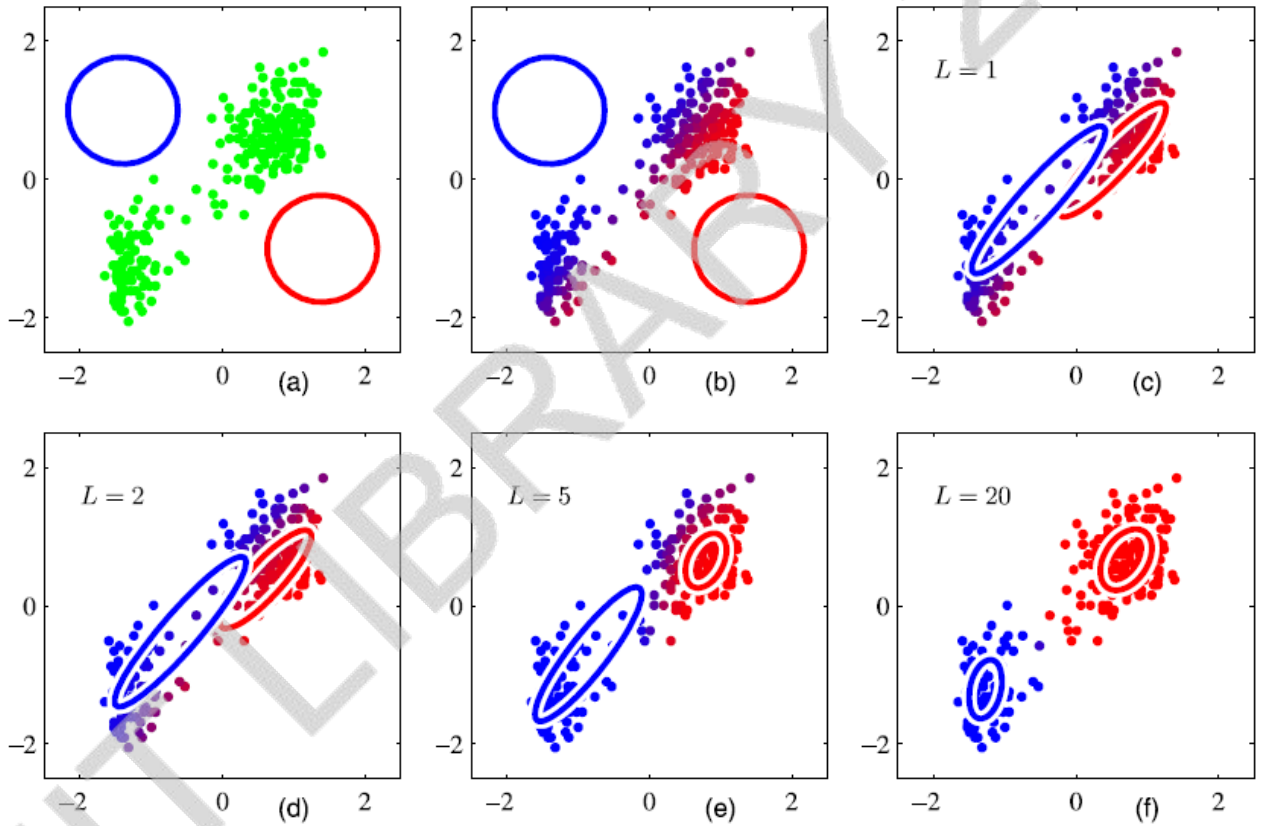


Fig. 2.4 Illustration of the EM algorithm [4]

When the number of components is not known a priori, it is typical to guess the number of components and fit that model to the data using the EM algorithm. The EM algorithm is illustrated in Fig. 2.4. Usually, the model with the best trade-off between fit and number of components (simpler models have fewer components) is kept.

Once the EM algorithm has run to completion, the fitted model can be used to perform various forms of inference. The two most common forms of inference done on GMMs are density estimation and clustering.

Since the GMM is completely determined by the parameters of its individual components, a fitted GMM can give an estimate of the probabilities of both in-sample and out-of-sample data points, known as density estimation. Furthermore, since numerically sampling from an individual Gaussian distribution is possible, one can easily sample from a GMM to create synthetic datasets.

Using Bayes theorem and the estimated model parameters, one can also estimate the a-posteriori component assignment probability. Knowing that a data point is likely from one component distribution versus another provides a way to learn clusters, where cluster assignment is determined by the most likely component assignment as shown in Fig. 2.5. Clustering has many uses in machine learning, ranging from tissue differentiation in medical imaging to customer segmentation in market research. Given a univariate model's parameters, the probability that a data point belongs to component is calculated using Bayes' Theorem [7].

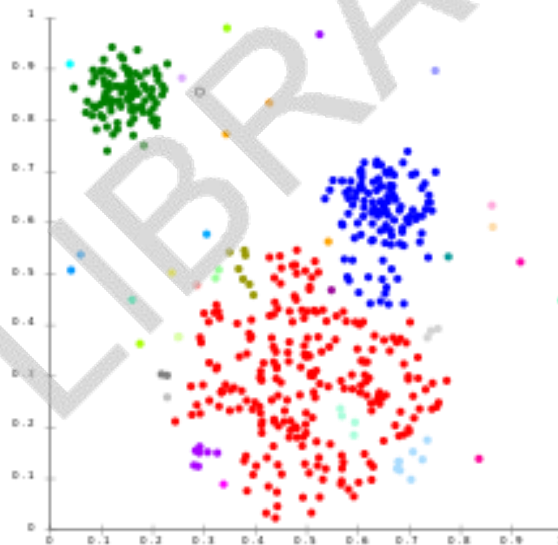


Fig. 2.5 Clustering using a Gaussian mixture model. Each color represents a different cluster according to the model [7].

GMMs have been used recently for feature extraction from speech data for use in speech recognition systems. They have also been used extensively in object tracking of multiple objects, where the number of mixture components and their means predict object

locations at each frame in a video sequence. The EM algorithm is used to update the component means over time as the video frames update, allowing object tracking [7].

2.4.1 Applications

The EM algorithm has many applications, including:

- Dis-entangling superimposed signals,
- Estimating Gaussian mixture models (GMMs),
- Estimating hidden Markov models (HMMs),
- Estimating parameters for compound Dirichlet distributions,
- Finding optimal mixtures of fixed models [10].

2.4.2 Limitations

The EM algorithm can be very slow, even on the fastest computer. It works best when we only have a small percentage of missing data and the dimensionality of the data isn't too big. The higher the dimensionality, the slower the E-step; for data with larger dimensionality, we may find the E-step runs extremely slow as the procedure approaches a local maximum [10].

2.5 Support Vector Regression (SVR)

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall [11]. Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize an algorithm. The Support Vector Regression (SVR) uses the same principles as the SVM for classification with only a few minor differences.

Instead of minimizing the observed training error, Support Vector Regression (SVR) attempts to minimize the generalization error bound so as to achieve generalized performance.

The idea of SVR is based on the computation of a linear regression function in a high dimensional feature space where the input data are mapped via a nonlinear function. SVR has been applied in various fields – time series and financial (noisy and risky) prediction, approximation of complex engineering analyses, convex quadratic programming and choices of loss functions, etc. [12].

A version of SVM for regression was proposed in 1996 by Vladimir N. Vapnik, Harris Drucker, Christopher J. C. Burges, Linda Kaufman and Alexander J. Smola. This method is called support vector regression (SVR). The model produced by support vector classification (as described above) depends only on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. Analogously, the model produced by SVR depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction. Another SVM version known as least squares support vector machine (LS-SVM) has been proposed by Suykens and Vandewalle. Training the original SVR means solving Eq. (2.8) which is given as

$$\begin{aligned} \min \quad & \frac{1}{2} \|\omega\|^2 \\ \text{subject to} \quad & y_i - \langle \omega, x_i \rangle - b \leq \varepsilon; \quad \langle \omega, x_i \rangle + b - y_i \leq \varepsilon \end{aligned} \quad (2.8)$$

where x_i is a training sample with target value y_i . The inner product plus intercept

$\langle \omega, x_i \rangle + b$ is the prediction for that sample, and ε is a free parameter that serves as a threshold: all predictions have to be within an ε range of the true predictions.

In machine learning, the (Gaussian) radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification. The RBF kernel on two samples x and x' , represented as feature vectors in some *input space*, is defined in Eq. (2.9).

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \quad (2.9)$$

$\|x - x'\|^2$ may be recognized as the squared Euclidean distance between the two feature vectors. σ is a free parameter [49]. Support Vector Regression (prediction) with different thresholds ε is as shown in Fig. 2.6. As ε increases, the prediction becomes less sensitive to errors. Slack variables are usually added into the above to allow for errors and to allow approximation in the case the above problem is infeasible [11].

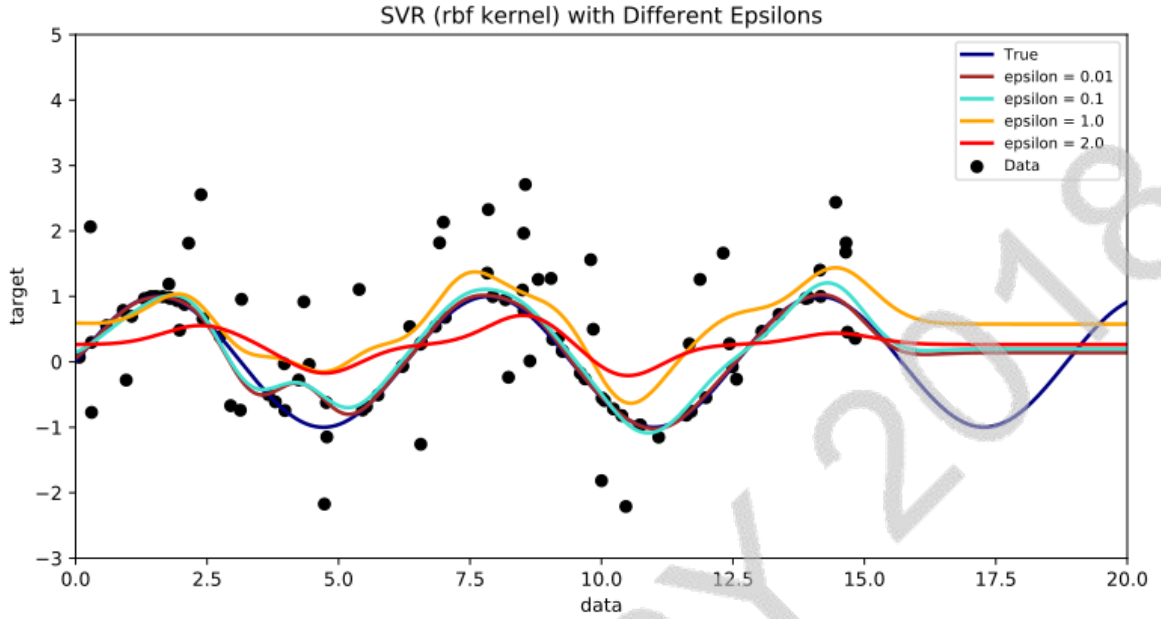


Fig. 2.6 Support Vector Regression (prediction) with different thresholds ε [11].

2.6 Vector Quantization

Vector quantization (VQ) model also known as centroid model, is one of the simplest text-independent speaker models. It was introduced to speaker recognition in the 1980s and its roots are originally in data compression. VQ is often used for computational speedup techniques and lightweight practical implementations.

VQ involves the process of taking a large set of feature vectors of a particular user and producing a smaller set of feature vectors that represent the centroids of the distribution, i.e. points spaced so as to minimize the average distance to every other point. VQ is used since it would be highly impractical to represent every single feature vector in feature space that we generate from the training utterance of the corresponding speaker. While the VQ algorithm does take a while to generate the centroids, it saves a lot of time during the testing phase as we are only considering few feature vectors instead of overloaded feature space of a particular user. Therefore, it is an economical compromise that we can live with.

A vector quantizer maps k -dimensional vectors in the vector space R_k into a finite set of vectors $Y = \{y_i: i = 1, 2, \dots, N\}$. Here k -dimension refers to the no of feature coefficients in each feature vector. Each vector y_i is called a code vector or a codeword and the set of all the code words is called a codebook. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown

voice is “vector-quantized” using each trained codebook and the total VQ distortion is computed [42].

VQ is a classical quantization technique from signal processing that allows the modelling of probability density functions by the distribution of prototype vectors. It was originally used for data compression. It works by dividing a large set of points (vectors) into groups having approximately the same number of points closest to them. Each group is represented by its centroid point [43].

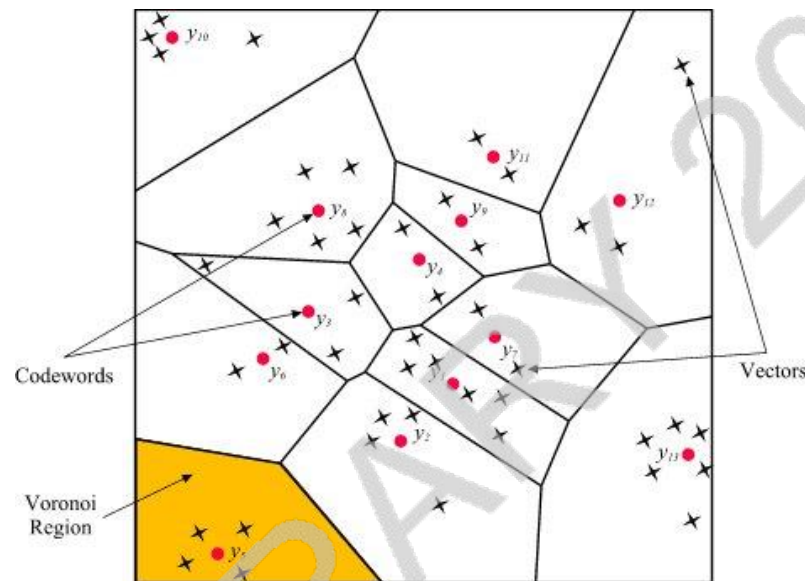


Fig. 2.7 Code vectors in 2-dimensional space. Source vectors are marked with a cross, code vectors are marked with circles, and the Voronoi regions are separated with boundary lines [44].

The representation for VQ is a collection of codebook vectors. A codebook vector is a list of numbers that have the same input and output attributes as the training data. The codebook is partitioned such that for each vector x a nearest neighbor codevector y_i exists. Mathematically spoken this partition is called *Voronoi* or *Dirichlet partition* and the code vectors are the *centroids* of each region. The following Fig. 2.7 represents the two-dimensional case. In general, the probability density function of the source is rarely known. To circumvent this problem the rate distortion theory and high-resolution theory offers each a solution. The later one is to choose a large number of code vectors. The associated regions are small and high structured as a *lattice*. So instead to find the nearest neighbor the number of code vectors is raised such that for each vector x the distortion is approximately constant. As mentioned, this results to a local uniformity of the source probability density function and a high bit rate.

The other solution is to use a training set representing best the source to optimize the codebook. To achieve this, a *clustering algorithm* is used. Such an algorithm is the Lloyd algorithm. It iteratively improves a codebook by alternately optimizing the encoder for the decoder -

subdividing the codebook in regions (Voronoi regions) in the manner that the average distortion for the given training set is minimal, and the decoder for the encoder - replacing the code vectors by the centroids. This is repeated until the average distortion and rate converges to an inferior limit set by the Shannon rate distortion theory. So, the codebook is optimal for the training set but not necessary for the set of source vectors. Other clustering algorithms are known such as the pairwise nearest neighbor algorithm by Ward and Equitz, k-means algorithm, neural net approaches, simulated annealing and stochastic relaxation algorithms just to mention some [44].

VQ is a data reduction method which means that it seeks to reduce the number of dimensions in the input data so that the models used to match unknowns can be as simple as possible. VQ reduces dimensionality quite drastically since it encodes each vector as a single number.

In order to make the VQ approach workable the number of clusters needs to be quite large to provide sufficient discriminatory power. It is common to use cluster sizes which are powers of two since a common consideration is the number of binary digits needed to represent the input string. Common value is 64 or 128 clusters which require 6 and 7 binary digits respectively. The set of cluster centers is described as a codebook since it is used to translate feature vectors into single numbers. Codebook size refers to the number of entries or clusters in the codebook.

When VQ is used to encode an input sequence, information is lost. There is grouping of dissimilar points and effective representation of every cluster member by the cluster mean. One way of measuring this loss of information is to look at the difference between the original input vector sequence and the quantized version. The average value of this difference is called the VQ distortion and gives a measure of how well a particular set of clusters fits a set of data. The larger the codebook the smaller the VQ distortion since each vector can be assigned to a closer cluster centre [45].

2.7 AKL and AED

2.7.1 Average KL Divergence (AKL)

To measure the difference between two probability distributions over the same variable x , a measure, called the Kullback-Leibler divergence is used. The KL divergence, which is closely related to relative entropy, information divergence, and information for discrimination,

is a non-symmetric measure of the difference between two probability distributions $p(x)$ and $q(x)$.

Specifically, the Kullback-Leibler (KL) divergence of $q(x)$ from $p(x)$, denoted $D_{KL}(p(x), q(x))$, is a measure of the information lost when $q(x)$ is used to approximate $p(x)$. Here, let $p(x)$ and $q(x)$ are two probability distributions of a discrete random variable x . That is, both $p(x)$ and $q(x)$ sum up to 1, and $p(x) > 0$ and $q(x) > 0 \forall x \in X$. $D_{KL}(p(x), q(x))$ is defined in Eq. (2.10).

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)} \quad (2.10)$$

Typically, $p(x)$ represents the “true” distribution of data, observations, or a precisely calculated theoretical distribution. The measure $q(x)$ typically represents a theory, model, description, or approximation of $p(x)$. The continuous version of the KL divergence is defined in Eq. (2.11)

$$D_{KL}(p(x)||q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx \quad (2.11)$$

Although the KL divergence measures the “distance” between two distributions, it is not a distance measure. This is because that the KL divergence is not a metric measure. It is not symmetric: the KL from $p(x)$ to $q(x)$ is generally not the same as the KL from $q(x)$ to $p(x)$. Furthermore, it need not satisfy triangular inequality. Nevertheless, $D_{KL}(P||Q)$ is a non-negative measure. $D_{KL}(P||Q) \geq 0$ and $D_{KL}(P||Q) = 0$ if and only if $P = Q$. For a good model the value of average KL Divergence should be as small as possible [46].

2.7.2 Average Euclidean Distance (AED)

The Euclidean distance between two points in either the plane or 3-dimensional space measures the length of a segment connecting the two points. It is the most obvious way of representing distance between two points. The Pythagorean Theorem can be used to calculate the distance between two points. If the points (x_1, y_1) and (x_2, y_2) are in 2-dimensional space, then the Euclidean distance between them is $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ [47].

The weighted Euclidean distance between A and B is defined in Eq. (2.12)

$$d(A, B) = \sqrt{\sum_i w_i (A_i - B_i)^2} \quad (2.12)$$

where A_i is the i -th feature for A and w_i is the weight to be given to feature i .

If there are many points in the space, one possible way to choose w_i is the inverse variance of the feature i [48].

In mathematics, the Euclidean distance or Euclidean metric is the “ordinary” straight-line distance between two points in Euclidean space. With this distance, Euclidean space becomes a metric space. A Euclidean space has some number of real-valued dimensions. There is a notion of average of two points. A Euclidean distance is based on the locations of points in such a space.

2.8 MIR ToolBox

MIR toolbox is a toolbox of MATLAB that supports various unique features like extraction of many key musical features from the audio files. It can also be used for statistical analysis, segmentation and clustering. MIR toolbox integrates a user-friendly syntax that enables to easily combine low-level and high-level operators into complex flowcharts. The modular design of MIR toolbox is guided by a philosophy of expertise capitalization: techniques developed for certain domains of music analysis are turned into general operators that could be used for different analytical purposes. Each feature extraction method can accept as argument an audio file, or any preliminary result from intermediary stages of the chain of operations. Also, the same syntax can be used for analysis of one or many audio files, series of audio segments, multi-channel signals, etc. For that purpose, the data and methods of the toolbox are organized in an object-oriented architecture [13].

We used the following features from MIR toolbox:

Various Dynamic Features extracted were:

- *RMS*: It gives global energy of the waveform. It can be used to classify exciting/relaxing music.

Various Spectral Features extracted were:

- *Centroid*: Weighted average of amplitude spectrum.
- *Spread*: Returns standard deviation of data.
- *Skewness*: Measure symmetry of distribution.
- *Kurtosis*: Kurtosis more commonly defined as the fourth cumulant divided by the square of the variance of the probability distribution, equivalent to:

$$\frac{\mu_4}{\sigma^4} - 3$$

which is known as excess kurtosis.

- *Spectentropy*: Provides general description of audio waveform.
- *Flatness*: Indicates whether distribution is smooth or spiky. It is used to extract statistics of the music.
- *Rolloff (85% and 95%)*: Finding a frequency such that a certain fraction of total energy is contained below that frequency.
- *Brightness*: Measuring the amount of energy above that frequency. The result is expressed as a number between 0 and 1.
- *Roughness*: Estimates the total roughness by computing the peaks of the spectrum, taking the average of all the dissonance between all possible pairs of peaks.
- *Irregularity*: The irregularity of a spectrum is the degree of variation of the successive peaks of the spectrum.

Various Timbre Features were:

- *Zerocross*: Counts the number of times the signal crosses the X-axis. It gives the indication of noisiness.
- *MFCC*: 13-dimensional h is used which is useful in speaker recognition process. It possesses human perception sensitivity w.r.t frequencies.

$$mel(f) = 1125 * \ln(1 + f/700) \quad (2.13)$$

Fig. 2.8 illustrates the generation of a 13-D MFCC. The input audio signal is first windowed using a Hamming window and then FFT is applied to it. The result is mapped to the Mel frequency scale. Inverse log operation is performed on it followed by DCT to obtain the MFCC.

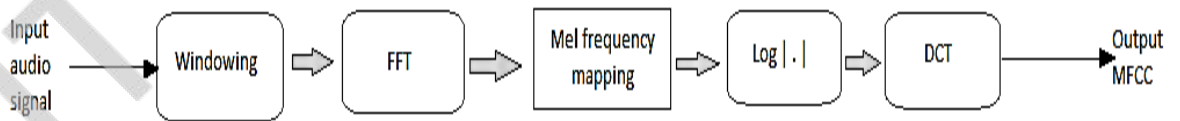


Fig. 2.8 Block diagram of generation of 13-D MFCC

- *DMFCC*: Extracted using first derivative of MFCC features. It represents change between frames.
- *DDMFCC*: Extracted by taking derivative of Delta features. It shows the changes between frames in the corresponding delta features.

Various Tonal Features were

- *Mode*: Estimates the modality, i.e. major vs. minor, returned as a numerical value between -1 and +1; the closer it is to +1, the more major the given excerpt is predicted to be, the closer the value is to -1, the more minor the excerpt might be.
- *Keyclarity*: It computes the key strength, a score between -1 and +1 associated with each possible key candidate.
- *HCDF*: The Harmonic Change Detection Function (HCDF) is the flux of the tonal centroid
- *Chromagram*: The chromagram, also called Harmonic Pitch Class Profile, shows the distribution of energy along the pitches or pitch classes. Other features such as Chromagram.peak, Chromagram.centroid, chromagram_data are extracted using chromagram [14].

Chapter 3

Literature Survey

In this chapter, we present basic emotion concepts followed by the survey of nine reference papers.

3.1 Classification of Emotions

The two approaches of computational modeling of music emotions are categorical approach and dimensional approach.

3.1.1 Categorical Approach

Basic emotions are described as "discrete" because they are believed to be distinguishable by an individual's facial expression and biological processes [15]. Categorical approach classifies the emotions into discrete mood classes such as happy, sad, angry or relaxed. It offers an atomic description of music which is easy to incorporate in conventional text-based retrieval systems [16]. The various mood classes in the categorical approach are depicted in Fig. 3.1.

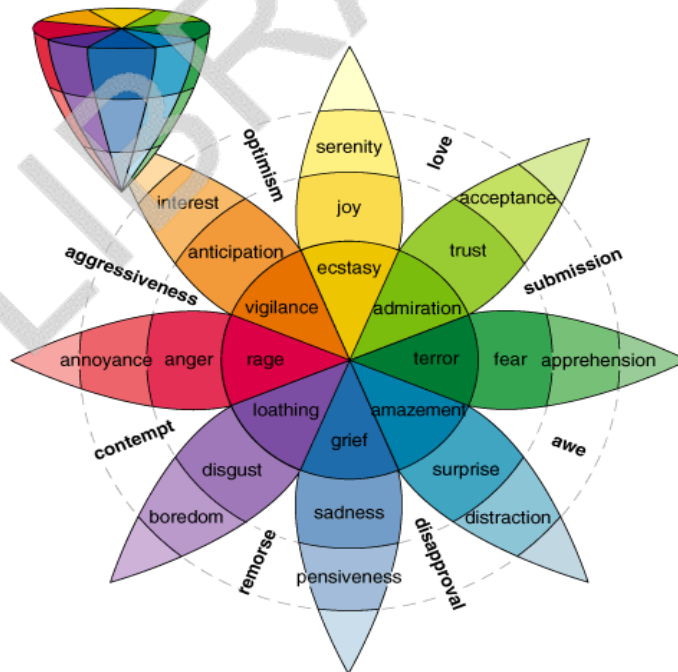


Fig. 3.1 Categorical approach [17]

The advantage of categorical approach is that it represents human emotions intuitively with easy to understand emotion labels. But according to psychological theories, basic

emotions are too few to capture the various sensations of human emotions and the richness of human perception. Also, the emotional terms differ with culture, linguistic, environmental and personal differences. Therefore, the same affective state can be expressed by different emotional categories causing poor agreement among them. The increasing the number of emotion classes can easily lead to overwhelming the user and can further introduce ambiguities making it difficult to obtain the ground-truth values. The major limitation of this model is it does not provide a precise identification of the emotional state perceived by people [18]. An example of categorical model is Hevner's model.

3.1.1.1 Hevner's Model

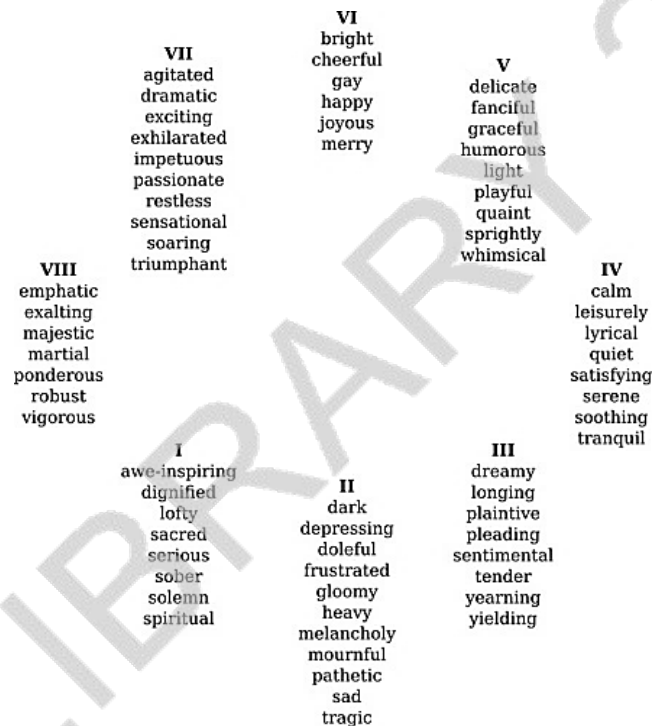


Fig. 3.2 Hevner's model [19]

Hevner's model shown in Fig. 3.2 is a categorical model which describes the different music mood states by employing a list of 66 adjectives arranged in eight clusters. These clusters are arranged in a circle. Adjectives with similar meaning are placed in same cluster while adjacent clusters slightly differ in their meaning. The difference in the meaning between clusters increases with the distance. Therefore, neighboring clusters express similar moods and opposite ones exhibit the most difference. The use of lot of adjectives increases the complexity of mood mapping [20].

In order to avoid the semantic ambiguity and possible overlaps of affective terms, it is preferred to represent the emotions with VA values instead of discrete emotion classes.

3.1.2 Dimensional Approach

For both theoretical and practical reasons researchers define emotions according to one or more dimensions. In the dimensional approach, emotions can be described by three dimensions: "pleasurable versus unpleasurable", "arousing or subduing" and "strain or relaxation". The human emotions are conceptualized by defining where they lie in two or three dimensions. Most dimensional models incorporate valence and arousal dimensions in which emotions are represented as numerical values over the two dimensions – valence (positive or negative affective state) and arousal (intensity or energy level) [15]. The dimensional model offers a simple means of creating a 2D user interface [16]. It provides a generalized MER from categorical domain to real-valued domain. The two-dimensional approach models are- Thayer's Model and Russell's Model.

3.1.2.1 Thayer's Model

It is a two-dimensional model wherein moods are positioned in a multi-dimensional space.

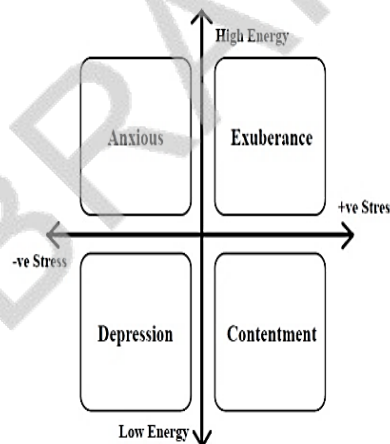


Fig. 3.3 Thayer's model [21]

As shown in Fig. 3.3, Thayer's mood model divides mood into four clusters namely, Contentment, Depression, Exuberance, Anxious (Frantic). Contentment refers to calm and happy music so it is placed in positive stress and low energy region, Depression refers to calm and anxious music so it is placed in negative stress and low energy region, Exuberance refers to happy and energetic music so it is placed in positive stress and high energy region while Anxious refers to frantic and energetic music so it is placed in negative stress and high energy region [21].

3.1.2.2 Russell's Model

It is a two-dimensional model as shown in Fig. 3.4, in which emotions are distributed in a two-dimensional circular space, containing arousal and valence dimensions. Arousal represents the vertical axis and Valence represents the horizontal axis, while the centre of the circle represents a neutral valence and a medium level of arousal [15].

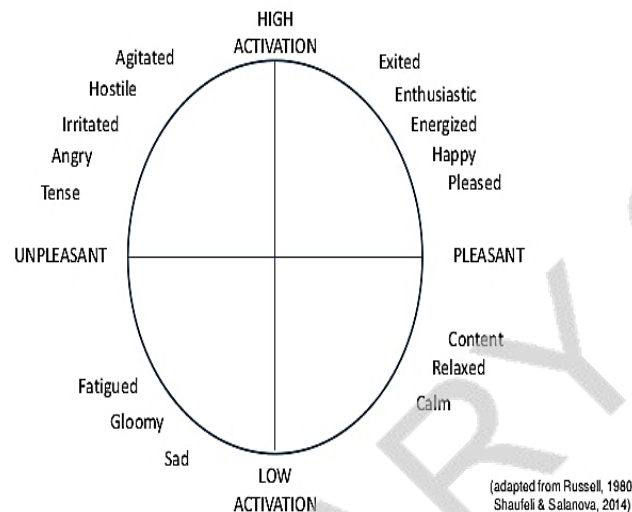


Fig. 3.4 Russell's model [22]

Emotions can be independently described by each of the primitives on a continuous scale from $[-1,1]$.

3.2 Valence and Arousal

Emotional experiences can be described by two terms: Valence & Arousal. The two terms are often used interchangeably. Valence is positive or negative affectivity, whereas arousal measures how calming or exciting the information is [23]. Valence is the "intrinsic attractiveness or aversiveness" of an emotion which is cognitive-psychologist speaking for whether or not people would want to feel something [24]. Most people want to be happy, and most people don't want to be sad, happiness good, sadness bad, happiness positive, sadness negative. Arousal is physiological and psychological state of being reactive to stimuli. It results in an observable change in the physical state of the body which causes you to become alert and a ready to move and respond. In other words, "**arousal**" is the level/amount of physical response and "**valence**" is the emotional "direction" of that emotion.

Emotions can be mapped out on a chart modeling the range of arousal (high to low) and valence (pleasure to displeasure) that is experienced during a particular emotion. This mapping

is depicted in Fig. 3.5. For example, in the top right corner are the emotions with high arousal and high valence, which include excited, astonished, delighted, happy, and pleased. These emotions are all examples of positive emotions that are high in arousal. In the opposite corner is the low valence and low arousal section, containing miserable, depressed, bored, and tired as some examples. An individual with high emotional granularity would be able to discriminate between their emotions that all fall within the same level of valence and arousal, labeling their experiences with discrete emotion words. Someone with low emotional granularity would report their emotions in global terms, usually of pleasure or displeasure. Emotional granularity is an individual's ability to differentiate between the specificity of their emotions [25].

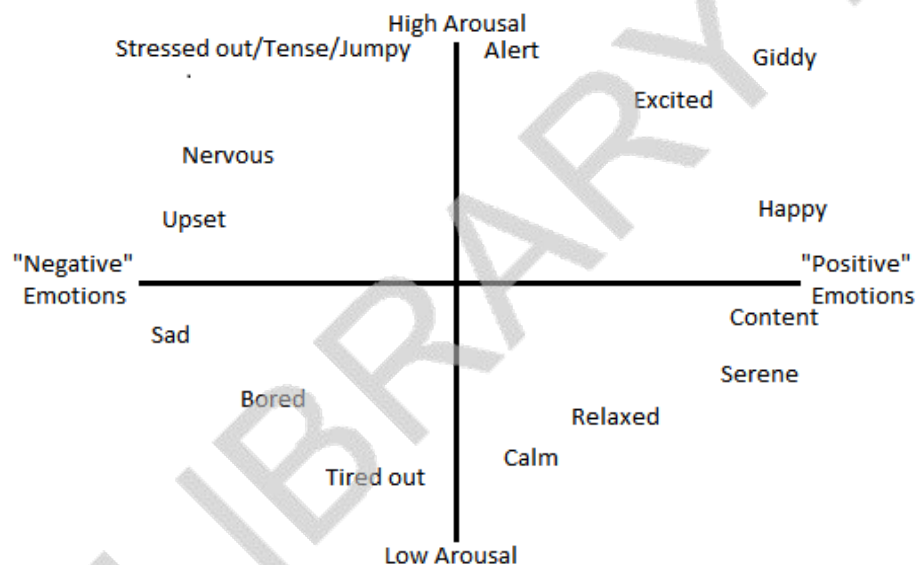


Fig. 3.5 VA space [24]

3.3 Approaches for modeling Valence and Arousal

3.3.1 VA Point Approach

In this method, the affective content of a music excerpt can be represented as a single point in the VA space shown in Fig. 3.6. The ground-truth VA values of a music excerpt are obtained by averaging the annotations of a number of listeners, without considering the covariance of the annotations. For simplicity, the two dimensions are usually assumed to be

independent, so that each dimension can be modelled by fitting a regression model that minimizes the error between the predicted and the ground-truth values [26].

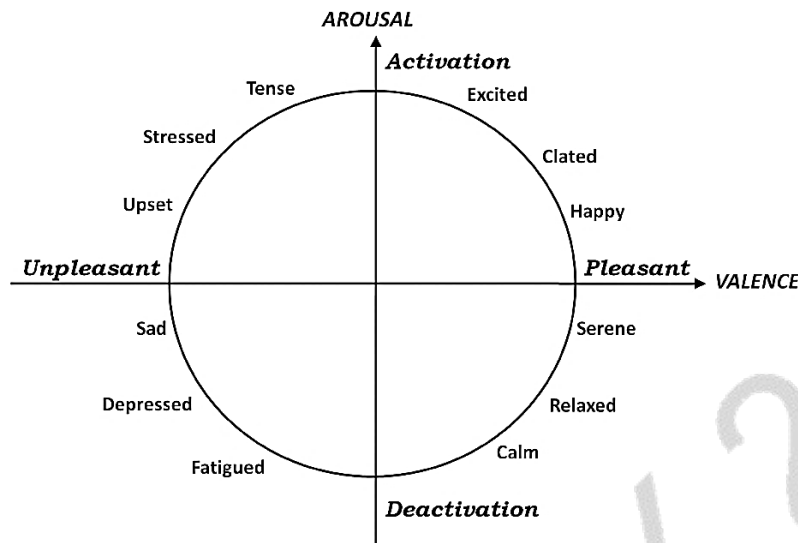


Fig. 3.6 VA point approach [27]

3.3.2 Heatmap Approach

A heat map is a two-dimensional representation of data in which values are represented by colors. A simple heat map provides an immediate visual summary of information. More elaborate heat maps allow the viewer to understand complex data sets. There can be many ways to display heat maps, but they all share one thing in common -- they use color to communicate relationships between data values that would be much harder to understand if presented numerically in a spreadsheet [28].

Machine learning models are usually characterized by very high predictive power, but in many case, are not easily interpretable by a human. Interpreting a nonlinear classifier is important to gain trust into the prediction, and to identify potential data selection biases or artifacts.

The project studies in particular techniques to decompose the prediction in terms of contributions of individual input variables such that the produced decomposition (i.e. explanation) can be visualized in the same way as the input data [29]. These visualizations are called "heatmaps".

3.3.3 Gaussian Parameter Approach

The Gaussian Parameter approach parameterizes emotion distribution as a Gaussian and uses regression models to predict from acoustic features the Gaussian parameters (i.e.,

mean and covariance) of a music excerpt. It is an intuitive extension of the VA-point approach, where in the lessons learned from previous work can be applied to construct effective models for the mean VA values. The Gaussian Parameter approach is mostly discriminative and does not offer a strict probabilistic interpretation. The correlation among the Gaussian parameters also remains unmodeled. The predicted covariance is not guaranteed to be a positive definite matrix because the parameters of covariance are modeled independently. Therefore, heuristic may be applied to adjust the predicted parameters to produce a valid covariance [26]. The emotion distribution is modeled by some known probability function and machine learning is then applied to predict the parameterization of the distribution from music features.

This single-Gaussian approach is computationally more efficient and allows easier performance analysis. But, its prediction accuracy is limited by the single Gaussian assumption. Instead of assigning an emotion label or an emotion value to a clip, the probability of the perceived emotion of the clip at any point in the emotion plane is computed [30].

3.4 Related Work

a) **K. F. MacDorman, S. Ough, and C.-C. Ho, “Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison,” *J. New Music Res.*, vol. 36, no. 4, pp. 281–299, 2007 [31].**

It contributes to psychology by refining an index to measure pleasure and arousal responses to music. It contributes to music visualization by developing a representation of pleasure and arousal with respect to the perceived acoustic properties of music, namely, bark bands (pitch), frequency of reaching a given sone (loudness) value, modulation frequency, and rhythm. It contributes to pattern recognition by designing and testing an algorithm to predict accurately pleasure and arousal responses to music. Various methods of automatic music classification are mentioned which includes:

1. Grouping by acoustic similarity: One of the most natural means of grouping music is to listen for similar sounding passages; however, this is time consuming and challenging, especially for those who are not musically trained.
2. Grouping by genre: They performed genre classification using statistical pattern recognition on training and sample music collections. They focused on three features of audio they felt characterized a genre: timbre, pitch, and rhythm.
3. Grouping by emotion: A number of studies were done on classification on based on emotion with some using Thayer’s model (Wang et al), one extracting timbre, rhythm, and pitch (Pohle et al), other extracting timbre, intensity and rhythm (Liu et al.) which achieved the highest

accuracy, 86.3%, but these results were limited to only four emotional categories. It can also be seen that despite all this small number of emotion categories, accuracy is also poor, never reaching 90%.

This paper reports the results and possible implications of a pilot study and survey used to construct an emotion index for subjective ratings of music combinations. The studies found agreement among listeners regarding the ability of pleasure and arousal to describe accurately the broad emotional categories expressed in music. However, the studies failed to discriminate consistently among nuances within an emotional category (e.g. discriminating sadness and depression, Livingstone & Brown, 2005). This difficulty in defining consistent emotional dimensions for listeners warranted the use of an index proven successful in capturing broad, basic emotional dimensions.

For constructing an index for the emotional impact of music, here they have selected Mehrabian and Russell's (1974) pleasure, arousal and dominance (PAD) model because of its established effectiveness and validity in measuring general emotional responses.

In the application of the PAD index to music, the following survey was conducted to understand the reliability of all the three dimensions. There were 72 participants, evenly split by gender, 52 of whom were between 18 and 25. Representative 30 s excerpts were extracted from 10 songs selected from the Thomson Music Index Demo corpus of 128 songs. The results are depicted in Fig. (3.7). The standard pleasure, arousal, and dominance values were calculated based on the 18 semantic differential item pairs used by the 72 participants to rate the excerpts from the 10 songs. Pleasure and arousal both showed high mutual consistency, with a Cronbach's α of 0.85 and 0.73, respectively. However, the Cronbach's α for dominance was only 0.64. The inconsistency in measuring dominance (Cronbach's $\alpha=0.64$) indicated the dimension to be a candidate for removal from the index, because values for Cronbach's α below 0.70 are generally not considered to represent a valid concept.

The development of an algorithm to predict emotional responses to music accurately: Predictor variables derived from a test excerpt's distance from emotion- weighted visualizations proved to be the most accurate among the compared methods at predicting mean ratings of pleasure and arousal. They also appear to have exceeded the accuracy of published methods though before making that claim direct comparisons should first be made using the same index, music corpus, and participant data. Thus, the proposed technique holds promise for serious commercial applications that demand high accuracy in predicting emotional responses to music.

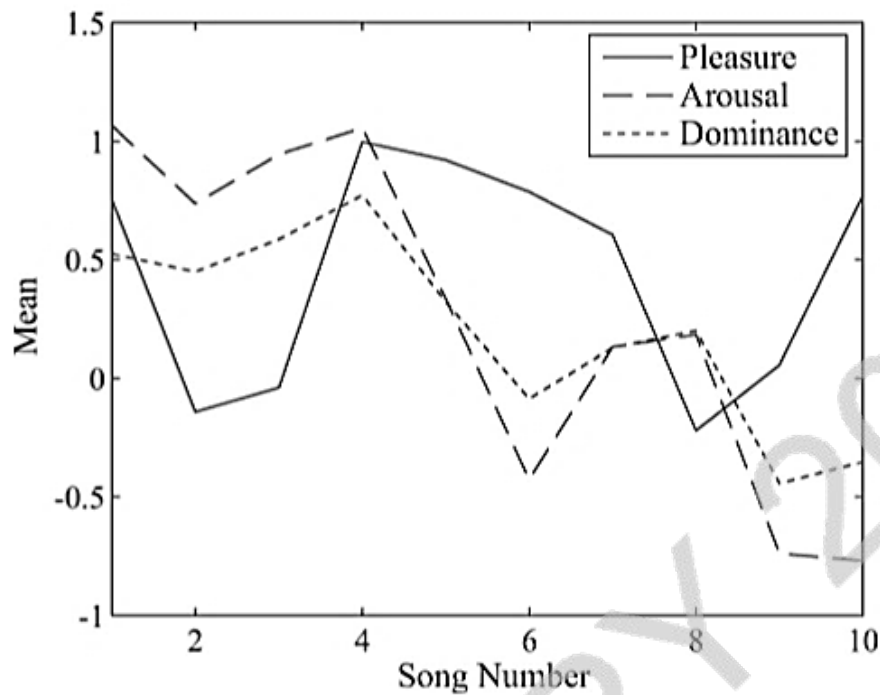


Fig. 3.7 Participants mean PAD ratings for the 10 songs [31]

b) Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, “A regression approach to music emotion recognition,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 2, pp. 448–457, Feb. 2008 [32].

Categorical classification suffers from the subjectivity issue. The subjectivity issue stems from the fact that music perception is intrinsically subjective and is under the influence of many factors such as cultural background, generation, sex, and personality, each music sample become each music sample becomes a point in the arousal-valence plane, so the users can efficiently retrieve the music sample by specifying a desired point in the emotion plane. The performance of existing AV computation methods such as AV modelling, The Fuzzy Approach, The System Identification Approach (System ID) are compared and this illuminated the reason of adopting the regression approach rather than these methods. The system block diagram is as shown in Fig. 3.8 and the system description for the same is given below

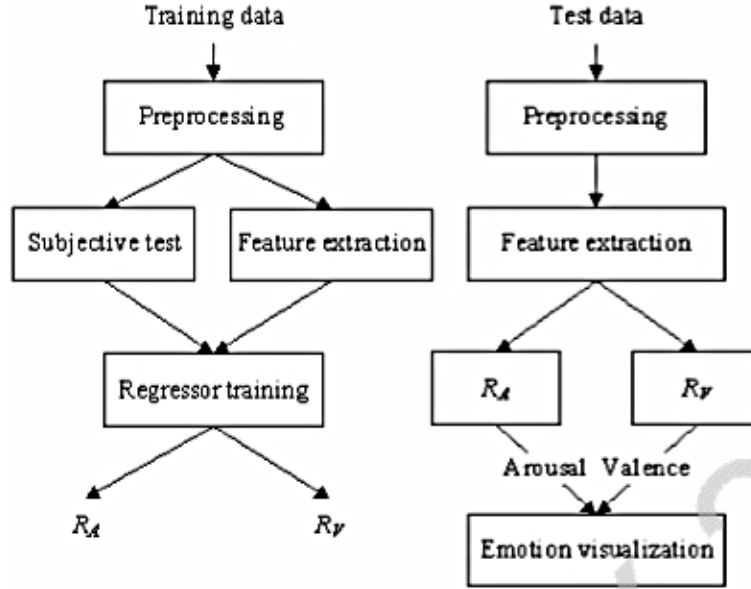


Fig. 3.8 System block diagram [32]

The music database is made up of 195 popular songs selected from a number of Western, Chinese, and Japanese albums. The genre of the database is popular music of different countries rather than the western classical music, which is commonly adopted most of the studies. Since the emotion within a music selection can vary over time, for each song a 25-second segment that is representative of the song and expresses a certain dominant emotion is manually selected. For feature extraction the spectral contrast algorithm, DWCH algorithm, and two computer programs PsySound and Marsyas are used construct a 114-dimension feature space. When all the above four methods are used together, the 114-dimension feature space obtained is represented as ALL. The subjective test sets the ground truth of the AV values. 253 volunteers were recruited and each of them were asked to listen to 10 music samples randomly drawn from the music database and to label the AV values from -1.0 to 1.0 . The subjects were asked to label the emotion based on their feelings of what the music sample is trying to evoke, rather than the emotion the subjects perceive at the test. For regression training, three regression algorithms: multiple linear regression (MLR), support vector regression (SVR), and AdaBoost. RT (BoostR) are used and the best among three will be considered. By evaluating the performance of the three models it was found that the best performance evaluated in terms of the R^2 statistics reaches 58.3% for arousal and 28.1% for valence by employing support vector machine as the regressor, where R is the terms of the R^2 statistics, which is a standard way for measuring the goodness of fit for regression models [17],

$$R^2 = 1 - \frac{N\varepsilon}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (3.1)$$

where \bar{y} is the mean of the ground truth, and the normalization of the total squared error ($N\varepsilon$).

A major issue of the continuous perspective is the dependency between the two dimensions arousal (a) and valence (v) in the arousal-valence (denoted as AV) data space. The Pearson's linear correlation coefficient between a and v in our dataset can reach 0.3368; therefore, the correlation between the two is reduced to check whether the performance is improved. Principal component analysis (PCA) is used for the same converting the data space into principal component space (PC). Also, features are not necessarily of equal importance or quality, and irrelevant or redundant features may lead to inaccurate conclusion. One solution for addressing this problem is to extract a number of musical features and then use a feature selection algorithm (FSA) to identify good features. The feature selection algorithm used over here is RReliefF. RReliefF for each data space and rank the features by importance.

Table 3.1 Comparison of performance of ALL, Psy15, RRF [32]

THE R^2 STATISTICS OF SVR WITH DIFFERENT DATA AND FEATURE SPACES						
	ALL		Psy15		RRF	
	a	v	a	v	a	v
AV	58.6%	14.6%	57.0%	22.2%	60.9%	25.4%
PC	60.2%	16.2%	58.5%	18.1%	58.3%	28.1%

From Table 3.1, we can observe that transforming the data to PC does not make significant difference to the prediction accuracy i.e. reducing the correlation between arousal and valence seems to have little influence. At the same time, selecting features by RRF greatly improves the accuracy (especially for valence), which shows the importance of feature selection. The best performance of the regression approach reaches 58.3% for arousal and 28.1% for valence by using PC+ RRF+SVR.

The distributions of the ground value and the predicted values by PC+ RRF+SVR in this paper is almost the same. Despite that regression approach has more freedom in describing a song, the it may fail to exactly resolve the subjectivity issue since personal difference in the perception of popular music is too high, and since the regressors are trained based upon the average opinions of the subjects.

To resolve this, group-wise MER scheme (GWMER) is developed which divides the users into a variety of user groups and train regressors for each group. The groups can be

defined according to user information such as generation, sex, occupation, personality, etc. to reduce the individual differences for each group.

c) **Y.-H. Yang, Y.-C. Lin, and H. H. Chen, “Personalized music emotion recognition,” in *Proc. ACM SIG Inf. Retrieval*, 2009, pp. 748–749 [33].**

Recently, there has been a rapid increase in the amount of information retrieval methods for subjective concepts such as aesthetic, emotion and preference. It is demanding to build a general model for retrieval that satisfies everyone. The paper proposes two methods namely bag-of-users model and residual modelling to suffice for the individual differences for music retrieval based on emotion and a personalized MER system as shown in Fig. 3.9 is built.

In a bag-of-users model, the opinions of the subjects are averaged to obtain the ground truth data typically needed to train a model for retrieval. The individual annotations assigned by each subject provide many details about the interpretation of a song, but are not used much in the procedure for bag-of-users model. A regression model $M_j(\cdot)$ is trained for each subject u_j using his/her annotations and a bag of models $M_1(\cdot), M_2(\cdot), \dots, M_U(\cdot)$ is obtained, where U denotes the number of subjects. The super model is trained by minimizing the error between the recognition result for each song and the estimate for that song. The estimate is the aggregation of the opinions of the U subjects. In residual modeling, a perception residual of a user u_j can be computed as the difference between the general perception and the personalized one, $r_{ij} = y_{ij} - y_i$, given the general perception y_i of a song s_i . The recognition of perception residual r_{ij} is called residual modelling.

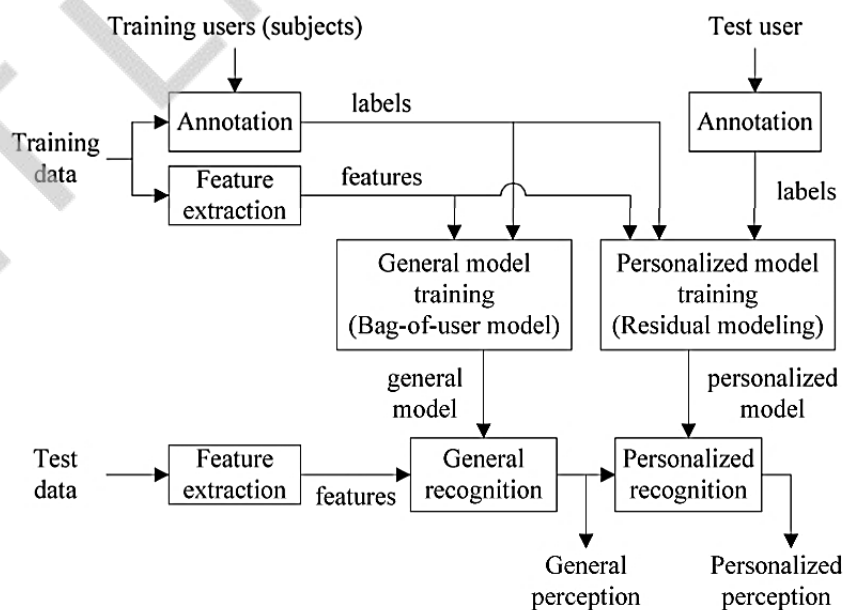


Fig. 3.9 A schematic diagram of the personalized system [33]

The regression models are trained using SVR. Squared sample correlation coefficient R^2 is measured between the ground truth and the estimated values. The value of R^2 ranges from 0 to 1; an R^2 of 1.0 means perfect fit. Comparing the bag-of-users and the baseline models, the R^2 of arousal is both around 0.70, the bag-of-users model improves the R^2 of valence from 0.149 to 0.158, a 6.4% relative improvement.

Aggregation of individual perceptions of the subjects is better accomplished using a bag-of-users model. Residual modelling focuses on the music content and user perception in various stages. These methods can also be used for other applications that involve subjective human perception.

d) **E. M. Schmidt and Y. E. Kim, “Prediction of time-varying musical mood distributions from audio,” in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2010, pp. 465–470 [36].**

The ambiguities in emotion perception can be removed by modelling emotional ground-truth as a probability distribution instead of viewing musical mood as a singular value. It provides a more realistic and accurate reflection of the perceived emotions conveyed by a song. In the given approach, the human response labels to music are modelled as a stochastic distribution in the arousal-valence (A-V) plane. Techniques of multi-variate parameter regression like Multiple Linear Regression (MLR), Partial Least-Squares (PLS) Regression, and Support Vector Regression (SVR) are used for mapping the acoustic features to the AV mood space. The data is represented by a single two-dimensional Gaussian distribution. A dataset consisting of 15-second music clips from 240 songs was selected using the original label set, to approximate an even distribution across the four primary quadrants of the A-V space. The system can track the changes in emotion on a short-time basis. The time varying features are taken into account for tracking of emotions over time. KL divergence and Euclidean distance are used for the comparison of probability distribution. The feature fusion system used in this approach is a combination of the outputs from the individual feature regression systems. For the time-varying approach, regressors are developed that can predict the emotion for a single second using only current and past audio data. Emotion regression over time is observed by plotting the collected and projected distributions of three music clips at one second intervals. Combining MLR in multiple stages produces results comparable to more computationally complex methods. The results are shown in Fig. 3.10.

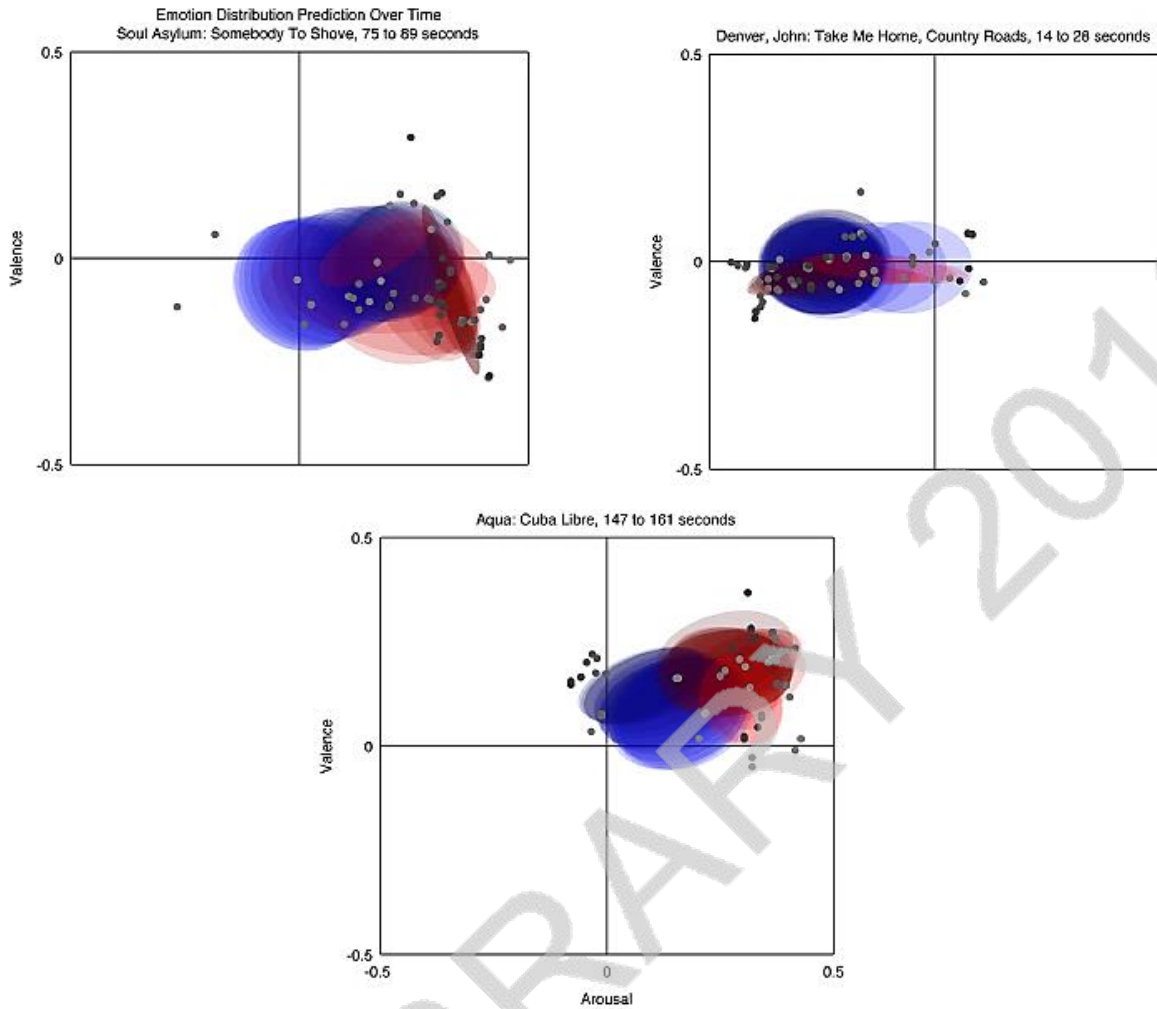


Fig. 3.10 Time-varying emotion distribution regression results for three example 15-second music clips (markers become darker as time advances) [36]

e) **E. M. Schmidt and Y. E. Kim, “Modeling musical emotion dynamics with conditional random fields,” in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2011, pp. 777–782 [35].**

Human emotion responses gradually change timely along with the music. This means that the dynamic nature of music needs to be predicted for every short interval of time. It is not just that for every interval we need to define some models but also training the model so as to adjust to the change in the music pattern. In this work they have modelled such relationships using a conditional random field (CRF), a powerful graphical model which is trained to predict the conditional probability $p(y|x)$ for a sequence of labels y given a sequence of features x . Treating the features as deterministic, they retained the rich local details present in the data, which can be used for content-based audio analysis when the data is available in abundance.

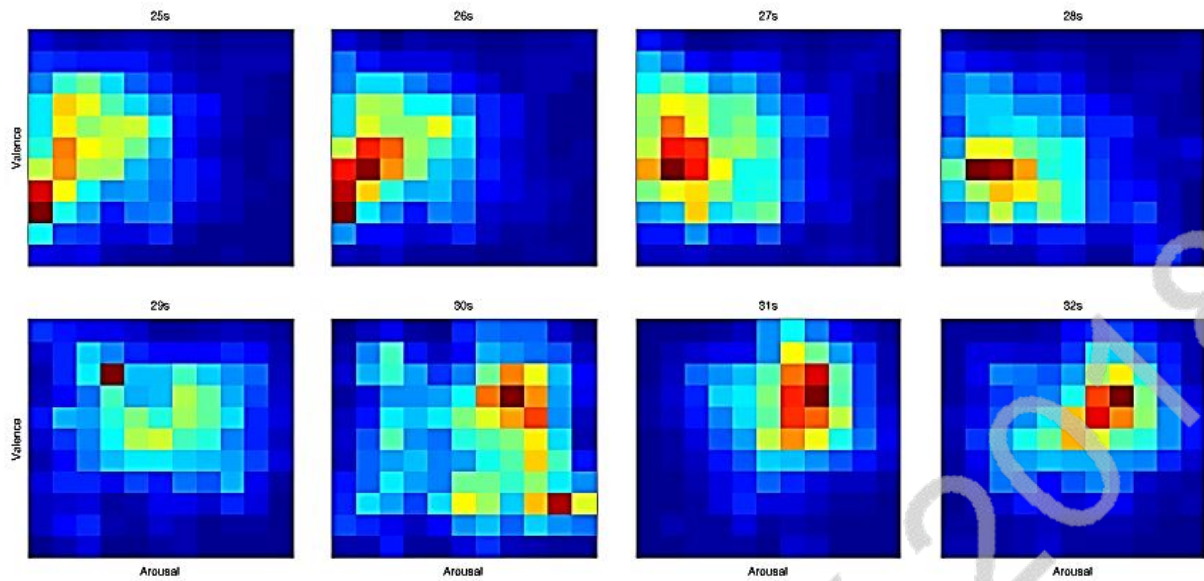


Fig. 3.11 Emotion space heatmap prediction using conditional random fields. Shown is the predicted emotion from the beginning of the song “Something About You,” by Boston [35].

They trained their graphical model on the emotional responses of individual annotators in an 11×11 quantized representation of the arousal-valence (A-V) space. Their model can produce estimates of the conditional probability for each A-V bin, allowing them to easily model complex emotion-space distributions (e.g. multimodal) as an A-V heatmap.

Fig. 3.11 demonstrates the system tracking the emotion through the low-energy, negative-emotion introduction, and through the transition at second 29 into a high-energy, positive emotion rock verse. In these figures, red indicates the highest density and blue is the lowest.

They successfully implemented their model based on the annotations for short-time intervals. Thus, for every interval, variations were seen in the A-V plot of heatmap. This provided sheer idea for determining the emotion contained in the song. This will greatly help us in designing a model which determines the emotion contained in the song thus telling the user if the song is happy, sad, relaxing, and so on.

f) **Y.-H. Yang and H. H. Chen, “Prediction of the distribution of perceived music emotions using discrete samples,” IEEE Trans. Audio, Speech Lang. Process., vol. 19, no. 7, pp. 2184 2196, Sep. 2011 [30].**

The perceived emotion of a music clip varies from person to person depending on several personal and situational factors. To take into account the subjectivity of emotion perception, the emotion distribution of a music clip is directly predicted from the music

features. In this approach, instead of assigning an emotion label or emotion value to the clip, the probability of the perceived emotion of a clip is computed at any point in the emotion plane. The emotion distribution is predicted by estimating the emotion mass at discrete samples in the emotion plane.

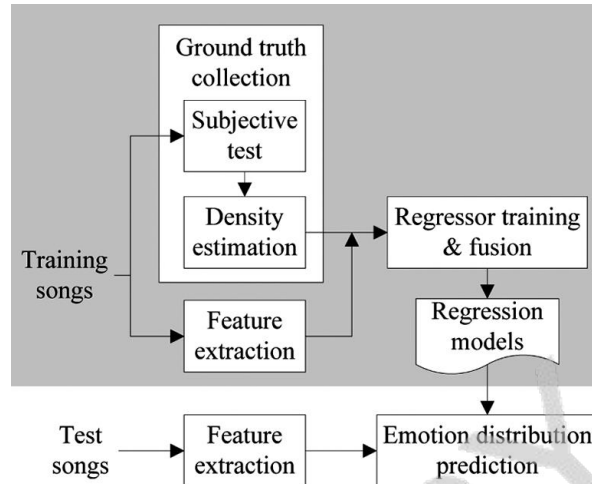


Fig. 3.12 Schematic diagram of emotion distribution prediction [30]

Emotion corresponds to dimensional perspective therefore each dimension is quantized by G equally spaced discrete samples, so a $G \times G$ grid representation of 2D emotion plane is obtained.

Fig. 3.12 illustrates the process of predicting the emotion distribution. The steps involved are:

1. **Ground Truth collection:** The VA values of a music clip are annotated by a point in the emotion plane and based on these annotations the emotion mass $y_{s,ij}$ of each discrete sample is estimated by using Kernel Density Estimation.
2. **Regressor Training and Regressor fusion:** G^2 regressors are trained for each emotion point. Support vector regression is employed for minimizing the mean square error between ground truth and prediction. The model fusion algorithms are developed to linearly fuse the features of different perceptual dimensions of music listening, such as melody, timbre, and rhythm.
3. **Prediction:** The features are extracted and trained regressors are applied to compute the emotion distribution of any new music clip.

The accuracy of this method is better as compared to Single Gaussian approach. Emotion distribution close to the ground truth, with average squared sample correlation coefficient (R^2) of 0.5439 for emotion prediction can be predicted by this approach. Since the music clips are

represented by distribution of emotions rather than points, a rich description of music emotion is achieved.

g) **M. Soleymani, M. N. Caro, E. Schmidt, C.-Y. Sha, and Y.-H. Yang, “1000 songs for emotional analysis of music,” in *Proc. Int. Workshop Crowdsourcing Multimedia*, 2013, pp. 1–6 [34].**

It becomes necessary to collect data from human subjects due to the perceptual nature of musical emotion. Complicated legal issues occur while sharing the contents of music libraries, even for academic purposes. A new publicly available dataset for music emotion recognition research and a baseline system is proposed in this paper. Creative commons music from the Free Music Archive is used in the dataset. It can be shared freely without penalty. The final dataset contains 1000 songs, each annotated by a minimum of 10 subjects, which is larger than many currently available music emotion datasets. The annotations were carried out via crowdsourcing using Amazon’s Mechanical Turk. The aim was to collect continuous V-A ratings, as well as a single discrete (9 point) A-V ratings applied to the entire clip. A personalized online annotation interface for music was developed which was a continuous annotation interface for valence, where workers need to slide the ball from left to right indicating the current emotion.

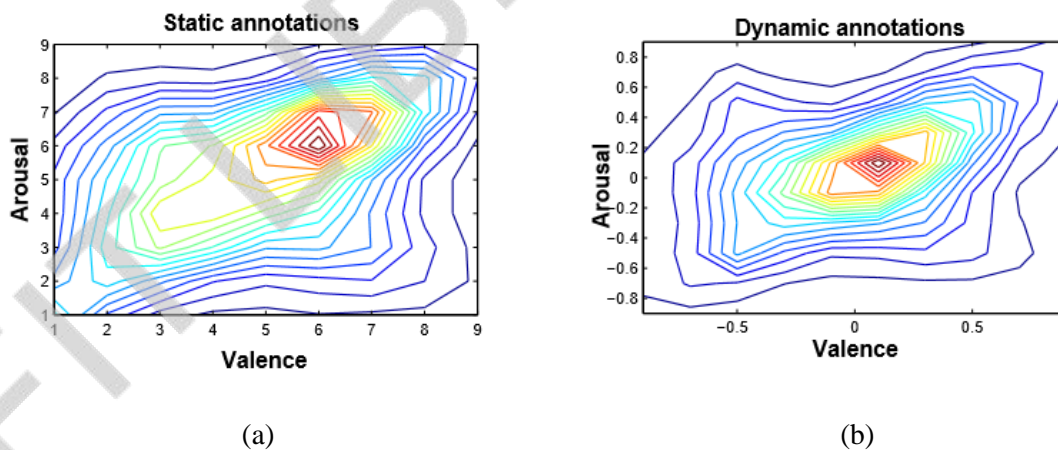


Fig. 3.13 Contours representing the distribution of annotation in case of static (left) and dynamic (right) annotations [34].

Multivariate linear regression, a simple and generalizable prediction method, was selected for the baseline system. The arousal estimations are far better than valence estimations which are in the order of chance level for static ratings. Consistently, arousal estimation results are superior to valence estimation on the continuous, dynamic affect estimation task.

The analysis on the annotations showed that there is a higher agreement in arousal ratings compared to the valence ratings. The time of the day and workers' reported "energetic" mood had a small but significant effect on the ratings. Fig.3.13 shows that both dynamic and static annotations were collected which would help the study of emotional trend on perception of music in future.

h) **Y. Vaizman, B. McFee, and G. Lanckriet, "Codebook-based audio feature representation for music information retrieval," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1483–1493, Oct. 2014 [37].**

This approach looks for compact audio content representations that will be powerful for two different MIR applications: query-by-tag and query-by-example. In query-by-tag, the system ranks music items according to relevance to a tag word describing some semantic meaning of the desired music (emotional content, specific instruments, musical style, etc.). In query-by-example the system ranks music items according to relevance or similarity to a given music example. For both these search interfaces some annotation or indexing of the songs in the repository is required. Pre-existing meta-data of the music (e.g. title, artist, lyrics, genre, instruments) is one source of such annotations which can assist in retrieving desired items. In collaborative filtering the past records of user preferences, either of specific users, for personalization purposes, or of crowds of users, are used for general recommendation.

The encoding pooling to get a compact representation for each song scheme is comprised of three stages:

- 1) Short time frame features: Each song is processed to a time series of low-level feature vectors, $X \in R^{dxT}$ (T time frames, from each a d dimensional feature vector is extracted).
- 2) Encoding: each feature vector $x_t \in R^d$ is then encoded to a code vector $c_t \in R^k$, using a pre-calculated dictionary $D \in R^{dxk}$, a codebook of k "basis vectors" of dimension d . We get the encoded song $\in R^{k \times T}$.
- 3) Pooling: the coded frame vectors are pooled together to a single compact vector $v \in R^k$.

This approach is also known as the bag of features (BoF) approach: where features are collected from different patches of an object to form a variable-size set of detected features. The pooling stage helps to have a unified dimension to the representations of all songs, regardless of the songs' durations.

In query-by-example, LASSO and CS can suffer sharp decrease of performance when adjusting their sparsity parameters, VQ is more robust, having smooth and controlled change

in performance when adjusting its density parameter. An efficient encoding method (VQ) can successfully compete with the more sophisticated method (the LASSO), achieving comparable, and even better performance, with much less computing resources.

i) Y.-A. Chen, J.-C. Wang, Y.-H. Yang, and H. H. Chen, “Linear regression-based adaptation of music emotion recognition models for personalization,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 2168–2172 [38].

Personalization techniques can be applied to address the subjectivity issue of music emotion recognition, which is important for music information retrieval. However, achieving satisfactory accuracy in personalized music emotion recognition for a user is difficult because it requires an impractically huge amount of annotations from the user. In this paper, they have adopted a probabilistic framework for *valence-arousal* music emotion modeling and proposed an adaptation method based on linear regression to personalize a background model in an online learning fashion. They also incorporated a component-tying strategy to enhance the model flexibility. Comprehensive experiments were conducted to test the performance of the proposed method on three datasets, including a new one created specifically in this work for personalized music emotion recognition. The results obtained demonstrate the effectiveness of the proposed method.

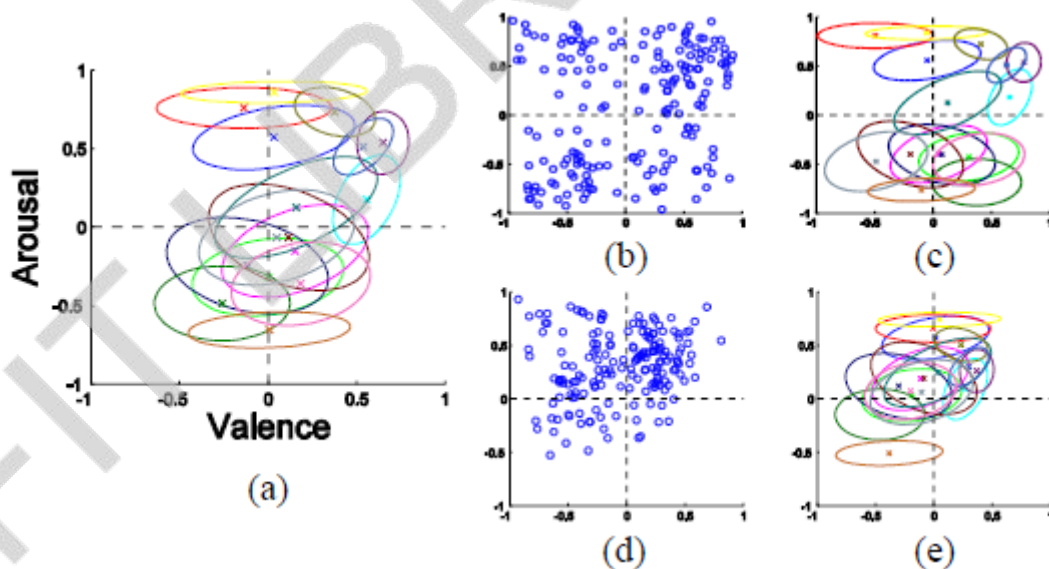


Fig. 3.14 Qualitative illustration of the performance of PMER. (a) The background VA GMM trained with MER60, each ellipse stands for a Gaussian component. (b) Distribution of the test annotations labeled by Subject #1 in AMG240, each circle corresponds to the VA annotation of a song. (c) VA GMM personalized by MAPLR with 16 annotations labeled by Subject #1. (d) Distribution of the test annotations labeled by Subject #2 in AMG240. (e) VA GMM personalized by MAPLR with 16 annotations labeled by Subject #2. [38]

In this paper, they had proposed an LR-based model adaptation method to personalize a background MER model in an online fashion. The qualitative illustration of the PMER performance is indicated in Fig. 3.14. The proposed method works effectively across a wide range of available data for adaptation and is particularly useful when only a limited amount of adaptation data is available from the user.

j) **Chen, Sih-Huei, Yuan-Shan Lee, Wen-Chi Hsieh, and Jia-Ching Wang.** "Music emotion recognition using deep Gaussian process." In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific*, pp. 495-498. IEEE, 2015 [50].

In this work, the categorical approach is utilized to detect emotions in music. The Gaussian process (GP) has been used as a powerful probabilistic framework for solving regression and classification problems with complex data. A GP can be specified by the mean vector and covariance matrix. The mean vector is commonly assumed to be zero. The covariance matrix, which is obtained from a kernel function, can express the relationship among data points. The field of deep learning is used. Deep hierarchies are constructed by stacking several models. Schmidt and Kim employed the deep belief network (DBN) to learn the sparse feature for music. Considering the advantage of GP and deep learning, Lawrence have GP can be used in deep hierarchical structure by stacking them. Deep Gaussian process provides structural learning in Gaussian process model.

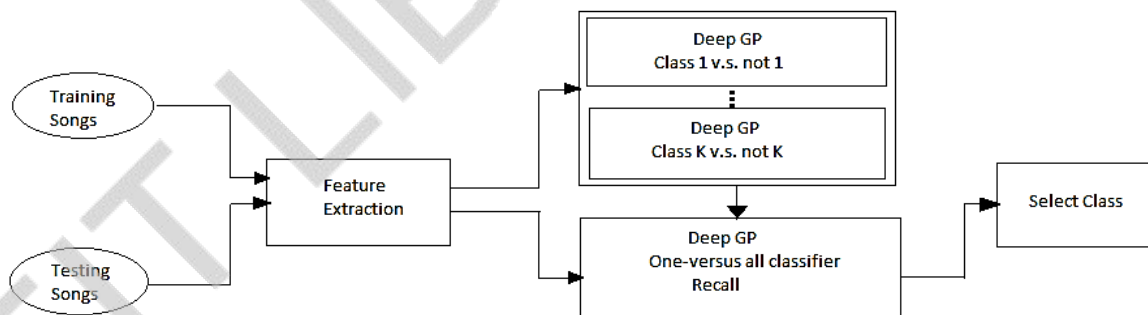


Fig. 3.15 Block diagram of Deep GP [50]

This paper presents a system for recognizing emotion in music that is based on a deep Gaussian process. The potential and efficiency of learning methodologies with deep architecture have been proved in statistical machine learning. Moreover, the Gaussian process can provide the uncertainty of predictions because of probabilistic properties. Therefore, the deep Gaussian process is a powerful approach to capture the relationship among nonlinear data. In this study, deep GP is utilized for music emotion recognition. This paper constructs a system for

recognizing emotion in music that is based on a deep Gaussian process. The procedure of proposed system is shown in Fig. 3.15. In the feature extraction part, 15 acoustical features that are often used in music emotion research are extracted from each music clip. In the classification part, a deep Gaussian process is used for recognizing emotion. For $K > 2$ classes, the use of K one versus-all classifiers, each of which solves a two-class problem, is considered. Two-fold cross-validation is used to evaluate the performance of the proposed system and compared with SVM and standard GP.

k) **Fukuyama, Satoru, and Masataka Goto.** "Music emotion recognition with adaptive aggregation of Gaussian process regressors." In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 71-75. IEEE, 2016 [51].

Music emotion recognition works by extracting features from music audio and applying a regression or classification model to map those features into an emotion representation. This approach is based on a multi-stage regression, which aggregates the results from different regressors trained with training data. However, after training, the aggregation happens in a fixed way and cannot be adapted to acoustic signals with different musical properties.

Therefore, the aggregation is adapted by taking into account new acoustic signal inputs. Gaussian process (GP) regression is used to adjust the importance of each feature depending on the audio input. GP regression can predict the mean and variance of an estimated Arousal Valence (AV) value from a new audio input as shown in Fig. 3.16. By preparing multiple GP regressors wherein each is trained with a different feature, the means and variances of the estimated AV values are obtained considering each feature. The feature is not important when the feature values in the training data take various values for similar AV values, and this leads to a large variance when an AV value is estimated with GP regression from a new piece of audio. Thus, the importance of each feature is obtained by calculating the variance of AV values with individual GP regressors. Since the importance of each feature varies depending on the audio input, the results from the individual regressors are aggregated adaptively to estimate the emotion.

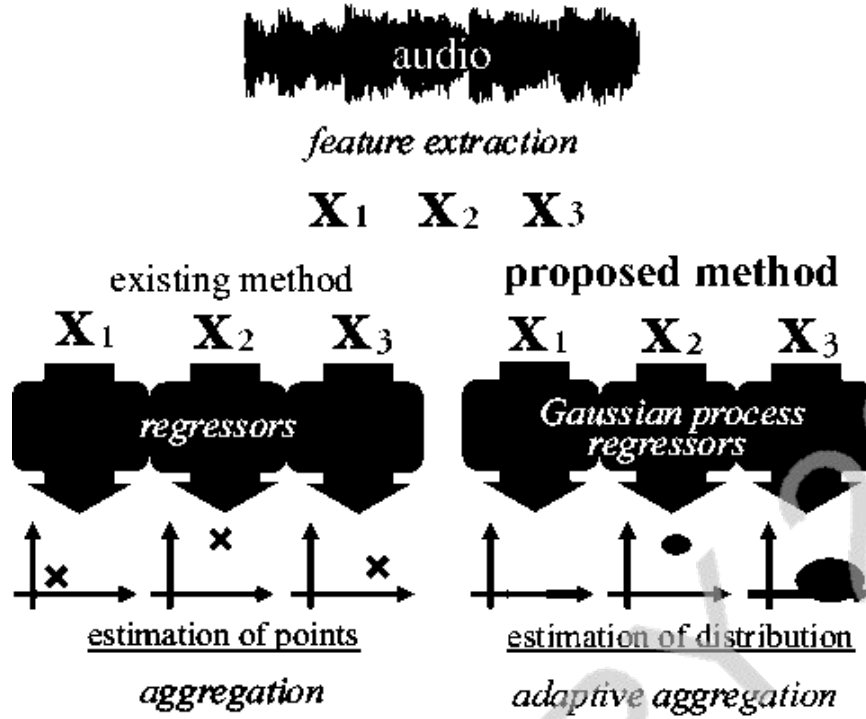


Fig. 3.16 Fixed aggregation vs. proposed adaptive aggregation of Gaussian process regressors for music emotion recognition [51].

The steps involved are:

- 1) Extracting low-level audio descriptors related to emotion from the audio.
- 2) Generating features by creating three different feature sets by dividing the original 6373-dimensional vector into three collecting descriptors related to the spectral descriptors, MFCCs, etc.
- 3) Constructing GP regressors using Emotion in Music Database as training data which consists of 744 audio clips and annotations of AV values on a scale from 1 to 9 normalized from -1.0 to 1.0. Out of 744 clips, 619 randomly chosen clips are used as training data and the rest (125 clips) as evaluation data.
- 4) Aggregating the results from regressors to obtain the AV value estimate The AV value are calculated as a weighted average of the means of the estimated Gaussian distributions from the GP regressors, wherein the weights were set to the normalized inversed square of the variances. When conducting the fixed aggregation, the training data (619 clips) is split into 309 clips (to construct the GP regressors) and 310 clips (for multivariate regression between the results from the GP regressors and ground-truth).

The performance is evaluated by using 10-fold cross validation and calculating R -squared value (R^2) and Root Mean Square Error (RMSE).

1) **An, Yunjing, Shutao Sun, and Shujuan Wang.** "Naive Bayes classifiers for music emotion classification based on lyrics." In *Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference on*, pp. 635-638. IEEE, 2017 [52].

For automatic classification of Chinese music emotions in a more effective manner, this work uses the lyrics of music to analyze and classify music based on emotion. It uses Naive Bayes algorithm which is simple but can classify text effectively. Naive Bayesian Classifier is a simple classifier based on applying Bayes theorem with independence assumptions.

The Naive Bayes classifier uses the Naive Bayesian formula to calculate the probability of each class A given the values B_i of all attributes for an instance to be classified. The conditional independence of the attributes is defined using Eq. (3.2).

$$P(A|B_1 \dots B_n) = P(A) \prod_i \frac{P(A|B_i)}{P(A)} \quad (3.2)$$

The work assumes that the words of the lyrics are independent and of equal weight. The music emotion labels used are from Baidu music emotion labels. The emotion labels of the Baidu music contain sad, passionate, quiet, comfortable, sweet, inspirational, lonely, miss, romantic, yearning, joyful, soulful, happy, nostalgic, relaxed. The Scrapy, a framework of Python, is used to achieve a simple crawler to crawl the related information of Baidu music which contain the singer, music name, lyrics and the category of the music. The emotion classification model used is the Thayer model.

As the transform of Baidu music labels to the category of anxiety is inappropriate, the emotion labels of Baidu are transformed to three categories, namely contentment, depression and exuberance as shown in Table 3.2.

Table 3.2. Map of the Emotion Labels [52]

Emotion labels of our work	Emotion labels of Baidu
Contentment	Quiet, Yearning, Romantic, Sweet, Healing
Depression	Waiting, Sad, Frustrated
Exuberance	Passionate, Inspirational, Happy, Joyful

The performance is evaluated using threefold cross validation. Each lyric of song used for experiment is segmented and the emotional words are picked up. A Python module named

Jieba is used for word segmentation. In order to validate the influence of different datasets on the classification performances, four different datasets are used for training and the results are compared. The result for one of the four datasets is as shown in Table 3.3

Table 3.3 Experimental Result of Dataset-2 [52]

Original Label \ Test Result	Depression	Contentment	Exuberance
Depression	584	57	160
Contentment	64	13	34
Exuberance	85	20	124

The final accuracy reported is approximately 68%.

Chapter 4

Music Emotion Recognition using Acoustic Gaussian Mixture Model

4.1 System Block Diagram:

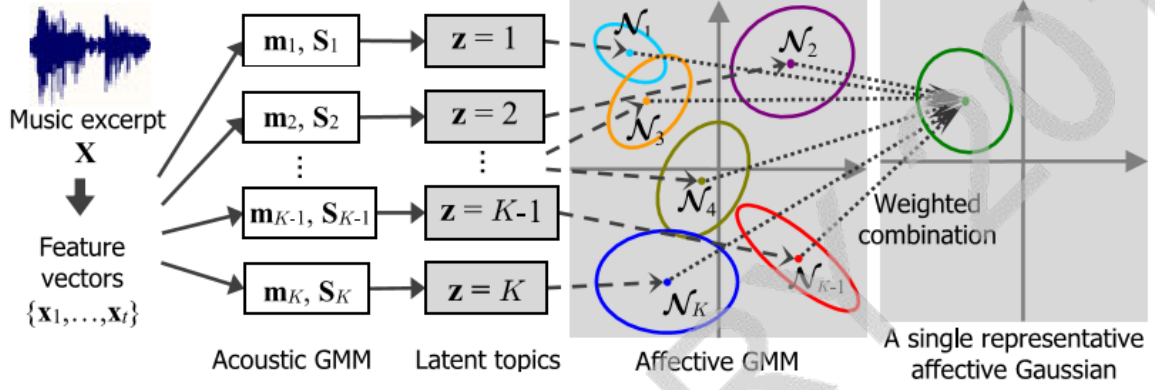


Fig. 4.1: System block diagram [26]

The song preview (.wav files) of 30 seconds each of 1608 songs from AMG1608 dataset was used for the implementation of the proposed method. This audio data of music excerpt \mathbf{X} $\{\mathbf{x}_1, \dots, \mathbf{x}_T\}$, $\mathbf{x}_t \in R^M$, the position of excerpt on the continuous valence and arousal emotional space $\mathbf{y} \in R^2$, and the associated discrete latent topic $\mathbf{z} \in \{1, 2, \dots, K\}$. The model assumptions made are:

- 1) We have the graphical structure $\mathbf{X} \rightarrow \mathbf{z} \rightarrow \mathbf{y}$, implying that the emotion \mathbf{y} is independent of the audio data \mathbf{X} when a given topic \mathbf{z} .
- 2) The distribution of an arbitrary frame \mathbf{x} given \mathbf{z} is Gaussian

$$p(\mathbf{x}|\mathbf{z} = k) \sim N(\mathbf{m}_k, \mathbf{S}_k) \quad (4.1)$$

Accordingly, we have the probability density function for \mathbf{x}

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k N(\mathbf{x}|\mathbf{m}_k, \mathbf{S}_k) \quad (4.2)$$

where π_k , \mathbf{m}_k and \mathbf{S}_k are the model parameters associated with the k^{th} latent topic.

Then, given an observed frame \mathbf{x}_t , the posterior probability of a topic is computed by

$$p(\mathbf{z} = k|\mathbf{x}_t) = \frac{\pi_k N(\mathbf{x}_t|\mathbf{m}_k, \mathbf{S}_k)}{\sum_{h=1}^K \pi_h N(\mathbf{x}_t|\mathbf{m}_h, \mathbf{S}_h)} \quad (4.3)$$

- 3) The excerpt-level posterior probability of $\mathbf{z} = k$ given \mathbf{X} can be approximated by averaging the frame-level posterior probabilities,

$$p(\mathbf{z} = k|\mathbf{X}) \approx \frac{1}{T} \sum_{t=1}^T p(\mathbf{z} = k|\mathbf{x}_t) \quad (4.4)$$

In other words, it is assumed that every frame of the excerpt has equal contribution.

- 4) The distribution of \mathbf{y} given a topic $\mathbf{z} = k$ is Gaussian.

$$p(\mathbf{y}|\mathbf{z} = k) \sim N(\boldsymbol{\mu}_k, \Sigma_k) \quad (4.5)$$

where the parameters μ_k, Σ_k are associated with the k^{th} latent topic as well.

Accordingly, the marginal distribution of \mathbf{y} is

$$p(\mathbf{y}|\mathbf{X}) = \sum_k p(\mathbf{y}|\mathbf{z} = k)p(\mathbf{z} = k|\mathbf{X}) \quad (4.6)$$

$$= \sum_k N(\mathbf{y}|\boldsymbol{\mu}_k, \Sigma_k)p(\mathbf{z} = k|\mathbf{X}) \quad (4.7)$$

We make the following observations:

To model the complicated relationship between the input and output data, a hidden layer is introduced. To achieve this \mathbf{z} uses K discrete latent topics to connect the acoustic feature space and emotion space.

The model $\{\pi_k, \mathbf{m}_k, \mathbf{S}_k\}_{k=1}^K$ is the acoustic GMM. The total number of parameters is $K + MK + M^2K$ which can be reduced to $2MK$ by assuming each $\pi_k = \frac{1}{K}$ and each \mathbf{S}_k to be diagonal which implies no correlation among the features.

The encoding approach similar to bag-of-frames (BoF) is adopted. The BoF approach uses a codebook of size K to quantize a frame-level input \mathbf{x}_t as a codeword and assumes that the excerpt-level information can be represented by the histogram over the codewords of the codebook. The frame-level encoding result is computed as a probability $p(\mathbf{z}|\mathbf{x})$ which has been proven to be more effective than the conventional BoF approach with smaller K , when the acoustic GMM is used.

Owing to the subjective nature of emotion perception, the Gaussian model of the emotion space is intended to parameterize the emotion distribution.

The model $\{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$ is the affective GMM since the K topics collectively define a GMM for the emotion data.

The four assumptions suggest that music excerpts sharing similar distributions in $p(\mathbf{z}|\mathbf{X})$ would also have similar distributions in the emotion space, so that the annotations of different excerpts can blend with one another.

The parameters $\Theta \equiv \{\pi_k, \mathbf{m}_k, \mathbf{S}_k, \boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$ are statistical and can be estimated from data using maximum likelihood estimation.

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \sum_{i=1}^N \log p(\mathbf{Y}^{(i)} | \mathbf{X}^{(i)}, \Theta) \quad (4.8)$$

where N denotes number of available excerpts for training and (i) denotes the i^{th} training excerpt.

As Fig. 4.1 depicts, AEG involves a generative process for the affective content of music. For a given excerpt, we first extract the frame-level feature vectors, compute $p(\mathbf{z}|\mathbf{X})$ according to Eqs. (4.3) and (4.4), and then compute $p(\mathbf{y}|\mathbf{X})$ according to Eq. (4.7). If an excerpt's acoustic content \mathbf{X} can be completely described by a single latent topic $\mathbf{z} = k$, i.e. $p(\mathbf{z} = k|\mathbf{X})=1$ and $p(\mathbf{z} = h|\mathbf{X})=0 \ \forall h \neq k$; its emotion distribution would exactly follow $N(\boldsymbol{\mu}_k, \Sigma_k)$. Otherwise, the emotion distribution would be a weighted combination of $\{N(\boldsymbol{\mu}_k, \Sigma_k)\}_{k=1}^K$ using $p(\mathbf{z}|\mathbf{X})$ as the weights [26].

4.2 Fitting the Acoustic GMM

The model fitting process can be simplified by dividing the parameter set to $\{\pi_k, \mathbf{m}_k, \mathbf{S}_k\}_{k=1}^K$ (acoustic GMM) and $\{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$ (affective GMM) and fitting them separately.

The acoustic GMM is conceptually similar to a codebook that is applicable to describe the acoustic feature bases for any music excerpt, but it is more general due to its probabilistic treatment. As the representative for a topic, moreover, the use of Gaussian distribution in fact conveys more semantic meanings than a single codeword does. An acoustic GMM can be learned from a collection of unlabeled frame-level acoustic feature vectors \mathbf{U} , whose size can be arbitrarily large since no human annotation is needed. The features include root-mean-square energy, zero-crossing rate, spectral flux, centroid, spread, skewness, kurtosis, entropy, flatness, 85 percent-rolloff, 95 percent-rolloff, brightness, roughness, irregularity, 13-dimensional MFCCs, delta MFCCs, delta-delta MFCCs, key clarity, musical mode, harmonic changes likelihood, 12-bin chroma vector, chroma peak and chroma centroid, leading to a 70-dimensional feature vector for a frame. This is a typical problem of learning a universal background model in the speech processing field and can be tackled by the expectation-

maximization (EM) algorithm. As aforementioned, we fix $\pi_k = \frac{1}{K}$ and assume each \mathbf{S}_k to be diagonal for simplicity. Once the parameters $\{\mathbf{m}_k, \mathbf{S}_k\}_{k=1}$ are learned, the resulting acoustic GMM can be used to compute $p(\mathbf{z}|\mathbf{X})$ for a music excerpt [26].

4.3 Model the Annotation Prior

To obtain the general emotion response of an excerpt in \mathcal{L} , we typically ask multiple listeners to annotate the excerpt. However, as some listeners' annotations might not be reliable. To improve the robustness of AEG, we can introduce a variable γ to weight the importance of different annotations in learning the affective GMM. For example, if we have known that a user may be biased or less consistent with others, his/her annotations can be considered as less reliable. In our prior work [29], we develop an intuitive approach to setup the annotation prior for each annotation by

$$\gamma_j^{(i)} \leftarrow \frac{N(\mathbf{y}_j^{(i)}|\mathbf{a}^{(i)}, \mathbf{B}^{(i)})}{\sum_h N(\mathbf{y}_h^{(i)}|\mathbf{a}^{(i)}, \mathbf{B}^{(i)})} \quad (4.9)$$

where $\mathbf{a}^{(i)}$ and $\mathbf{B}^{(i)}$ are the sample mean and covariance of $\mathbf{Y}^{(i)}$ computed beforehand. For simplicity, we use a single Gaussian instead of a GMM as the prior, so there is no need to determine the number of components of the GMM. Note that this setting does not contradict our motivation to model the affective content of music as a GMM, as explained below. In some sense, $\gamma_j^{(i)}$ can be viewed as a regularizer of the algorithm to fit the affective GMM in addition to reflecting the annotation importance. The intuition is that $\gamma_j^{(i)}$ tends to regularize the parameters $\{\boldsymbol{\mu}_k, \Sigma_k\}$ to stay close to $\{\mathbf{a}^{(i)}, \mathbf{B}^{(i)}\}$ if $p(\mathbf{z} = k|\mathbf{y}_j^{(i)})$ large. This shows that the resulting $\{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$ will be diverse enough, because we always have multiple training excerpts that generate a variety of Gaussian priors $\{\mathbf{a}^{(i)}, \mathbf{B}^{(i)}\}_{i=1}^N$. We can also set a parameter $0 \leq \lambda \leq 1$ to control the trade-off between regularity and data fidelity given by [26]

$$\gamma_j^{(i)} \leftarrow (1 - \lambda) \cdot 1 + \lambda \cdot \gamma_j^{(i)} \quad (4.10)$$

4.4 Fitting the Affective GMM

Fitting the affective GMM parameters $\{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$, on the contrary, requires a labeled dataset $\mathcal{L} = \{\mathbf{X}^{(i)}, \mathbf{Y}^{(i)}\}_{i=1}^N$, where $\mathbf{Y}^{(i)} = [\mathbf{y}_1^{(i)}, \dots, \mathbf{y}_{U^{(i)}}^{(i)}]$ denotes the set of VA values entered by listeners, $\mathbf{y}_j^{(i)} \in \mathbb{R}^2$ the individual annotation of the j th listener, and $U^{(i)}$ the number of annotations available for the i th excerpt. Based on the acoustic GMM, for each training excerpt

we compute $p(\mathbf{z}|\mathbf{X}^{(i)})$, which is called the *acoustic prior* and will stay fixed in the learning process of affective GMM. Then, the data log-likelihood can be derived by

$$\begin{aligned} L &= \log \prod_{i=1}^N \prod_{j=1}^{U^{(i)}} p(\mathbf{y}_j^{(i)} | \mathbf{X}^{(i)}) \\ &= \sum_{i,j} \log \sum_k N(\mathbf{y}_j^{(i)} | \boldsymbol{\mu}_k, \Sigma_k) p(\mathbf{z} = k | \mathbf{X}^{(i)}) \end{aligned} \quad (4.11)$$

In practice, we might want to introduce the *annotation prior* for modeling the reliability of each annotation $\mathbf{y}_j^{(i)}$, giving rise to

$$\hat{L} = \sum_{i,j} \gamma_j^{(i)} \log \sum_k N(\mathbf{y}_j^{(i)} | \boldsymbol{\mu}_k, \Sigma_k) p(\mathbf{z} = k | \mathbf{X}^{(i)}) \quad (4.12)$$

where $0 \leq \gamma_j^{(i)} \leq 1$ and $\sum_{i,j} \gamma_j^{(i)} = 1$. This equation reduces to Eq. (4.11) when a uniform setting $\gamma_j^{(i)} = \frac{1}{\sum_h U^{(h)}}$ is adopted. We will describe a model for $\gamma_j^{(i)}$. One can observe from Eq. (4.12) a very important attribute of AEG, i.e., the affective GMM is learned based on the raw emotion annotations from each listener instead of the aggregated ones across subjects. This scheme directly takes the subjectivity into account, making AEG fundamentally different from the Gaussian-parameter approach.

Although maximizing \hat{L} is intractable, we can employ the EM algorithm to find an approximated solution. In the E-step, we compute the posterior probability of $\mathbf{z} = k$ given $\mathbf{y}_j^{(i)}$,

$$p(\mathbf{z} = k | \mathbf{y}_j^{(i)}) = \frac{p(\mathbf{z} = k | \mathbf{X}^{(i)}) N(\mathbf{y}_j^{(i)} | \boldsymbol{\mu}_k, \Sigma_k)}{\sum_h p(\mathbf{z} = k | \mathbf{X}^{(i)}) N(\mathbf{y}_j^{(i)} | \boldsymbol{\mu}_h, \Sigma_h)} \quad (4.13)$$

In the M-step, the updating forms for the mean vector and covariance matrix are as follows:

$$\boldsymbol{\mu}'_k \leftarrow \frac{\sum_{i,j} \gamma_j^{(i)} p(\mathbf{z} = k | \mathbf{y}_j^{(i)}) \mathbf{y}_j^{(i)}}{\sum_{i,j} \gamma_j^{(i)} p(\mathbf{z} = k | \mathbf{y}_j^{(i)})} \quad (4.14)$$

$$\Sigma'_k \leftarrow \frac{\sum_{i,j} \gamma_j^{(i)} p(\mathbf{z} = k | \mathbf{y}_j^{(i)}) (\mathbf{y}_j^{(i)} - \boldsymbol{\mu}'_k) (\mathbf{y}_j^{(i)} - \boldsymbol{\mu}'_k)^T}{\sum_{i,j} \gamma_j^{(i)} p(\mathbf{z} = k | \mathbf{y}_j^{(i)})} \quad (4.15)$$

The EM algorithm iteratively maximizes the value of \hat{L} defined in Eq. (4.12) until convergence. One can fix the number of maximal iterations or set a stopping criterion according to the relative increase in \hat{L} .

As Eqs. (4.14) and (4.15) show, the parameter update is collectively determined by $\mathbf{y}_j^{(i)}, \gamma_j^{(i)}$ and $p(\mathbf{z}|\mathbf{y}_j^{(i)}), \forall i, j$. In this way, the learning process jointly takes the data likelihood, annotation prior and acoustic prior over the current affective GMM into consideration, so that the annotations of different excerpts can share with one another according to their corresponding probabilities. This is another unique attribute of AEG. The algorithm summarizes the learning process of affective GMM. The initialization of the parameters $\{\boldsymbol{\mu}_k^0, \boldsymbol{\Sigma}_k^0\}_{k=1}^K$ can be obtained by, for example, using the sample mean vector and covariance matrix $\boldsymbol{\mu}_{\mathcal{L}}, \boldsymbol{\Sigma}_{\mathcal{L}}$ computed over the whole data set \mathcal{L} .

Algorithm to fit the affective GMM:

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

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

initial model $\{\mu_k^0 = \mu_L, \Sigma_k^0 = \Sigma_L\}_{k=1}^K$;

maximal number of iterations R or

threshold of stopping ratio Γ ;

OUTPUT: Model parameters $\{\mu'_k, \Sigma'_k\}_{k=1}^K$

1. Initialize $r \leftarrow 0$ and L_0 using equation (4.12);
2. repeat
3. Compute the posterior probability using equation (4.13) with $\{\mu_k^r, \Sigma_k^r\}_{k=1}^K$;
4. $r \leftarrow r + 1$;
5. Update $\{\mu_k^r, \Sigma_k^r\}_{k=1}^K$ using equations (4.14) and (4.15).
6. Compute L_r using equation (4.12)
7. Until $r = R$ or $\frac{L_r - L_{r-1}}{|L_{r-1}|} < \Gamma$;
8. Let $\mu'_k \leftarrow \mu_k^r$ and $\Sigma'_k \leftarrow \Sigma_k^r$; [26]

4.4.1 Singularity Issue in Learning the Affective GMM

In practice, as the affective GMM is getting fitted to the data, a small number of affective Gaussian components might overly fit to some emotion annotations, leading to singularity.

When this occurs, some covariance matrices become nonpositive definite (non-PD), making the corresponding affective Gaussians ill-defined. For instance, if a component affective Gaussian is contributed by only one or two annotations, the shape of its covariance becomes a point or a straight line. It is particularly important to avoid the singularity issue

when the size of training examples is too small, when there are prevalent outliers in the set of annotations, or when the value of K is set to an overly large value. A straightforward approach to circumvent this issue is to perform early stop in Algorithm 1 by setting a smaller R (e.g., 8) or a larger Γ (e.g., 0.01), or to stop the EM update whenever a non-PD covariance matrix appears. However, both methods might lead to an insufficient model. Alternatively, one can i) regularize the covariance matrices by adding small values to the diagonal, or ii) remove that ill-conditioned Gaussian component, which in effect dynamically decreases the value of K . We adopt the latter approach [26].

4.5 Predicting Emotions

1. The song preview (.wav files) of 30 seconds each of 1608 songs from AMG1608 dataset is used for the implementation of the proposed method.
2. 70 D features are extracted from each wav file to generate 1608 .mat files of dimension 70 x 1199.
3. To reduce the computational complexity, 5% values from each .mat file were extracted and stacked one after the other to obtain a feature matrix of dimension 70 x 192,960.
4. Expectation maximization algorithm is used to obtain mean, co-variance and mixing coefficients that trains the acoustic model.
5. Using the above obtained mean, co-variance and mixing coefficients, the responsibility and label of size $1 \times k$ of each song is obtained where k is the cluster size.
6. In order to account for the reliability of the annotations provided by the annotators, annotation prior is used.
7. The valence-arousal annotations for 1608 clips are stacked to form a 1608 x 1 cell containing the annotations for each clip inside a structure.
8. The annotations and the annotation prior are used to train the affective model which produces a hybrid model by combining the models with and without annotation prior.
9. For testing purpose i.e. for predicting the emotion of a given clip, it is necessary to compute the posterior probability of that clip.
10. This posterior probability and the trained affective model are used to predict the song emotion by taking the weighted average of all clusters thus producing a single representative affective Gaussian for each song.

Chapter 5

Results and Discussion

5.1 Histogram

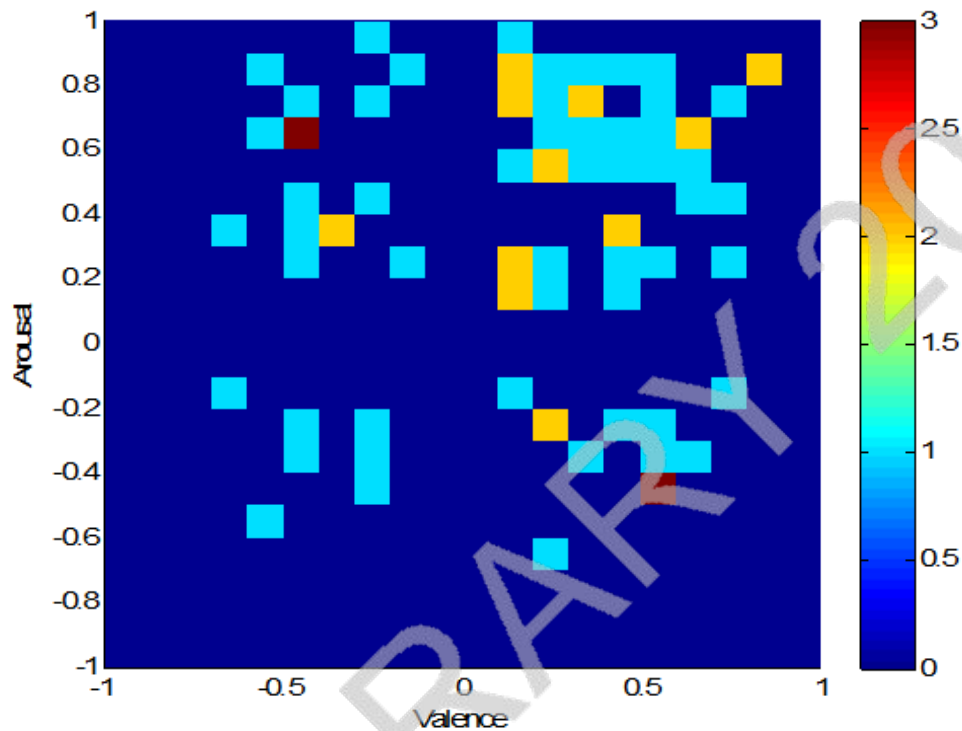


Fig. 5.1 Histogram of experimental annotation

As an experiment, we collected annotations for 20 different songs from 4 people and those annotations were plotted to form the above histogram. As observed from Fig. 5.1, region with red color has the highest number of annotations and they are present approximately in both second and fourth quadrants which mean that most of the songs had either negative valence and positive arousal or positive valence and negative arousal. The dark blue shaded regions are the one with least number of annotations because such songs are usually not preferred or usually not composed. And it can also be observed that because the number of annotations is less, we won't obtain a wide range of opinion. For a better system performance, it is preferred that annotations are taken from a large group of people, so that all possible human perception is considered.

5.2 AMG1608 Dataset

Automated recognition of musical emotion from audio signals has received considerable attention recently. To construct an accurate model for music emotion prediction, the emotion-annotated music corpus has to be of high quality. It is desirable to have a large number of songs annotated by numerous subjects to characterize the general emotional response to a song. Due to the need for personalization of the music emotion prediction model to address the subjective nature of emotion perception, it is also important to have a large number of annotations per subject for training and evaluating a personalization method. AMG1608 is a dataset for music emotion analysis. It contains frame-level acoustic features extracted from 1608 30-second music clips and corresponding valence-arousal (VA) annotations provided by 665 subjects. Furthermore, 46 subjects annotated more than 150 songs, making this dataset the largest of its kind to date [39].

The dataset consists of two parts:

- (1) The campus subset. It is a 240-song subset annotated by 22 subjects recruited from the National Taiwan University and the Academia Sinica.
- (2) The Amazon Mechanical Turk (AMT) subset. This subset contains annotations of all the 1608 songs provided by the 643 subjects using AMT. Each song receives a total of 15 emotion annotations from each subject in this subset [40].

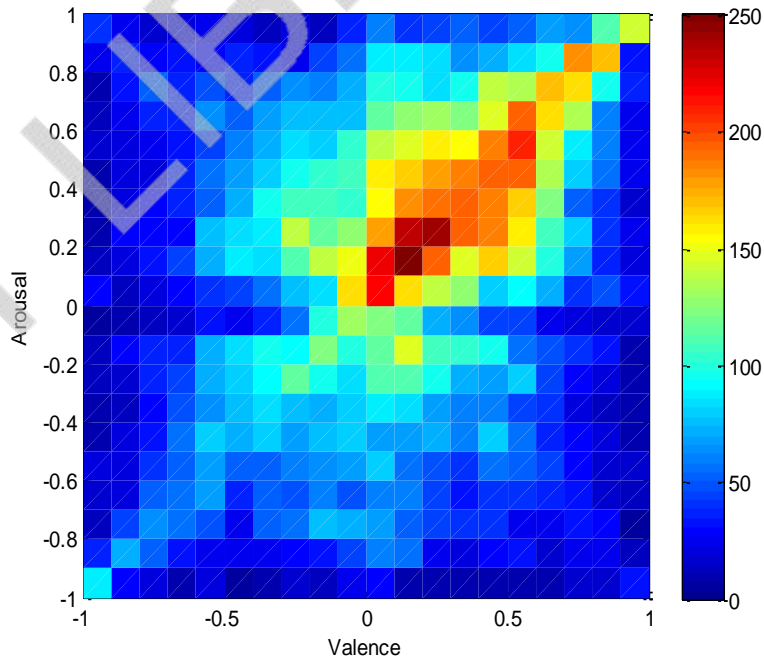


Fig. 5.2 Histogram obtained from AMG1608 dataset

Fig. 5.2 shows the histogram obtained using the annotations from AMG1608 dataset. It is observed that maximum number of annotations (indicated by red and yellow regions) is present in the first quadrant which implies that most of the songs have a positive valence and high arousal. The dark blue regions present at the edges indicate that a very few songs have extreme values of valence and arousal.

5.3 Feature Extraction

AMG1608 dataset contains the song preview of 30 seconds each (.wav files). These .wav files are processed to obtain the corresponding .mat files. These wave files are divided into frames using mirframes. These frames have 50% overlap between the consecutive frames to prevent the loss of low frequency components during windowing for DFT. By default mirframe uses hamming window with 50% overlap.

For each of these frames, 70 D features were extracted which comprised of spectral, dynamic, tonal and timbre features. Thus $70 \times N$ feature matrix was obtained for each song which are stored in .mat files.

5.4 Generation of Posterior probabilities

The 1608 .mat files each containing $70 \times N$ feature matrix were processed to obtain $1 \times k$ array. To reduce the computational complexity only 10% of the values are considered from each .mat file. Since the number of frames in each song were 1199, we randomly extracted 10% of the frames (approximately 120). These 120 frames each with 70D features for all 1608 songs were then stacked one after the other to obtain a feature matrix of dimension $70 \times 192,960$ ($192,960 = 120 \times 1608$). The system is first trained and then used to predict. Expectation maximization algorithm is used on this feature matrix to obtain the mean, co-variance and mixing coefficients of each song. This trained model is then processed to obtain the responsibility and label of size $1 \times k$. 60 frames from each song are randomly selected and the posterior probabilities are obtained that are as depicted in Fig. 5.3. The horizontal axis indicates the cluster number while the vertical axis indicates the song number. Posterior probabilities indicate the probability of a song being generated from a particular cluster. It is observed that the region for cluster number 2 is shaded with dark blue. This indicates that the probability of any song being generated from cluster 2 is very low.

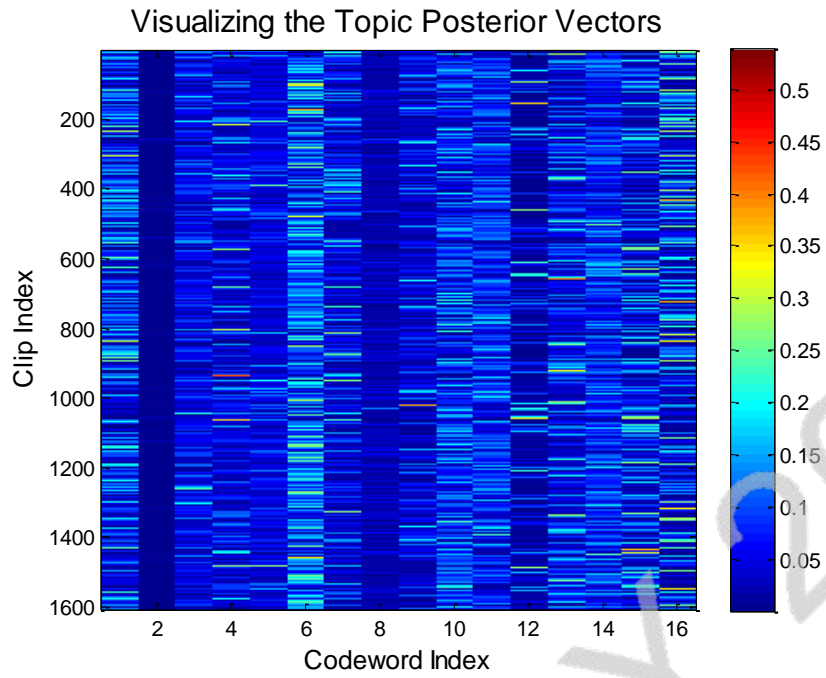
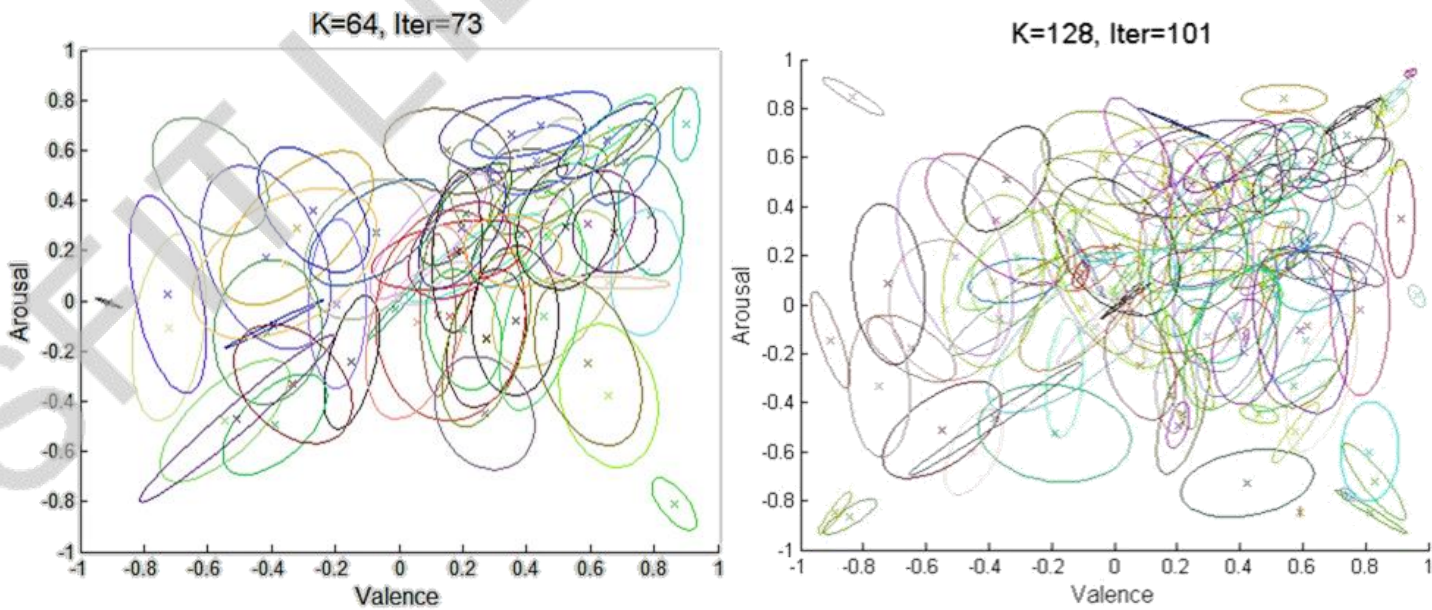


Fig. 5.3 Acoustic Posterior Probabilities

5.5 Affective GMMs

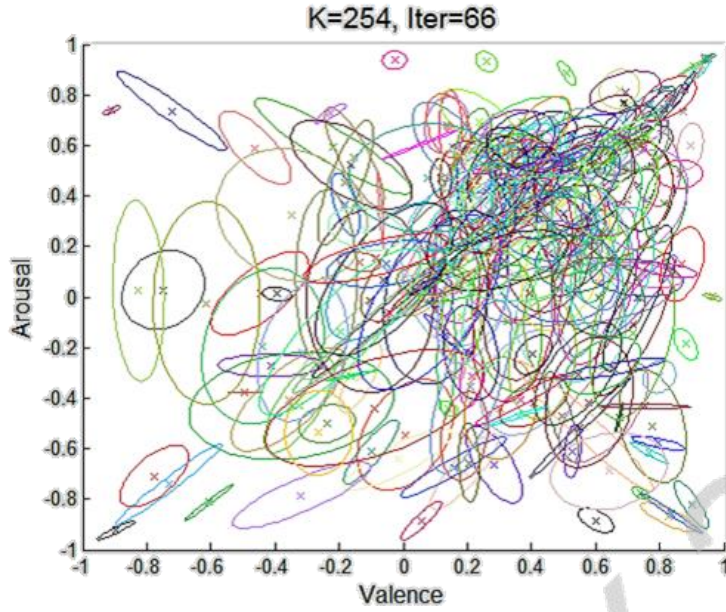
5.5.1 Affective GMM using AEG

Based on acoustic GMM, the affective GMM was learned on the annotations of AMG1608 dataset. The learned affective GMM was examined for different values of latent topics K (32, 64, 128, 256). Fig. 5.4 shows these affective GMMs obtained by using AEG.



(a)

(b)



(c)

Fig. 5.4 Affective GMMs using AEG (a) $K=64$ (b) $K=128$ (c) $K=256$

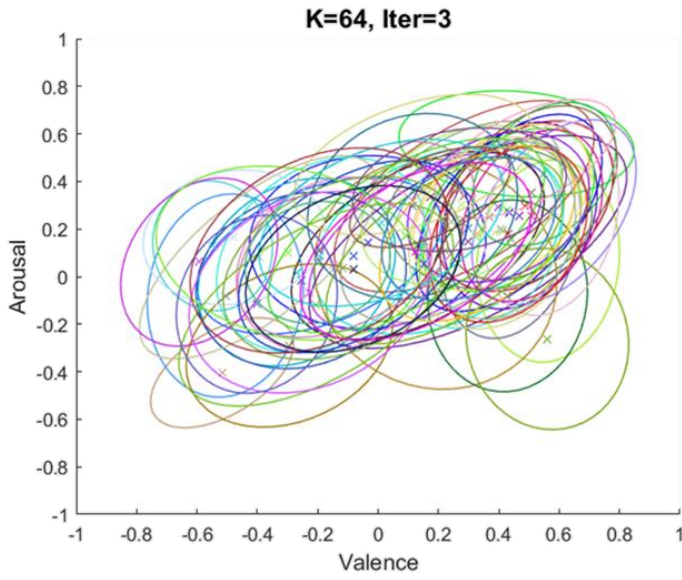
The specific area covered by a Gaussian suggests the mapping from acoustic space to the emotion space governed by a specific latent topic.

It is observed from Fig. 5.4 that,

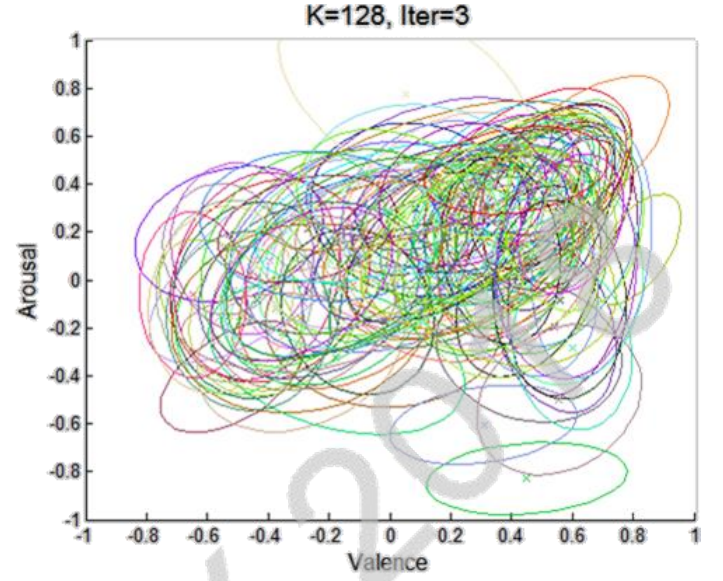
1. The Gaussians are close together in the beginning and as iterations proceed, they gradually separate from each other.
2. The covariance (size of ellipse) of each Gaussian gets increasingly smaller until convergence, when the Gaussians collectively cover the different areas in the VA space, making it possible to approximate all kinds of emotion distribution by combining the learned affective GMM with different weights set according to acoustic prior $p(\mathbf{z}|\mathbf{X})$ for each music excerpt individually.
3. The Gaussians with horizontally elongated ellipses suggest that it is more difficult to discriminate positive/negative valence as compared to high/low arousal.
4. As the number of latent topics K increases, the model covers almost the full VA space.
5. The smaller and diverse Gaussians suggest that the association between music and emotion is clearer for the listeners and thus a higher inter-user agreement.
6. The singularity issue is observed for $K=256$ at iterations 30 and 65.

5.5.2 Affective GMM using VQ

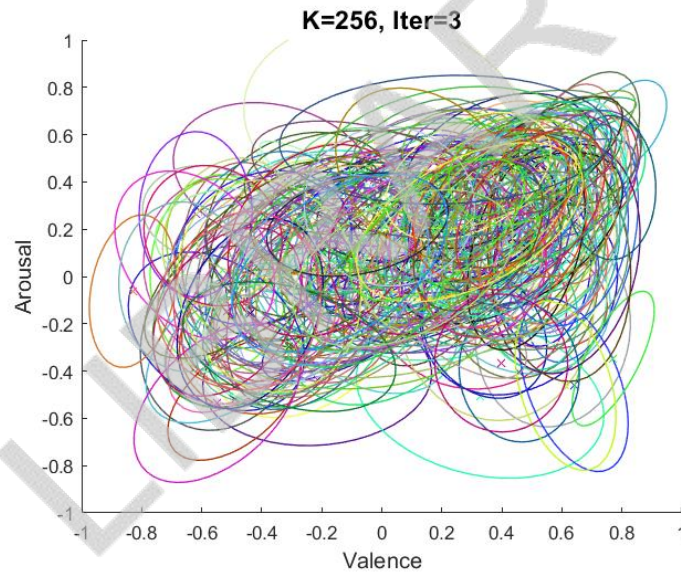
Vector Quantization is used instead of acoustic GMM to compute the posterior probability and then the affective model was learned using annotation prior and this posterior probability. Fig. 5.5 shows these affective GMMs obtained by using VQ for $K=64, 128, 256$.



(a)



(b)



(c)

Fig. 5.5 Affective GMM using VQ (a) $K=64$ (b) $K=128$ (c) $K=256$

On comparing Affective GMM clusters using AEG and VQ, as seen from Fig. 5.4 and Fig. 5.5, it is observed that the clusters obtained using AEG cover more of the VA space as compared to those obtained using VQ. Congestion and overlapping of Gaussians in a small area indicates a low inter-user agreement.

5.6 Cross Validation

Cross validation checks how well a model generalizes to new data. Cross Validation is used for sub-sampling the training data, avoiding the overfitting and to make predictions generalizable.

APPROACH:

1. Use the training set
2. Split it into training/test set
3. Build a model on the training set
4. Evaluate on the test set
5. Repeat and average the estimated errors

USED FOR:

1. Picking variables to include in a model
2. Picking the type of prediction function to use
3. Picking the parameters in the prediction function
4. Comparing different predictors

Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once. For classification problems, one typically uses stratified k-fold cross-validation, in which the folds are selected so that each fold contains roughly the same proportions of class labels. In repeated cross-validation, the cross-validation procedure is repeated n times, yielding n random partitions of the original sample. The n results are again averaged (or otherwise combined) to produce a single estimation [45].

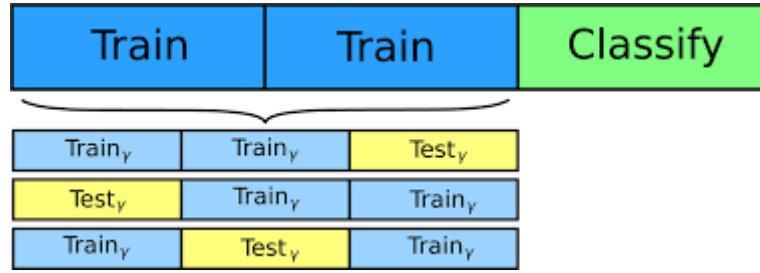


Fig. 5.6 Threefold cross validation [45]

In the threefold cross validation used, the entire dataset is divided into 3 blocks. In each fold, two blocks are used for training and one block is used for testing as shown in Fig. 5.6.

5.7 Performance metrics

The performance of AEG and VQ was compared by using the metrics Average Kullback-Leibler divergence (AKL) and Average Euclidean distance (AED) for different values of number of clusters ($K=32, 64, 128, 256$).

Table 5.1 indicates the values of AKL obtained for different number of clusters (K) with and without considering the annotation prior. These values are graphically represented in Fig. 5.7.

Table 5.1 AKL

	AEG				VQ			
	$K=32$	$K=64$	$K=128$	$K=256$	$K=32$	$K=64$	$K=128$	$K=256$
AKL values With Annotation Prior	0.8101	0.7813	0.7887	0.8065	0.8198	0.8282	0.8456	0.9111
AKL values Without Annotation Prior	0.8629	0.8403	0.8451	0.8629	0.8577	0.8474	0.8787	1.0640

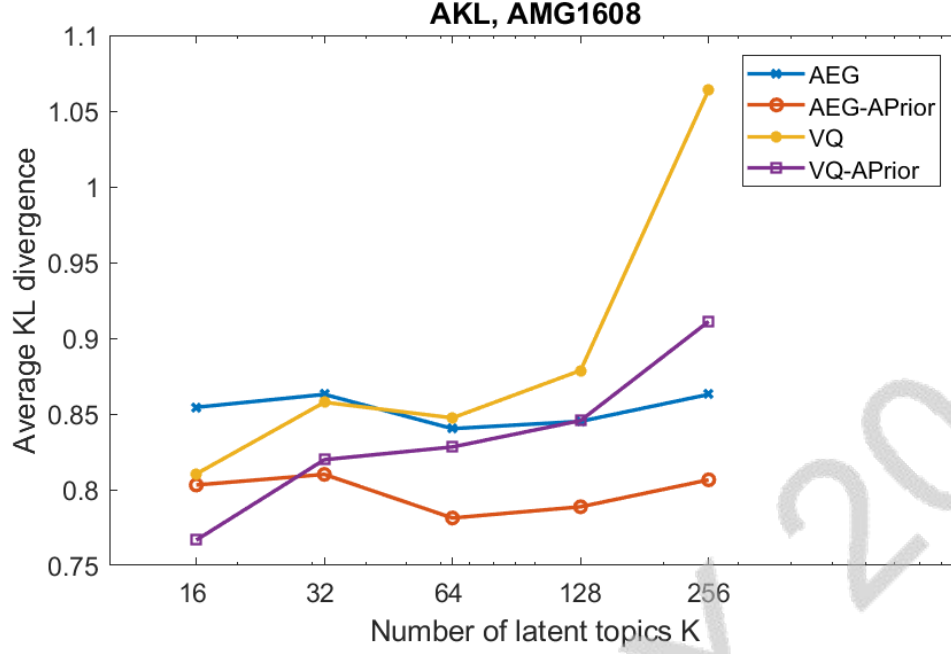


Fig. 5.7 AKL

Table 5.2 indicates the AED values obtained for different number of clusters (K) with and without considering the annotation prior. These values are graphically represented in Fig. 5.8.

Table 5.2 AED

Parameters	AEG				VQ			
	$K=32$	$K=64$	$K=128$	$K=256$	$K=32$	$K=64$	$K=128$	$K=256$
AED values With Annotation Prior	0.3272	0.3232	0.3234	0.3299	0.3176	0.3206	0.3284	0.3383
AED values Without Annotation Prior	0.3157	0.3124	0.3127	0.3179	0.3114	0.3159	0.3222	0.3438

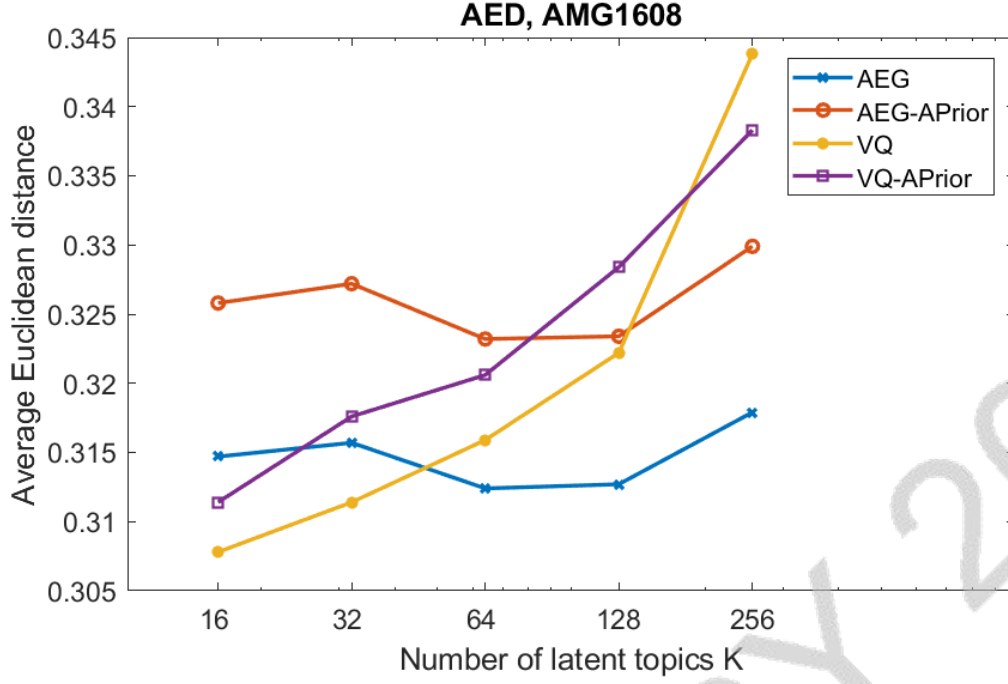


Fig. 5.8 AED

The following observations can be made from Tables 5.1 and 5.2, Fig. 5.7 and Fig. 5.8,

1. Considering the annotation prior, on an average, the AKL values for AEG are 6.846% times less than those of VQ. This implies that the difference between the Ground Truth Gaussian and the Gaussian obtained by using AEG is lesser as compared to the difference between the Ground Truth Gaussian and the Gaussian obtained by using VQ. Annotation prior accounts for the reliability of annotations provided by the annotators and thus reduce the AKL values.
2. The AED values for AEG are 0.07% times less than those of VQ. This indicates that, the distance between the mean of Ground Truth Gaussian and AEG Gaussian is lesser than the distance between the mean of Ground Truth Gaussian and VQ Gaussian.
3. Using the annotation prior consistently improves AKL but not AED. AKL is improved possibly because the annotation prior adds information regarding the annotation covariance of each training excerpt to the update of the model parameters. However, this may introduce bias to the original annotation mean of a training excerpt and slightly harm AED.

Thus, the performance of AEG is better as compared to that of VQ due to smaller values of AKL and AED. AKL is considered as a better performance indicator as it takes both mean and covariance into consideration. Also, the value of AED is sensitive to the numerical range of emotion space, whereas AKL is not.

5.8 Predicted Gaussians Using AEG and VQ

The ground-truth emotion annotations from listeners and their corresponding Gaussians and the predicted Gaussians by AEG and VQ for three randomly selected song clips of AMG1608 are presented in Fig. 5.9.

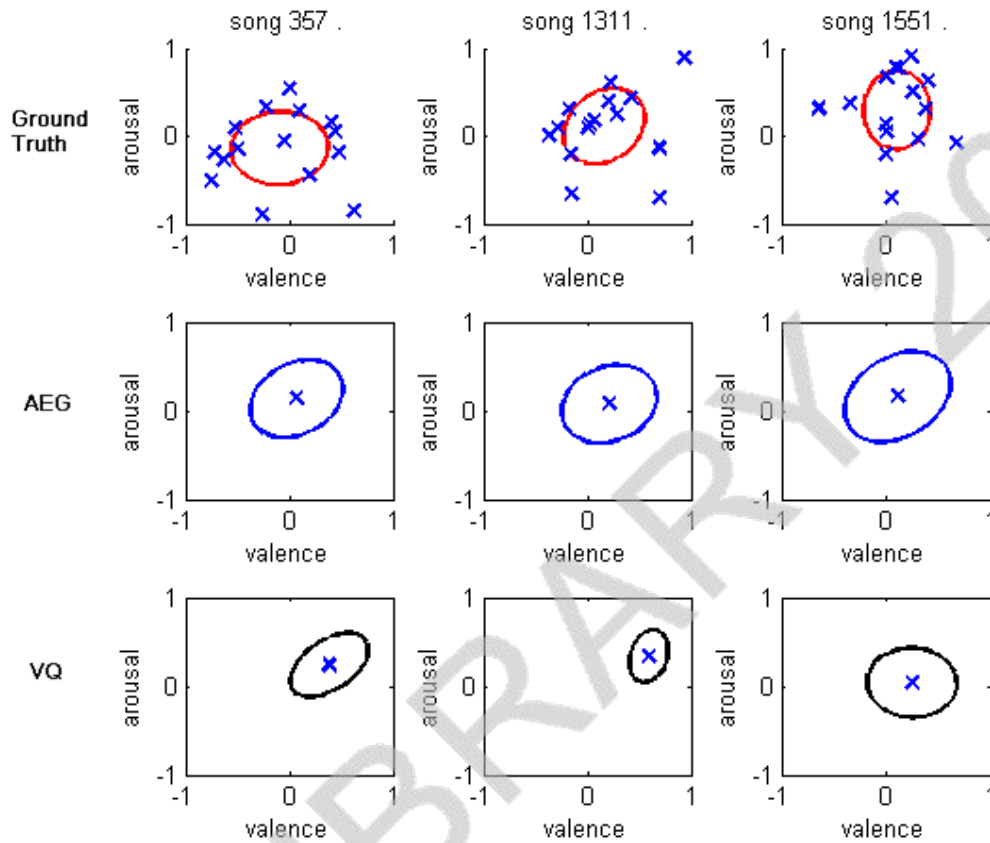


Fig. 5.9 Predicted Gaussians using AEG & VQ

It is observed that the mean of Gaussian obtained by using AEG has a relatively more accurate position as compared to the mean of Gaussian obtained by using VQ. The covariance of Gaussian obtained by using AEG is more similar to that of ground-truth as compared to the covariance of Gaussian obtained by using VQ. Thus, the Gaussian for emotion prediction using AEG is closer to Ground Truth than that using VQ. This validates the effectiveness of AEG.

Chapter 6

Conclusion

6.1 Conclusion

AEG is a principled probabilistic framework which utilizes the computation processes for MER and emotion-based music retrieval for dimensional emotion representations such as valence and arousal. The benefit of AEG is that it better takes into account the subjective nature of music emotional responses through the use of probabilistic inference and model adaptations thus making it possible to personalize an emotion-based MIR system. Acoustic GMM was used for implementing an improved multiclass emotion recognition system. AMG1608 database was used to test the robustness of the proposed method. GMM based emotion recognition method was used because it provides the advantage of diminishing the high within-class variance and escalating the between-class variance. GMM was employed as it enhances the entropy features and hence the performance of emotion recognition system could be enhanced particularly in the recognition of multiclass emotions. VQ gives a hard decision by representing a cluster by a single centroid. This limits the resolution of subjectivity issue. This is overcome using AEG as it considers the covariance along with the mean for each cluster thus providing a soft decision. Overall, the emotion-based music system helps in reducing the searching time for music thereby reducing the unnecessary computational efforts, thus increasing the overall accuracy and efficiency of the system. Apart from reducing the physical stress it will also assist the music therapies.

6.2 Future Scope

The current model designed, is for a general purpose, but can sometimes be prototypical. In order to go into more details, we would need to put the user in the loop. We can start to personalize these models and adapt them to user. Innovative personalization methods can be designed by applying the model to tasks such as context-aware recommendation, implicit tagging, and computer-generating emotional music. To develop a mood-enhancing music player, start with user's current emotion (which may be sad) and then play the music of positive emotions thereby eventually giving a joyful feeling to the user. Another future work would be to account for the effect of listeners' mood while collecting emotion annotations and while evaluating the performance of the system. Emotion recognition using facial expressions can be incorporated to enhance the user interface. By using the available source codes, AEG can be used in studying emotions in fields such as psychology or neurobiology.

Appendix

Timeline Chart of the Project

TIMELINE CHART FOR SEMESTER VII																	
MONTH	JULY				AUGUST					SEPTEMBER				OCTOBER			
WEEK NO.	W1	W2	W3	W4	W1	W2	W3	W4	W5	W1	W2	W3	W4	W1	W2	W3	W4
WORK TASKS																	
Study of abstract, introduction and related work from base paper.																	
Designing of code to obtain histogram.																	
Study of spectrogram concepts.																	
Study of K mean clustering and GMM.																	
Study of approaches of classifying emotions and techniques for modelling valence and arousal																	
Study of functions of MIR toolbox																	
Generation of 70D feature vector																	
Documentation																	

TIMELINE CHART FOR SEMESTER VIII																	
MONTH	JANUARY				FEBRUARY					MARCH				APRIL			
WEEK NO.	W1	W2	W3	W4	W1	W2	W3	W4	W5	W1	W2	W3	W4	W1	W2	W3	W4
WORK TASKS																	
Study of affective GMM																	
Implementation of affective GMM using AEG																	
Study and implementation of affective GMM using VQ																	
Testing using threefold cross validation																	
Black book documentation																	
Presentation and Documentation																	

References

- [1] https://en.wikipedia.org/wiki/Probability_distribution
- [2] https://en.wikipedia.org/wiki/Normal_distribution
- [3] <http://www.investopedia.com/terms/n/normaldistribution.asp>
- [4] Christopher M. Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics), Feb 2015
- [5] <http://www.itl.nist.gov/div898/handbook/eda/section3/eda3661.htm>
- [6] <http://scikit-learn.org/stable/modules/mixture.html>
- [7] <https://brilliant.org/wiki/gaussian-mixture-model>
- [8] <https://www.quora.com/What-are-the-advantages-to-using-a-Gaussian-Mixture-Model-clustering-algorithm>
- [9] <http://scikit-learn.org/stable/modules/mixture.html>
- [10] <http://www.statisticshowto.com/em-algorithm-expectation-maximization>
- [11] https://en.wikipedia.org/wiki/Support_vector_machine
- [12] Basak, Debasish, Srimanta Pal, and Dipak Chandra Patranabis. "Support vector regression." *Neural Information Processing-Letters and Reviews* 11, no. 10 (2007): 203-224.
- [13] <https://in.mathworks.com/matlabcentral/fileexchange/24583-mirtoolbox>
- [14] Olivier Lartillot. *MIRtoolbox 1.3.3*. Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Finland, July, 12th, 2011.
- [15] https://en.wikipedia.org/wiki/Emotion_classification
- [16] Wang, Ju-Chiang, Yi-Hsuan Yang, Kaichun Chang, Hsin-Min Wang, and Shyh-Kang Jeng. "Exploring the relationship between categorical and dimensional emotion semantics of music." In *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pp. 63-68. ACM, 2012.
- [17] <http://www.6seconds.org/2017/04/27/plutchiks-model-of-emotions>
- [18] sydney.edu.au/engineering/latte/docs/11-Kim-Master_Thesis_Final.pdf
- [19] article.sapub.org/10.5923.j.jgt.20130203.02.html
- [20] Millie Pant, Kusum Deep, Jagdish Chand Bansal, Atulya Nagar, Kedar Nath Das. Proceedings of Fifth International Conference on Soft Computing for Problem Solving, SocProS 2015, Volume 1

- [21] Harshali Nemade, Deipali Gore. (2017, june). "Music and Mood Detection using BoF Approach." International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization). Vol. 5, Issue 6. Available: www.ijircce.com
- [22] <https://www.slideshare.net/MaijuVuolle/emotion-measurement-services-for-knowledge-work>
- [23] <https://conversionxl.com/blog/valence-arousal-and-how-to-ignite-an-emotional-fire>
- [24] <http://www.thinkingzygote.com/2013/06/monitoring-emotions-valence-vs-arousal.html>
- [25] https://en.wikipedia.org/wiki/Emotional_granularity
- [26] Wang, Ju-Chiang, Yi-Hsuan Yang, Hsin-Min Wang, and Shyh-Kang Jeng. "Modeling the affective content of music with a Gaussian mixture model." *IEEE Transactions on Affective Computing* 6, no. 1 (2015): 56-68.
- [27] <http://emotiondevelopmentlab.weebly.com/circumplex-model-of-affect.html>
- [28] <http://power-map.com/what-are-heat-maps>
- [29] <http://www.heatmapping.org/index.html>
- [30] Y.-H. Yang and H. H. Chen, "Prediction of the distribution of perceived music emotions using discrete samples," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 7, pp. 2184-2196, Sep. 2011.
- [31] K. F. MacDorman, S. Ough, and C.-C. Ho, "Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison," *J. New Music Res.*, vol. 36, no. 4, pp. 281-299, 2007.
- [32] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 2, pp. 448-457, Feb. 2008.
- [33] Y.-H. Yang, Y.-C. Lin, and H. H. Chen, "Personalized music emotion recognition," in *Proc. ACM SIG Inf. Retrieval*, 2009, pp. 748-749.
- [34] M. Soleymani, M. N. Caro, E. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proc. Int. Workshop Crowdsourcing Multimedia*, 2013, pp. 1-6.
- [35] E. M. Schmidt and Y. E. Kim, "Modeling musical emotion dynamics with conditional random fields," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2011, pp. 777-782.
- [36] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions from audio," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2010, pp. 465-470.
- [37] Y. Vaizman, B. McFee, and G. Lanckriet, "Codebook-based audio feature representation

- for music information retrieval,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1483–1493, Oct. 2014.
- [38] Y.-A. Chen, J.-C. Wang, Y.-H. Yang, and H. H. Chen, “Linear regression-based adaptation of music emotion recognition models for personalization,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 2168–2172.
- [39] Chen, Yu-An, Yi-Hsuan Yang, Ju-Chiang Wang, and Homer Chen. "The AMG1608 dataset for music emotion recognition." In *Acoustics, Speech and Signal Processing (ICASSP)*, 2015 IEEE International Conference on, pp. 693-697. IEEE, 2015.
- [40] <http://mpac.ee.ntu.edu.tw/dataset/AMG1608/Description.htm>
- [41] <https://www.openml.org/a/estimation-procedures/1>
- [42] Kamale, Hemlata Eknath, and R. S. Kawitkar. "Vector quantization approach for speaker recognition." *International Journal of Computer Technology and Electronics Engineering*(2008):110-114.
- [43] https://en.wikipedia.org/wiki/Vector_quantization
- [44] http://www.oocities.org/stefangachter/VectorQuantization/chapter_1.htm
- [45] web.science.mq.edu.au/~cassidy/comp449/html/ch10s03.html
- [46] <http://web.engr.illinois.edu/~hanj/cs412/bk3/KL-divergence>
- [47] <http://rosalind.info/glossary/euclidean-distance/>
- [48] <https://math.stackexchange.com/questions/917066/calculating-weighted-euclidean-distance-with-given-weights>
- [49] https://en.wikipedia.org/wiki/Radial_basis_function_kernel
- [50] Chen, Sih-Huei, Yuan-Shan Lee, Wen-Chi Hsieh, and Jia-Ching Wang. "Music emotion recognition using deep Gaussian process." In *Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2015 Asia-Pacific, pp. 495-498. IEEE, 2015.
- [51] Fukuyama, Satoru, and Masataka Goto. "Music emotion recognition with adaptive aggregation of Gaussian process regressors." In *Acoustics, Speech and Signal Processing (ICASSP)*, 2016 IEEE International Conference on, pp. 71-75. IEEE, 2016.
- [52] An, Yunjing, Shutao Sun, and Shujuan Wang. "Naive Bayes classifiers for music emotion classification based on lyrics." In *Computer and Information Science (ICIS)*, 2017 IEEE/ACIS 16th International Conference on, pp. 635-638. IEEE, 2017.

Acknowledgement

The satisfaction that accompanies the completion of our project would be incomplete without the mention of the people who made it possible, without whose constant guidance and encouragement our efforts would go in vain. We consider ourselves privileged to express gratitude and respect towards all those who guided us throughout the completion of this project.

We convey our immense gratitude to our project guide Mr. Santosh Chapaneri, Assistant Professor of Electronics and Telecommunications Engineering Department for providing encouragement, constant support and guidance which was of a great help to work on this project successfully.

We would like to express sincere thanks to our Director, Bro. Jose Thuruthiyil, **Head of Electronics and Telecommunication Dept., Dr. Gautam Shah**, and the entire Electronics and Telecommunication Dept. for their valuable comments and constructive criticism during various stages of this project.

Last but not the least, we wish to thank our parents for constantly encouraging us to learn engineering. Their personal sacrifice in providing this opportunity to us to learn engineering is gratefully acknowledged.



Rutuja Girmal

J. G. Kamat
Jitali Kamat



Sayali Martal



Pooja Nair