The Acoustic Emotion Gaussians Model for Emotion-based Music Annotation and Retrieval

Ju-Chiang Wang[1,2], Yi-Hsuan Yang[2,3], Hsin-Min Wang[2,3] and Shyh-Kang Jeng[1]
[1] Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan
[2] Institute of Information Science, Academia Sinica, Taipei, Taiwan
[3] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
{asriver, yang, whm}@iis.sinica.edu.tw; skjeng@cc.ee.ntu.edu.tw

ABSTRACT

One of the most exciting but challenging endeavors in music research is to develop a computational model that comprehends the affective content of music signals and organizes a music collection according to emotion. In this paper, we propose a novel *acoustic emotion Gaussians* (AEG) model that defines a proper generative process of emotion perception in music. As a generative model, AEG permits easy and straightforward interpretations of the model learning processes. To bridge the acoustic feature space and music emotion space, a set of *latent feature classes*, which are learned from data, is introduced to perform the end-to-end semantic mappings between the two spaces. Based on the space of latent feature classes, the AEG model is applicable to both automatic music emotion annotation and emotion-based music retrieval. To gain insights into the AEG model, we also provide illustrations of the model learning process. A comprehensive performance study is conducted to demonstrate the superior accuracy of AEG over its predecessors, using two emotion annotated music corpora MER60 and MTurk. Our results show that the AEG model outperforms the state-of-the-art methods in automatic music emotion annotation. Moreover, for the first time a quantitative evaluation of emotion-based music retrieval is reported.

Categories and Subject Descriptors

H.5.5 [**Sound and Music Computing**]: Methodologies and Techniques, Systems

General Terms

Algorithms, Experimentation, Performance

Keywords

Computational emotion model, automatic music emotion recognition, music retrieval, Gaussian mixture model, EM algorithm.

MM'12, October 29–November 2, 2012, Nara, Japan.
Copyright 2012 ACM 978-1-4503-1089-5/12/10 ...$15.00.

1. INTRODUCTION

One of the most exciting but challenging endeavors in music research is to develop a computational model that comprehends the affective content of music signals and organizes a music collection according to emotion [12, 29]. Such a model is desirable as the pursuit of emotional experience is the primary motivation for music listening [11]. Behavioral studies have also identified emotion as one of the most important attributes used by people for music retrieval [25].

Automatic annotation of music emotion is challenging because the perception of emotion is in nature subjective.[1] Oftentimes people perceive differently when listening to the same song [5]. Consequently, one cannot assume that there is an objective, single ground truth label that applies equally well to every listener. Instead, one needs to learn directly from multiple labels provided by different annotators [17] and present as the final result a *soft* (probabilistic) emotion assignment instead of a *hard* (deterministic) one.

Modeling the time-varying dynamics of music emotion poses another challenge. A music piece can express different emotions as time unfolds, and it has been argued that music's most expressive qualities are related to its structural changes across time [6]. To capture the continuous changes of emotional expression, the *dimensional* representation of emotion is found superior to its *categorical* counterpart [5, 23]. In this representation, emotions are considered as numerical values (instead of discrete labels) over a number of emotion dimensions, such as *valence* (positive/negative affective states) and *activation* (or arousal; energy level) – the two most fundamental dimensions found by psychologists [18].[2] In this way, music emotion recognition becomes the prediction of the moment-to-moment valence and activation (VA) values of a music piece corresponding to a series of points in the emotion space [29].

While early approaches to emotion recognition neglected the above two properties and represented the emotion of a song by a single, song-level discrete emotion label [9, 15, 24], recent years have witnessed a growing number of attempts that modeled the emotion of a song or a segment as a probabilistic *distribution* in the emotion space [26, 30], as shown in Figure 1, to better account for the subjective nature of

[1] We define music emotion as the emotion human *perceives* as being expressed in a piece of music, rather than the emotion *feels* in response to the piece. This distinction is made as we may not feel sorrow when listening to a sad tune [5].

[2] For example, happiness is an emotion associated with a positive valence and a high activation, while sadness is an emotion associated with a negative valence and a low activation.

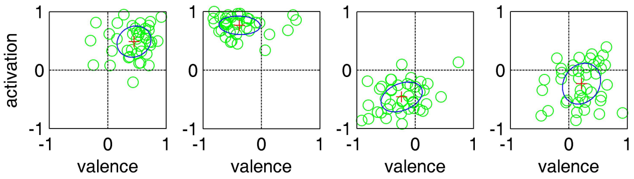


Figure 1: Subjects' annotations in the emotion space [18] for four 30-second clips.[4] Each circle corresponds to a subject's annotation for a clip, and the overall annotations for a clip can be modeled by a 2-D Gaussian distribution (i.e., the blue ellipse) [30].

emotion perception [30]. This can be approached by modeling the emotion distribution of each temporal segment as a bivariate Gaussian distribution and using regression algorithms to predict the mean, variance, and covariance of valence and activation directly from the acoustic features [30]. In this way, developers of an emotion-based music retrieval system can better understand how likely a specific emotional expression (defined in terms of VA values) would be perceived at each temporal moment. For a music piece whose perceived emotion is more listener-dependent, its emotion distribution in the emotion space would be sparser.

From the application point of view, this approach also creates a simple user interface for music retrieval through the specification of a point or a trajectory in the emotion space [29]. With this interface, users can easily retrieve music pieces of certain emotions without specifying the titles and can discover new pieces whose emotion is similar to that of a favorite piece. Users can also draw a trajectory in the display of a mobile device to indicate the desired emotion variation within a music piece or a playlist (e.g., changing from aggressive to tender).

From the theoretical point of view, the shift of music emotion modeling from a fixed label towards a time-varying, stochastic distribution is significant as a functional music retrieval system cannot be dissociated from the underlying psychological implications [2].

In this paper, we propose a novel *acoustic emotion Gaussians* (AEG) model that realizes a proper generative process of music emotion perception in a probabilistic and parametric framework. The AEG model learns from data two sets of Gaussian mixture models (GMMs), namely *acoustic GMM* and *VA GMM*, to describe the low-level acoustic feature space and high-level emotion space, respectively. A set of *latent feature classes* is introduced to play the end-to-end linkage between the two spaces and aligns the two GMMs. As a principled probabilistic model, AEG is applicable to both automatic music emotion annotation and emotion-based music retrieval.

The proposed AEG framework has three additional advantages. First, as a generative model, AEG permits easy and straightforward interpretation of the model learning and semantic mapping processes. Second, as AEG mainly involves light-weight computations of emotion prediction, it can track the moment-to-moment emotion distributions of a music piece in real-time. Third, due to its parametric and probabilistic nature, AEG provides great flexibility to future extension, such as personalizing the AEG model via model adaptation techniques or incorporating music tags for advanced music browsing and retrieval purposes.

A comprehensive performance study is conducted to demonstrate the superior accuracy of AEG over its predecessors, using two emotion annotated music corpora MER60 [30] and MTurk [26]. Moreover, for the first time a quantitative evaluation of emotion-based music retrieval is reported.

The remainder of this paper is organized as follows: First, we briefly review related work in Section 2. Next, we provide an overview of the AEG model in Section 3. The details of model learning, emotion annotation, and emotion-based retrieval are described in Sections 4 and 5. We then offer insights into the model learning process of AEG with empirical data in Section 6. Section 7 presents and discusses the evaluation results on both emotion-based music annotation and retrieval. Finally, we conclude the paper in Section 8.

2. RELATED WORK

A great amount of effort has been made by psychologists to study the relationship between music and emotion [7]. To identify the internal human representations of emotion, psychologists have applied factor analysis techniques such as multidimensional scaling to the emotion ratings of music stimuli. Although differ in names, existing studies give very similar interpretations of the resulting fundamental factors, most of which correspond to valence and activation [18].

Early approaches to automatic music emotion recognition [4, 16, 31] usually assumed that the perceived emotion of a music clip can be represented as a *single point* in the emotion space, in which VA values are considered as numerical values. The pair of ground truth VA values of a music clip is obtained by averaging the annotations of a number of subjects, without considering the variance of the annotations. Then, a regression algorithm, such as multiple linear regression (MLR) or support vector regression (SVR) [22], can be applied to train regression models for predicting the VA values. To exploit the temporal continuity of emotion variation within a clip, techniques such as Kalman filtering [20] or system identification [13] have also been used. Associated with the VA values, a music clip is visualized as a point in the VA space, and the similarity between clips are measured by the Euclidean distance in the emotion space.

However, because emotion perception is inherently subjective, simply representing a clip as a point according to the mean VA values is not enough to capture the nuance of human perception, as illustrated in Figure 1. Moreover, it has been argued that the VA space may not be an Euclidean space [27] and that it is better to measure the similarity of two songs according to the divergence between the corresponding two emotion distributions [30].

To better account for the subjective nature of emotion perception, algorithms for predicting the emotion distribution of a music clip from acoustic features have been proposed recently. Existing approaches can be divided into two categories. The *heatmap* approach [21, 30] quantizes each emotion dimension by G equally spaced discrete samples, leading to a $G \times G$ grid representation of the emotion space, and trains G^2 regression models for predicting the *emotion intensity* of each emotion point. Higher emotion intensity at an emotion sample indicates higher probability for a listener to perceive that emotion when listening to the music segment. The emotion intensity over the VA space cre-

[4] The four clips from left to right are *Dancing Queen* by ABBA, *Civil War* by Guns N' Roses, *Suzanne* by Leonard Cohen, and *All I Have To Do Is Dream* by the Everly Brothers, respectively.

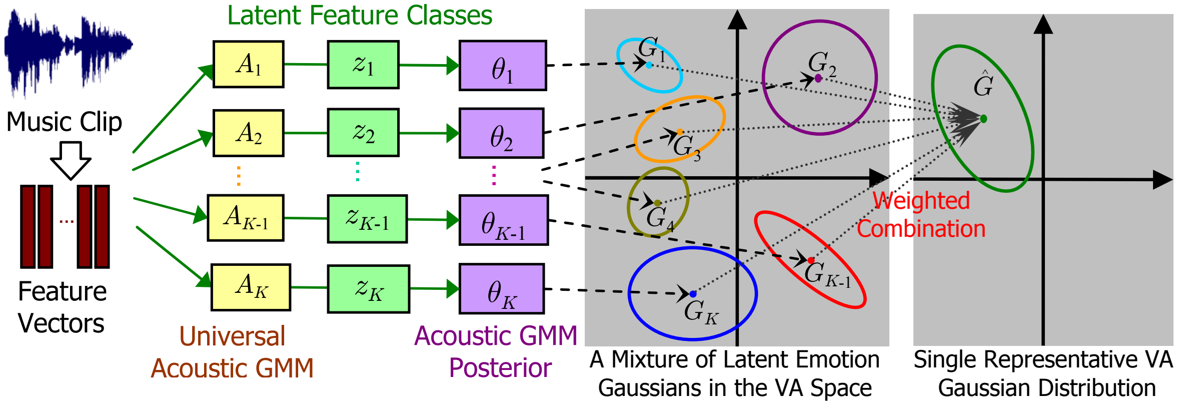


Figure 2: Illustration of AEG. Music emotion distribution can be generated from the acoustic features.

ates a heatmap like representation of emotion distribution. The drawback of this approach is that heatmap is a discrete rather than continuous representation of the VA space, and that emotion intensity cannot be regarded as a probability estimate in a strict sense.

The *Gaussian-parameter* approach [19,26,30] directly learns five regression models to predict the mean, variance, and covariance of valence and activation, respectively. This approach allows easier performance analysis; one can analyze the impact of separate acoustic features for the prediction of each Gaussian parameter. In addition, the modeling of mean VA values has been studied for years [4,16,31]. However, neither this approach nor the heatmap approach renders a strict probabilistic foundation. Moreover, the correlation among the mean and variance of valence and activation is not exploited, as the regression models are trained independently.

3. FRAMEWORK OVERVIEW

This section introduces AEG from a high-level point of view and outlines the system architecture. The details of each system component are described in Sections 4 and 5.

3.1 The Acoustic Emotion Gaussians Model

There are two basic assumptions underlying AEG. First, we assume that the emotion annotations of a music clip can be parameterized by a *bivariate Gaussian distribution.*[5] Second, we assume that music clips similar in acoustic features (e.g., loudness, timbre, rhythm, and harmony) also result in similar emotion distributions in the VA space.

As shown in Figure 2, the AEG model realizes the generation of music emotion in the emotion space. At the most abstract level, the relationship between the acoustic feature space and music emotion space can be observed from an annotated music corpus. However, such a relationship may be sometimes complicated and difficult to identify directly. Therefore, we introduce a set of *latent feature classes*, $\{z_k\}_{k=1}^K$, which functions as a linkage between the two spaces, into the generative process of music emotion.

Suppose that each z_k, which is defined by a latent acoustic classifier A_k, can map a specific pattern of acoustic features to a specific area G_k in the VA space. The set of latent acoustic classifiers, $\{A_k\}_{k=1}^K$, can be implemented by a universal acoustic Gaussian mixture model (GMM), denoted as the *acoustic GMM*, in which each component Gaussian A_k represents a specific acoustic pattern discovered by the GMM learning in the frame-based acoustic feature space. In the meanwhile, G_k can be modeled by a bivariate Gaussian distribution, denoted as a latent VA Gaussian. The mixture of latent VA Gaussians is called the *VA GMM* hereafter.

3.2 Generation of Music Emotion

As shown in the left hand side of Figure 2, the acoustic features of a music clip can be represented by computing the posterior probabilities over the acoustic GMM (i.e., each component Gaussian A_k leads to a posterior probability θ_k) based on its frame-based feature vectors. We denote this clip-level acoustic feature representation as the *acoustic GMM posterior*, $\{\theta_k\}_{k=1}^K$, subject to $\sum_k \theta_k = 1$. Therefore, the acoustic GMM posterior is able to capture the acoustic characteristics of every music clip in a K-dimensional probabilistic space. In the right hand side of Figure 2, the emotion distribution of a music clip in the VA space can be generated by the weighted combination of all latent VA Gaussians as $\sum_k \theta_k G_k$ using $\{\theta_k\}_{k=1}^K$ as weights. Consequently, the generative process of emotion of a music clip can be interpreted as that, for example, if human has perceived only the acoustic pattern of z_1 from a music clip's acoustic features (this will lead to $\theta_1 = 1$, and $\theta_k = 0$, $\forall k \neq 1$), then his/her emotion perception would exactly follow G_1.

Since it is more intuitive to interpret the final emotion prediction of a music clip to the users with only a set of mean (center) and covariance (uncertainty) rather than a complicated VA GMM, the $\boldsymbol{\theta}$-weighted VA GMM is finally combined into a single 2-D VA Gaussian $\hat{G}$ in the emotion prediction phase, as shown in the rightmost of Figure 2.

3.3 Emotion-based Music Retrieval

As shown in Figure 3, the proposed emotion-based music retrieval system can be divided into two phases: the *feature indexing phase* and the *music retrieval phase*. In the indexing phase, each music clip in the un-labeled music database is indexed with two indexing approaches based on the music clip's acoustic features: indexing with the *acoustic GMM posterior* (a fixed-dimensional vector) of a clip using the acoustic GMM, or indexing with the *predicted emotion distribution* (a single 2-D Gaussian) of a clip given by automatic

[5] The assumptions that the VA space can be considered as Euclidean and that the VA values can be modeled as a Gaussian distribution are made in prior work (cf. Section 2). We will not discuss this issue further but instead empirically validate the effectiveness of emotion-based music information systems built upon these assumptions for practical uses.

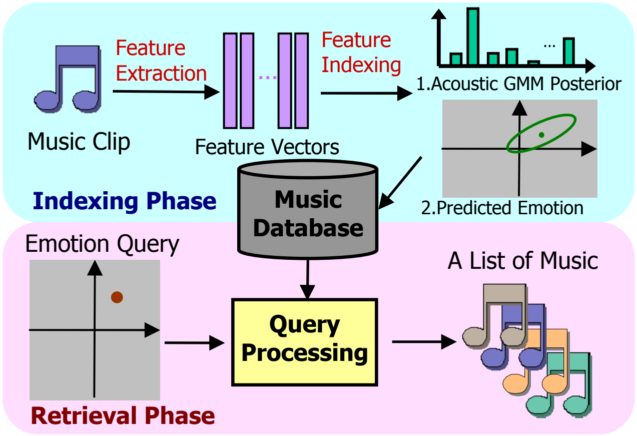


Figure 3: The flowchart of the content-based music search system using a VA-based emotion query.

Table 1: The two implementations of the emotion-based music search system.

Approach	Indexing	Retrieval
Fold-In	Acoustic GMM Posterior (K-dim Vector $\{\theta_k\}_{k=1}^K$)	Pseudo Song $\{\lambda_k\}_{k=1}^K$
Emotion Prediction	Predicted Emotion Distribution (2-D Gaussian $\hat{G}$)	Distribution Likelihood

emotion prediction. In the retrieval phase, given a point query from the emotion (VA) space, the system will return a ranked list of relevant music clips. We apply two matching methods, namely *pseudo song-based matching* and *distribution likelihood-based matching*; each corresponding to one of the two indexing approaches. In pseudo song-based matching, the point query is first transformed into a pseudo song (i.e., the estimated acoustic GMM posterior for the point query), and then matched with clips in an unlabeled music database. In distribution likelihood-based matching, a point query is fed into the predicted emotion distribution of each clip in an unlabeled music database, and the system ranks all the clips according to the estimated likelihoods. Table 1 summarizes the two implementations of the emotion-based music retrieval system.

3.3.1 Pseudo Song Estimation

In the Fold-In method, a given query point is transformed into probabilities $\{\lambda_k\}_{k=1}^K$, s.t. $\sum_k \lambda_k = 1$, as shown in Figure 4. The resulted λ_k represents the importance of the k-th latent VA Gaussian for the input query point. Suppose that the estimated $\{\lambda_k\}_{k=1}^K$ have a similar property with the acoustic GMM posterior $\{\theta_k\}_{k=1}^K$ that are used to index music clips in the music database. For instance, in the case shown in Figure 4, the 2-nd latent VA Gaussian is extremely likely to generate the input query point, the Fold-In process will assign a dominative weight λ_2 to z_2, e.g., $\lambda_2 = 1$, and $\lambda_k = 0$, $\forall k \neq 2$, which means that the query point is extremely relevant to the song whose acoustic GMM posterior is dominated by θ_2. Therefore, the pseudo song can be used to match with the clips in the un-labeled music database.

4. LEARNING THE AEG MODEL

This section will first present the mathematical generative process of the AEG model, i.e., introducing the learning from the acoustic GMM to the VA GMM, and then coming up with a summarized learning algorithm of the VA GMM.

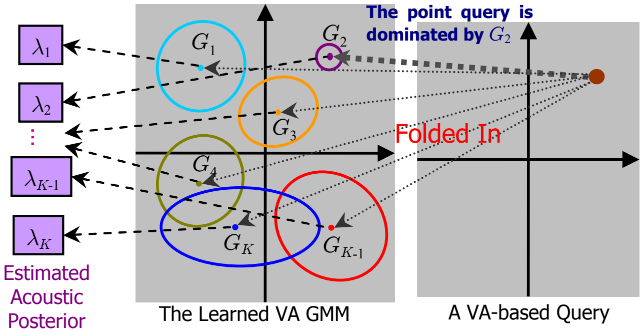


Figure 4: The Fold-In process. A point query is projected into the space of acoustic GMM posterior.

4.1 Acoustic GMM Posterior Representation

To start the generative process of the AEG model, we utilize a universal acoustic GMM, which is pre-learned using the expectation-maximization (EM) algorithm on a universal set of frame vectors $\mathcal{F}$, to span a probabilistic space with a set of diverse acoustic Gaussians. The learned acoustic GMM is expressed as follows,

$$p(\mathbf{x}) = \sum\nolimits_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|z_k, \mathbf{m}_k, \mathbf{S}_k), \tag{1}$$

where $\mathbf{x}$ is a frame-based feature vector; π_k, $\mathbf{m}_k$, and $\mathbf{S}_k$ are the prior weight, mean vector, and covariance matrix of the k-th component Gaussian A_k, respectively.

Suppose that we have an annotated music corpus $\mathcal{X}$ with N clips, each is denoted as s_i, and its t-th frame vector is denoted as $\{\mathbf{x}_{it}\}_{t=1}^{T_i}$, where T_i is the number of frames. The acoustic posterior probability of z_k for $\mathbf{x}_{it}$ is computed by,

$$p(z_k|\mathbf{x}_{it}) = \frac{\mathcal{N}(\mathbf{x}_{it}|z_k, \mathbf{m}_k, \mathbf{S}_k)}{\sum_{h=1}^{K} \mathcal{N}(\mathbf{x}_{it}|z_h, \mathbf{m}_h, \mathbf{S}_h)}. \tag{2}$$

In our implementation, the mixture prior (i.e., π_k and π_h) in Eq. 2 is replaced by $\frac{1}{K}$, because it is not useful in the previous work [28].

The clip-level acoustic GMM posterior probability θ_{ik} can be summarized by averaging the frame-level ones,

$$\theta_{ik} \leftarrow p(z_k|s_i) = \frac{1}{T_i} \sum\nolimits_{t=1}^{T_i} p(z_k|\mathbf{x}_{it}). \tag{3}$$

Finally, the acoustic GMM posterior of s_i is represented by vector $\boldsymbol{\theta}_i$, whose k-th component is θ_{ik}.

4.2 Prior Model for Emotion Annotation

To cover the emotion perception of different subjects, typically a clip in $\mathcal{X}$ is annotated by multiple subjects. However, as some annotations may be unreliable, we introduce a user prior model to express the contribution of each individual subject. Given the emotion annotations $\{\mathbf{e}_{ij}\}_{j=1}^{U_i}$ of s_i, where $\mathbf{e}_{ij}$ denotes the annotation by the j-th subject u_{ij} and U_i denotes the number of subjects who have annotated s_i, we build a label confidence model γ with the following Gaussian distribution,

$$\gamma(\mathbf{e}_{ij}|u_{ij}, s_i) \equiv \mathcal{N}(\mathbf{e}_{ij}|s_i,\mathbf{a}_i, \mathbf{B}_i), \tag{4}$$

where $\mathbf{a}_i = \frac{1}{U_i}\sum_j \mathbf{e}_{ij}$ and $\mathbf{B}_i = \frac{1}{U_i}\sum_j (\mathbf{e}_{ij} - \mathbf{a}_i)(\mathbf{e}_{ij} - \mathbf{a}_i)^T$. The confidence of $\mathbf{e}_{ij}$ can be estimated based on the likelihood calculated by Eq. 4. Therefore, if an annotation is far away from other annotations for the same song, it would be considered less reliable. Note that any criterion that is able to reflect the importance of a user's subjective annotation of a clip can be applied to model γ. In our preliminary study, we found that a single Gaussian empirically performs better than a GMM in describing γ.

Then, the posterior probability of user u_{ij} can be calculated by normalizing the confidence likelihood of $\mathbf{e}_{ij}$ over the cumulative confidence likelihood of all users for s_i,

$$p(u_{ij}|s_i) \equiv \frac{\gamma(\mathbf{e}_{ij}|u_{ij}, s_i)}{\sum_{r=1}^{U_i} \gamma(\mathbf{e}_{ir}|u_{ir}, s_i)}. \qquad (5)$$

$p(u_{ij}|s_i)$ is referred to as the *clip-level user prior*, as it indicates the confidence of each annotation for the clip. Based on the clip-level user prior, we further define the *corpus-level clip prior* to describe the importance of each clip as follows,

$$p(s_i|\mathcal{X}) \equiv \frac{\sum_{j=1}^{U_i} \gamma(\mathbf{e}_{ij}|u_{ij}, s_i)}{\sum_{q=1}^{N} \sum_{r=1}^{U_q} \gamma(\mathbf{e}_{qr}|u_{qr}, s_q)}. \qquad (6)$$

If the annotations of a clip are more consistent (i.e., the covariance values in $\mathbf{B}_i$ are small), the song is considered less subjective. If a song is annotated by more subjects, the corresponding γ model should be more reliable. The above two cases both lead to larger cumulative annotation confidence likelihoods.

With the two priors $p(u_{ij}|s_i)$ and $p(s_i|\mathcal{X})$, we define the *annotation prior* γ_{ij} by multiplying Eqs. 5 and 6,

$$\gamma_{ij} \leftarrow p(u_{ij}, s_i|\mathcal{X}) = \frac{\gamma(\mathbf{e}_{ij}|u_{ij}, s_i)}{\sum_{q=1}^{N} \sum_{r=1}^{U_q} \gamma(\mathbf{e}_{qr}|u_{qr}, s_q)}. \qquad (7)$$

The probabilities derived from Eqs. 5–7 are computed beforehand and then fixed in the learning process of AEG.

4.3 Learning the VA GMM

As described in the model overview, we assume that $\mathbf{e}_{ij}$ of s_i for u_{ij} in $\mathcal{X}$ can be generated from a weighted VA GMM governed by the acoustic GMM posterior $\boldsymbol{\theta}_i$ of s_i,

$$p(\mathbf{e}_{ij}|u_{ij}, s_i, \boldsymbol{\theta}_i) = \sum_{k=1}^{K} \theta_{ik} \mathcal{N}(\mathbf{e}_{ij}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad (8)$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ denote the mean vector and covariance matrix of the k-th latent VA Gaussian G_k shown in Figure 2 to be learned by means of the following algorithm.

First, for each clip s_i, we derive its acoustic prior $\boldsymbol{\theta}_i$, which is the acoustic GMM posterior vector, and the annotation prior γ_{ij} for each annotator u_{ij} of s_i. These two priors are computed beforehand and stay fixed in the learning process. Using $\mathbf{E}$ to denote the whole set of annotations in $\mathcal{X}$, the total likelihood can be derived by,

$$\begin{aligned} p(\mathbf{E}|\mathcal{X}) &= \textstyle\sum_i p(\mathbf{E}_i, s_i|\mathcal{X}) = \sum_i p(s_i|\mathcal{X}) p(\mathbf{E}_i|s_i, \mathcal{X}) \\ &= \textstyle\sum_i p(s_i|\mathcal{X}) \sum_j p(\mathbf{e}_{ij}, u_{ij}|s_i, \mathcal{X}) \\ &= \textstyle\sum_i p(s_i|\mathcal{X}) \sum_j p(u_{ij}|s_i, \mathcal{X}) p(\mathbf{e}_{ij}|u_{ij}, s_i, \mathcal{X}) \\ &= \textstyle\sum_i \sum_j p(u_{ij}, s_i|\mathcal{X}) \sum_k \theta_{ik} \mathcal{N}(\mathbf{e}_{ij}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \end{aligned} \qquad (9)$$

Taking the logarithm of Eq. 9 and replacing $p(u_{ij}, s_i|\mathcal{X})$ by γ_{ij} leads to,

$$L = \log \sum_i \sum_j \gamma_{ij} \sum_k \theta_{ik} \mathcal{N}(\mathbf{e}_{ij}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad (10)$$

where $\sum_i \sum_j \gamma_{ij} = 1$. To learn the VA GMM, we can maximize the log-likelihood in Eq. 10 with respect to the parameters of VA GMM. However, direct maximization of L is intractable as two-layer summation over the latent variables z_k appears inside the logarithm [1]. We therefore first derive a lower bound of L according to Jensen's inequality,

$$L \geq L_{\text{bound}} = \sum_{i,j} \gamma_{ij} \log \sum_k \theta_{ik} \mathcal{N}(\mathbf{e}_{ij}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \qquad (11)$$

Then, we treat L_{bound} as a surrogate of L. Although maximizing L_{bound} is still intractable, the logarithm now acts only on one-layer summation over the latent variables z_k. Therefore, we can maximize it with the EM algorithm [1].

In the E-step, we derive the expectation over the posterior probability of z_k given a user's annotation $\mathbf{e}_{ij}$ for s_i,

$$Q = \sum_{i,j} \gamma_{ij} \sum_k p(z_k|\mathbf{e}_{ij}, s_i) \Big(\log \theta_{ik} + \log \mathcal{N}(\mathbf{e}_{ij}|\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k) \Big), \qquad (12)$$

where

$$p(z_k|\mathbf{e}_{ij}, s_i) = \frac{\theta_{ik} \mathcal{N}(\mathbf{e}_{ij}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{h=1}^{K} \theta_{ih} \mathcal{N}(\mathbf{e}_{ij}|\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}. \qquad (13)$$

In the M-step, we maximize Eq. 12 and derive the following update rules:

$$\hat{\boldsymbol{\mu}}'_k \leftarrow \frac{\sum_i \sum_j \gamma_{ij} p(z_k|\mathbf{e}_{ij}, s_i) \mathbf{e}_{ij}}{\sum_i \sum_j \gamma_{ij} p(z_k|\mathbf{e}_{ij}, s_i)}, \qquad (14)$$

$$\boldsymbol{\Sigma}'_k \leftarrow \frac{\sum_i \sum_j \gamma_{ij} p(z_k|\mathbf{e}_{ij}, s_i) (\mathbf{e}_{ij} - \boldsymbol{\mu}'_k)(\mathbf{e}_{ij} - \boldsymbol{\mu}'_k)^T}{\sum_i \sum_j \gamma_{ij} p(z_k|\mathbf{e}_{ij}, s_i)}. \qquad (15)$$

At the last iteration of the EM learning process, we keep the prior model of the VA GMM,

$$\rho_k \leftarrow p(z_k) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{U_i} \gamma_{ij} p(z_k|\mathbf{e}_{ij}, s_i), \qquad (16)$$

for the Fold-In method in the later retrieval work. The prior ρ_k of the VA GMM involves all the prior information of$\mathcal{X}$ over the set of latent feature classes. The learning process of the VA GMM is summarized in Algorithm 1.

Algorithm 1: The learning process of the VA GMM

```
Input: Initial model $\{\boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}\}_{k=1}^{K}$; acoustic prior
       $\{\boldsymbol{\theta}_i\}_{i=1}^{N}$; annotation prior $\{\gamma_{ij}\}_{i=1,j=1}^{N,U_i}$;
Output: Model parameters $\{\rho_k, \boldsymbol{\mu}'_k, \boldsymbol{\Sigma}'_k\}_{k=1}^{K}$;
Iteration index $t \leftarrow 0$;
while $t \leq ITER$ do
    Compute the posterior probability using Eq. 13
    with $\boldsymbol{\mu}_k^{(t)}$ and $\boldsymbol{\Sigma}_k^{(t)}$;
    $t \leftarrow t + 1$;
    Update $\boldsymbol{\mu}_k^{(t)}$ and $\boldsymbol{\Sigma}_k^{(t)}$ using Eqs. 14 and 15;
    if $\boldsymbol{\Sigma}_k^{(t)}$ is non-positive definite then break;
end
Compute the prior model $\rho_k$ using Eq. 16;
Let $\boldsymbol{\mu}'_k \leftarrow \boldsymbol{\mu}_k^{(t)}$ and $\boldsymbol{\Sigma}'_k \leftarrow \boldsymbol{\Sigma}_k^{(t)}$;
```

4.3.1 Discussion

As shown in Eqs. 14 and 15, the parameters $\boldsymbol{\mu}'_k$ and $\boldsymbol{\Sigma}'_k$ of each Gaussian component in the learned VA GMM are softly contributed by all annotations $\mathbf{e}_{ij}$, and the responsibility is governed by the product of γ_{ij} and $p(z_k|\mathbf{e}_{ij}, s_i)$. As a result, the learning process seamlessly takes the annotation prior, acoustic prior, and likelihood over the current VA

GMM into consideration. As the VA GMM is getting fitted to the data, it may appear that two far-separate groups of annotations jointly have fairly large responsibilities on the covariance of a component VA Gaussian. This would make the covariance matrix non-positive definite (non-PD), which means the covariance shape of the VA Gaussian will become a straight line. The learning process will stop if the non-PD case occurs, as shown in line 6 in Algorithm 1.

5. EMOTION-BASED ANNOTATION AND RETRIEVAL USING AEG

In this section, we first explain how to use the AEG model for distribution-based emotion prediction and analyze the complexity of prediction. Then, we introduce how the model effects content-based music retrieval from the Emotion Prediction and the Fold-In perspectives, respectively.

5.1 Automatic Emotion Prediction

As described in Section 3.2, the VA GMM can generate the emotion annotation of an unseen music clip $\hat{s}$ given the clip's acoustic GMM posterior $\hat{\boldsymbol{\theta}}$ as follows,

$$p(\mathbf{e}|\hat{\boldsymbol{\theta}}) = \sum\nolimits_{k=1}^{K} \hat{\theta}_k \mathcal{N}(\mathbf{e}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{17}$$

where $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ are the learned VA GMM. The predicted emotion distribution of an unseen clip shown in Eq. 17 may be unnecessarily complicated, and presenting the result as a mixture of Gaussians also makes it difficult for a user to interpret the result of emotion prediction. Instead, a single and representative VA Gaussian $\mathcal{N}(\mathbf{e}|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ is practically more useful as explained in Section 3.2. This can be resorted to the information theory to calculate the mean and covariance of the representative VA Gaussian by solving the following optimization problem,

$$\mathcal{N}(\mathbf{e}|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \underset{\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}}{\operatorname{argmin}} \sum\nolimits_{k} \hat{\theta}_k D_{\mathrm{KL}}(\mathcal{N}(\mathbf{e}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \| \mathcal{N}(\mathbf{e}|\boldsymbol{\mu}, \boldsymbol{\Sigma})), \tag{18}$$

where $D_{\mathrm{KL}}(\mathcal{N}_A \| \mathcal{N}_B)$ denotes the one-way KL divergence from $\mathcal{N}(\mathbf{e}|\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A)$ to $\mathcal{N}(\mathbf{e}|\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$:

$$\begin{aligned} D_{\mathrm{KL}}(\mathcal{N}_A \| \mathcal{N}_B) =& \frac{1}{2}\left(\mathrm{tr}(\boldsymbol{\Sigma}_A \boldsymbol{\Sigma}_B^{-1}) - \log |\boldsymbol{\Sigma}_A \boldsymbol{\Sigma}_B^{-1}|\right) \\ &+ \frac{1}{2}(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)^T \boldsymbol{\Sigma}_B^{-1} (\boldsymbol{\mu}_A - \boldsymbol{\mu}_B) - 1, \end{aligned} \tag{19}$$

The optimal mean vector and covariance matrix for Eq. 18 are obtained by [3]:

$$\hat{\boldsymbol{\mu}} = \sum\nolimits_{k=1}^{K} \hat{\theta}_k \boldsymbol{\mu}_k, \tag{20}$$

$$\hat{\boldsymbol{\Sigma}} = \sum\nolimits_{k=1}^{K} \hat{\theta}_k \left(\boldsymbol{\Sigma}_k + (\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}})^T\right). \tag{21}$$

The AEG-based emotion prediction is very efficient. The complexity mainly depends on the size of latent feature classes K. One only needs to compute the acoustic GMM posterior representation of a music clip and use it to generate a bivariate Gaussian distribution by using Eqs. 20 and 21. The method is efficient enough to be applied to real-time music emotion tracking on a sequence of short music segments on a mobile device. The application can also be incorporated into a music player to create a real-time visualization of music content to enrich the experience of music listening.

5.2 Emotion-based Music Retrieval

As described in Section 3.3, two methods with the AEG model can be used for music retrieval from an un-labeled database given a point query in the VA space. In the Emotion Prediction method, each music clip is indexed as a single bivariate Gaussian distribution estimated by automatic emotion prediction. Given a point query, the music clips in the database are ranked according to the likelihoods of their automatically predicted Gaussian PDFs.

As for the Fold-In method, given a query point $\hat{\mathbf{e}}$, the system has to fold in the point into the AEG model and estimate a pseudo song $\hat{\boldsymbol{\lambda}}$ in an online fashion. This is achieved by maximizing the $\boldsymbol{\lambda}$-weighted VA GMM as follows:

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}} \log \sum\nolimits_{k} \lambda_k \mathcal{N}(\hat{\mathbf{e}}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{22}$$

That is to say, $\hat{\boldsymbol{\lambda}}$ corresponds to the most likely weighted combination of VA GMM for generating $\hat{\mathbf{e}}$. Eq. 22 can also be solved by the EM algorithm. In the E-step, the posterior probability of z_k given the query is computed by

$$p(z_k|\hat{\mathbf{e}}) = \frac{\lambda_k \mathcal{N}(\hat{\mathbf{e}}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_h \lambda_h \mathcal{N}(\hat{\mathbf{e}}|\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}. \tag{23}$$

Then, in the M-step, we only update λ_k by $\hat{\lambda}_k \leftarrow p(z_k|\hat{\mathbf{e}})$. Note that in estimating the pseudo song, the prior model of the VA GMM $\{\rho_k\}_{k=1}^{K}$ can be used for initialization. In our preliminary study, initializing with $\{\rho_k\}_{k=1}^{K}$ exhibits the best performance among other settings, such as uniform or random initialization.

Since each music clip in the un-labeled database has been indexed by an acoustic GMM posterior vector and the estimated pseudo song $\hat{\boldsymbol{\lambda}}$ of a query lies in the same vector space, the retrieval system ranks all the clips in descending order of cosine similarity in response to the query.

5.2.1 Remark

The Emotion Prediction method for retrieval is intuitive since we developed the system from the emotion prediction perspective. The Fold-In method goes one step further and leverages the learned relationship between audio patterns and emotions to represent a music clip and an emotion query in the same latent class space. Although the Fold-In method needs an additional step for pseudo song estimation, it is actually more flexible and efficient for incorporating external information. For example, when user feedback is available, online personalization is possible by performing adaptation on the learned VA GMM without retraining the model and re-annotating the songs in the database.

6. INTERPRETATION OF AEG

This section describes the annotated music corpora and the frame-based acoustic features utilized in this work. An in-depth analysis of the model learning process of AEG is conducted using the annotated corpora to offer insights to the VA GMM.

6.1 Music Corpora

Two corpora are employed in this work.[6] The first corpus MER60 consists of 60 clips (each is 30-second long) collected

[6] The datasets are available at `http://mac.iis.sinica.edu.tw/~yang/MER/NTUMIR-60/` and `http://music.ece.drexel.edu/research/emotion/moodswingsturk`.

from the chorus parts of English pop songs [30]. A total of 99 subjects were recruited for emotion annotation, making each clip annotated by 40 subjects. The subjects were asked to rate the VA values that best describe their general (instead of moment-to-moment) emotion perception of each clip in a silent computer lab. The VA values are entered by clicking on the emotion space on a computer display.

The second corpus, MTurk, is composed of 240 clips (each is 15-second long) drawn from the well-known uspop2002 database [26]. The emotion annotation is collected via Amazon's Mechanical Turk,[7] an online crowdscourcing engine. Each subject was asked to rate the per-second VA values for 11 randomly-selected clips using a graphical interface. After an automatic verification step that removes unreliable annotations, each clip is annotated by 7 to 23 subjects.

6.2 Frame-based Acoustic Features

In this work, we adopt the bag-of-frames modeling and extract frame-based musical features for acoustic modeling [28]. A frame that captures detailed temporal features can facilitate the ability of clip-level acoustic modeling of the acoustic GMM posterior representation. Instead of analyzing the emotion of a specific frame, we aggregate all the frames in a clip into the acoustic GMM posterior vector $\boldsymbol{\theta}$ (cf. Eq. 3) and perform our analysis of emotion at the clip level. Although it may be interesting to extract long-term mid-level features such as melody, rhythm, chord, or structural segments that directly characterizes the musical information, such features are not used because the extraction of them is still not perfect and they may introduce noises to the system. We leave this issue in our future work.

We use MIRToolbox 1.3 [14] to extract the following four types of frame-based acoustic features: *dynamic* (root-mean-squared energy), *spectral* (centroid, spread, skewness, kurtosis, entropy, flatness, rolloff 85%, rolloff 95%, brightness, roughness, and irregularity), *timbre* (zero crossing rate, spectral flux, 13 MFCCs, 13 delta MFCCs, and 13 delta-delta MFCCs), and *tonal* (key clarity, key mode possibility, HCDF, 12-bin chroma, chroma peak, and chroma centroid). All of the frame-based features are extracted with the same frame size of 50ms and 50% hop size to ensure easy alignment. Each dimension in all extracted frame vectors is normalized to have zero mean and one standard deviation. Two frame vector representations are considered in the performance evaluation: a 39-D vector that consists of MFCC-related features only and a 70-D vector that concatenates all the features.

For the MTurk corpus, since the audio waveforms are not available, we use the following four types of features that are kindly provided by the authors of [19, 26].

- **MFCCs** (20-D): Low-dimensional representation of the spectrum warped according to Mel-scale.
- **Chroma** (12-D): Autocorrelation of chroma is used, providing an indication of modality.
- **Spectrum descriptors** (4-D): Spectral centroid, flux, rolloff and flatness. Often related to timbral texture.
- **Spectral contrast** (14-D): Rough representation of the harmonic content in the frequency domain.

In the experiments, the individual feature sets are either used separately or concatenated into a 50-D vector.

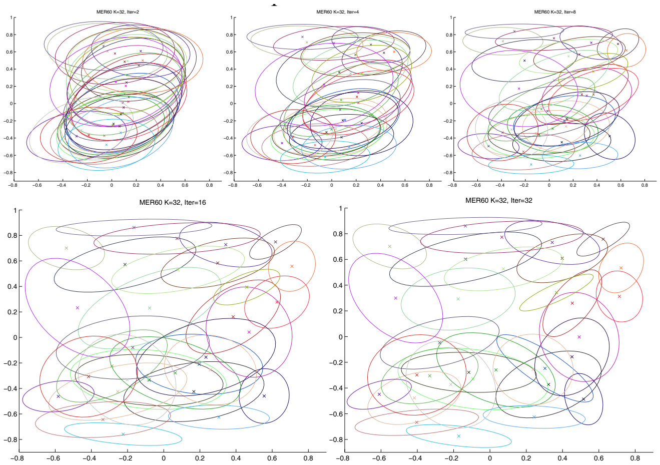


Figure 5: The learned VA GMMs (K=32) at different iterations (Iter=2, 4, 8, 16, and 32). Each component Gaussian is represented by an ellipse with a unique color. The horizontal and vertical axes correspond to valence and activation, respectively.

6.3 Interpretation of the Learned VA GMM on the MER60 Corpus

To observe the learning of AEG and interpret the resulting VA GMM, we fit the AEG model on MER60 and examine the learned VA GMM at each iteration of Algorithm 1. We first use an external music collection to form the global frame vector set $\mathcal{F}$ containing 235K 70-D frame vectors. Then, the acoustic GMMs are learned with several K values using the EM algorithm. We initialize all the Gaussian components of the VA GMM with the sample mean vector and covariance matrix of the emotion annotations of the whole training set.

Figure 5 illustrates the VA GMM learned at a number of iterations, from which one can observe that, while being tied together in the beginning, the VA Gaussians gradually separate one from the others as the learning process progresses. The ellipse of each VA Gaussian gets smaller and smaller after each iteration until convergence. As shown in the last sub-figure in Figure 5, in the end, different Gaussians cover different areas in the emotion space, and the learned VA GMM appears to be able to approximate all kinds of emotion distributions by using different sets of acoustic GMM posterior probabilities. Since our preliminary study indicates that over-small VA Gaussians may not help for emotion prediction, early stop is required.

The specific area covered by a VA Gaussian actually gives a semantic meaning to the corresponding latent feature class defined by the acoustic GMM. That is, the mapping from the acoustic space to the emotion space can be easily observed and interpreted. We also notice that there are many Gaussians with horizontally elongated ellipses, suggesting that the annotations along the valence dimension are more difficult to model from acoustic features. This is not surprising as much previous work has pointed out that valence perception is in nature much more subjective than activation perception and, thus, is more difficult to model [29].

Recall the definition of latent feature classes, the mapped VA Gaussians can show us the discriminability of each latent acoustic Gaussian. We could take away those ill-conditioned VA Gaussians as well as their mapped acoustic Gaussians ei-

[7] http://www.mturk.com

ther manually or via a stacked discriminative learning, e.g., by minimizing the prediction error with respect to Gaussian selection or adaptation on the VA GMM. In addition, the current latent acoustic Gaussian classifiers can also be replaced by kernel-based classifiers that map the frame vector space to a higher dimensional feature space, so that they may help generate a set of more discriminative VA Gaussians. We plan to leverage these insights to further improve the proposed framework in the future.

7. EVALUATION AND RESULTS

In this section, we first describe the evaluation setups for the emotion-based annotation and retrieval tasks, respectively. We then present the evaluation results of AEG with different settings of acoustic features and parameters, and compare them with that of two state-of-the-art methods.

7.1 Evaluation Setting and Metric

In Section 6.3, we have described the training setup for MER60. As for the MTurk corpus, the frame-based feature vectors of a song are provided by the authors of [26]. Since each piece has 15 second-by-second annotation sets (each annotation set is aligned with a specific second), we extract the frames corresponding to each one-second segment of each song. The resulting corpus contains 240×15=3,600 segments with corresponding annotations, which were normalized to $[-0.5, 0.5]$[8]. In addition, we select 300K frames at random from the remaining frames (excluding those annotated frames) of all songs to form the global frame set $\mathcal{F}$. The VA GMM is initialized with the global sample mean vector and covariance matrix of the emotion annotations of the whole training set.

The AEG model can predict the emotion of an input music piece as a mixture of VA Gaussians (i.e., a VA GMM) as well as a single VA Gaussian. Since measuring the distance between two GMMs is complicated and sometimes inaccurate [8],[9] we finally chose to evaluate the accuracy of emotion prediction based on a single Gaussian rather than a GMM. To evaluate the emotion annotation task, each set of ground truth annotations of a song is summarized by a ground truth Gaussian. The performance can be evaluated by two evaluation metrics, namely, the one-way KL divergence (cf. Eq. 19) and the Euclidean distance of mean vectors between the predicted and ground truth VA Gaussians for a clip.

For the emotion-based music retrieval task, we follow the music retrieval scenario described in Section 3, i.e., a user inputs a point query on the VA plane and receives a ranked list of pieces from an un-labeled music database. We randomly generate 500 2-D points uniformly covering the available VA space to form a test query set. To evaluate the retrieval performance using the test query set, we first derive the ground truth relevance $R(i)$ between a query point and the i-th piece in the database by feeding the query point into the ground truth Gaussian PDF of the i-th piece. For each test query point, the ground truth relevances is incorporated into the normalized discounted cumulative gain (NDCG) [10] to evaluate the ranking of pieces. The NDCG@P, which represents the quality of ranking of the top P retrieved music pieces for the query, is formulated as follows:

$$\text{NDCG@}P = \frac{1}{Q_P}\left\{R(1) + \sum_{j=2}^{P}\frac{R(j)}{\log_2 j}\right\}, \qquad (24)$$

where $R(j)$ is the ground truth relevance of the j-th piece on the ranked list, and Q_P is the normalization term that guarantees the ideal NDCG@P to be 1.

In the train/test evaluation scenario, we perform different settings for MER60 and MTurk. For MER60, we only execute the leave-one-out setting for the annotation task due to its limited number of songs [30]. That is, each song is taken as a test song in turn to predict its emotion annotation, and the remaining songs are used for training the models. For Mturk, in order to follow the work in [26], we randomly select 70% of songs (168 songs) for training and 30% (72 songs) for testing. The test set with 72×15=1,080 segments is used for emotion prediction and serves as the unseen music database for emotion-based retrieval. Since each segment has its unique audio content and the corresponding ground truth VA Gaussian, we treat each segment as an individual music piece for retrieval evaluation.

7.2 Results of Automatic Emotion Prediction

We evaluate the performance of emotion annotation in terms of average KL divergence (AKL) and average mean Euclidean distance (AMD) over the test set. Smaller AKL and AMD correspond to better performance. For MER60, the following factors in AEG are considered: the frame-level feature representation (either 39-D MFCCs or 70-D concatenated features), the number of latent feature classes K, and whether to use the annotation prior described in Section 4.2 or not. For example, “AEG-APrior-70DConcat” means using the annotation prior with the 70-D concatenated features. We test the AEG model with K ranging from 8 to 512. When the annotation prior is not used, we simply replace all γ_{ij} by 1 in the learning process. We compare the AEG method with support vector regression (SVR) [22] with different acoustic features. The SVR-Melody method using the melody features was the best performed setting reported in [30]. We also investigate the performance of SVR using the acoustic features used in our method.

From Figures 6 (a) and (b), we observe that the AEG method consistently outperforms the SVR method in almost all cases. Particularly, AEG-APrior-70DConcat (K=32) significantly outperforms SVR-Melody in terms of both AKL and AMD (p-value< 1% under the two-tailed t-test). The AKL shows a consistent tendency in performance difference for different AEG settings. In general, the annotation prior model improves the performance, and the 70-D concatenated features outperform the 39-D MFCCs when K is small. As for AMD, a general tendency is not observed. Interestingly, it seems that the performance of AEG-APrior-39DMFCC straightly improves as K increases, and that the performance does not saturate when K is 512. AEG obtains the lowest AKL when K=32. Such a result demonstrates its superior efficiency.

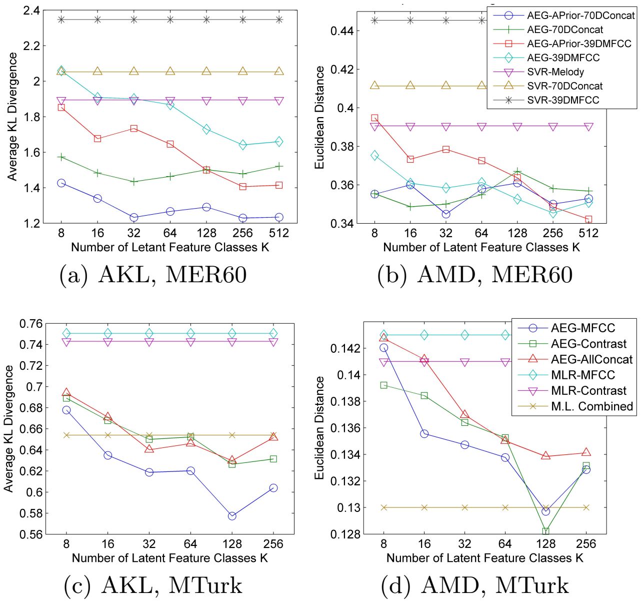
For MTurk, the annotation prior is adopted because of its superior performance on MER60. For each setting, we repeat the one-second prediction experiment 20 times and compute the average performance. We compare the AEG method with the systems in [26], including the multiple lin-

[8] The valence and activation of raw annotations derived from MTurk are both in the range of $[-200, 200]$.

[9] A single Gaussian representation is practically useful as explained in Secion 3.2.

Table 2: The comparison of automatic emotion annotation methods, each with the optimal setting.

Corpus	Method	AKL	AMD
MER60	SVR	1.894 ± 1.697	0.391 ± 0.177
	AEG	$\mathbf{1.230 \pm 1.421}$	$\mathbf{0.344 \pm 0.168}$
MTurk	MLR	0.654 ± 0.066	0.130 ± 0.006
	AEG	$\mathbf{0.577 \pm 0.031}$	0.128 ± 0.002



(a) AKL, MER60

(b) AMD, MER60

(c) AKL, MTurk

(d) AMD, MTurk

Figure 6: Results of automatic emotion annotation.

ear regression (MLR) method and the multi-layer combined (M.L. Combined) method, which employs multiple acoustic feature regressors and two layer prediction. M.L. Combined is considered as the ultimate fusion system in [19].

The results in Figures 6 (c) and (d) indicate that the AEG method significantly outperforms the MLR method using either MFCCs or Contrast features alone in almost all cases.[10] In many cases, AEG-MFCC and AEG-Contrast even outperform the M.L. Combined method in terms of AKL. We find that AEG does not benefit from the concatenated features on MTurk. On the contrary, AEG-MFCC performs consistently better than almost all other settings, except for AEG-Contrast in terms of AMD. The performance of AEG is observed to saturate when K is 128.

In general, the performance of AEG grows along with K when K is small, and then saturates after K is sufficiently large. This is reasonable as the value of K is closely related to the resolution in acoustic modeling and the model complexity. Nevertheless, the results suggest that the optimal value of K is corpus dependent, and it seems to be helpful to use a large K when the scale of the training corpus is large.

We summarize the comparison of best performance of AEG versus SVR and MLR fusion in Table 2. We also demonstrate the results of short-time emotion tracking in Figure 7. Empirically we found that it is easy to obtain better average performance and lower standard deviation for MTurk than for MER60 since the ground truth annotation

Table 3: The retrieval performance of different methods in terms of NDCG@5, 10, and 20.

Method (best K)	P=5	P=10	P=20
Predict MFCC (K=256)	0.886	0.825	0.785
Predict Contrast (K=256)	0.894	0.841	0.811
Predict AllConcat (K=128)	0.892	0.839	0.805
Fold-In MFCC (K=32)	0.874	0.814	0.767
Fold-In Contrast (K=32)	0.902	0.831	0.781
Fold-In AllConcat (K=64)	0.898	0.832	0.786

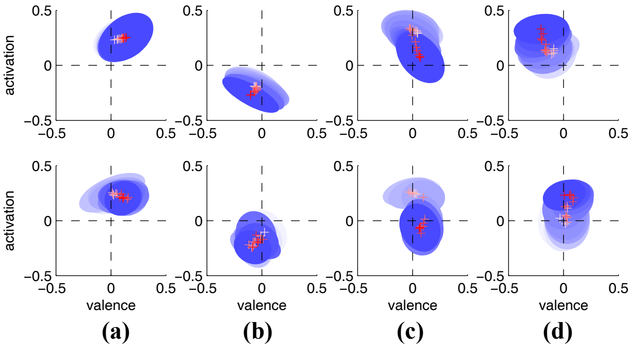


Figure 7: Demonstration for automatic emotion tracking of four 15-second clips on MTurk.[12] The top row shows the second-by-second ground truth 2-D annotation Gaussians (blue ellipse with a red cross representing the center), and the bottom row presents the corresponding predicted ones. The color gets darker as time unfolds.

Gaussians of MTurk are less diverse. This is evident from the fact that, when we measure the pair-wise KL divergence (PWKL) between the ground truth annotation Gaussians of each pair of clips in a corpus, the average PWKL for MTurk (1.985) is much smaller than that for MER60 (5.095). We also found that the annotation Gaussians of many clips in MTurk can be simply approximated by a Gaussian centered at the origin. But this is not the case for MER60, which consists of many clips whose annotation Gaussians are centered far away from the origin. Therefore, the evaluation results of MER60 should be considered more important and representative. In summary, the evaluation results demonstrate the effectiveness of the AEG model, and suggest that the proposed method indeed offers a potential solution for automatic VA-based emotion annotation.

7.3 Results of Emotion-Based Music Retrieval

The retrieval performance over the test query set of the pre-generated 500 points is given in Table 3. Each retrieval method with a type of acoustic features is evaluated with K ranging from 8 and 256, as done in the annotation task. Due to the space limit, we only show the best performance in Table 3 among different settings of K.

In general, the two retrieval methods exhibit similar performance. Interestingly, the Emotion Prediction method favors a larger K while Fold-In favors a smaller one. This makes sense since a larger K leads to better annotation performance, which would in turn benefit the Emotion Prediction based retrieval strategy. As for the Fold-In method, a

[10] Comparing to the standard formulation of KL divergence, the one used in [26] is two times larger. We therefore divided the KL divergences listed in [26] by two for fairness.

[12] The four clips are (a) *Same Old Song and Dance* by Aerosmith, (b) *Save Me* by Queen, (c) *Waitress* by Live, and (d) *Microphone Fiend* by Rage Against the Machine.

larger K might introduce randomness to the estimation of the corresponding pseudo song and, therefore, degrade the retrieval performance. When K is large, there will be multiple overlapping VA Gaussians in Figure 5. When a query point is located in an area with multiple overlapping VA Gaussians, the pseudo song estimated by Eq. 22 would be highly sensitive to the initialization of the model parameters. This would reduce the discriminative power in transforming a proper mapped pseudo song. We also notice that Fold-In achieves better performance when P (i.e., the number of retrieved pieces) in NDCG is small.

8. CONCLUSION AND FUTURE WORK

This paper has presented the novel acoustic emotion Gaussians model, which provides a principled probabilistic framework for music emotion analysis. We have interpreted the learning of AEG and provided insights. A comprehensive evaluation of automatic music emotion annotation on two emotion annotated music corpora has demonstrated the effectiveness of AEG over state-of-the-art methods. Moreover, we have also demonstrated the potential application of AEG for emotion-based music retrieval.

Our future work is four-fold. First, the probabilistic framework provides a blueprint for multi-modal music emotion modeling. The latent feature classes can not only be defined from music signals, but also from other relevant modalities such as lyrics, tags, and music video. Second, with a set of learned latent feature classes, multi-modal features can be aligned to one another. For example, if we have a *tagged* corpus and a *VA-annotated* corpus, we can automatically position the mood tags in the VA space by means of the fixed acoustic GMM. Third, we will investigate personalized music emotion recommendation. The AEG model can be used as a background model for general users. A collection of personal feedback data can then be used to adapt either acoustic GMM or VA GMM of the background model to create a personalized one. Finally, it would be interesting to study non-parametric Bayesian methods for the modeling of music emotion, and to investigate whether we can get better results without assuming that the VA space is Euclidean.

9. ACKNOWLEDGMENTS

This work was supported by the National Science Council of Taiwan under Grants: NSC-101-2221-E-001-019, NSC-101-2631-H-001-007, and NSC-100-2218-E-001-009. We thank the reviewers for their helpful comments.

10. REFERENCES

[1] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer-Verlag New York, Inc., 2006.

[2] R. A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Trans. Affective Computing*, 1(1):18–37, 2010.

[3] J. V. Davis and I. S. Dhillon. Differential entropic clustering of multivariate Gaussians. In *NIPS*, 2006.

[4] T. Eerola, O. Lartillot, and P. Toiviainen. Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *ISMIR*, 2009.

[5] A. Gabrielsson. Emotion perceived and emotion felt: Same or different? *Musicae Scientiae*, pages 123–147, 2002.

[6] P. Gomez and B. Danuser. Relationships between musical structure and psychophysiological measures of emotion. *Emotion*, 7(2):377–87, 2007.

[7] S. Hallam, I. Cross, and M. Thaut. *The Oxford Handbook of Music Psychology.* Oxford University Press, 2008.

[8] J. Hershey and P. Olsen. Approximating the Kullback Leibler divergence between Gaussian mixture models. In *ICASSP*, pages 317–320, 2007.

[9] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann. The 2007 MIREX audio mood classification task: Lessons learned. In *ISMIR*, 2008.

[10] K. Jarvelin and J. Kekalainen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. on Info. Syst.*, 20(4):422–446, 2002.

[11] P. N. Juslin and P. Laukka. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *J. New Music Res.*, 33(3):217–238, 2004.

[12] Y. E. Kim and et al. Music emotion recognition: A state of the art review. In *ISMIR*, 2010.

[13] M. D. Korhonen, D. A. Clausi, and M. E. Jernigan. Modeling emotional content of music using system identification. *IEEE Trans. System, Man and Cybernetics*, 36(3):588–599, 2006.

[14] O. Lartillot and P. Toiviainen. A Matlab toolbox for musical feature extraction from audio. In *DAFx*, 2007.

[15] L. Lu, D. Liu, and H. Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Trans. Audio, Speech and Lang. Proc.*, 14(1):5–18, 2006.

[16] 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.*, 36(4):281–299, 2007.

[17] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *J. Machine Learning Res.*, 11:1297–1322, 2010.

[18] J. A. Russell. A circumplex model of affect. *J. Personality and Social Science*, 39(6):1161–1178, 1980.

[19] E. M. Schmidt and Y. E. Kim. Prediction of time-varying musical mood distributions from audio. In *ISMIR*, 2010.

[20] E. M. Schmidt and Y. E. Kim. Prediction of time-varying musical mood distributions using Kalman filtering. In *ICMLA*, 2010.

[21] E. M. Schmidt and Y. E. Kim. Modeling musical emotion dynamics with conditional random fields. In *ISMIR*, 2011.

[22] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[23] E. Schubert. Modeling perceived emotion with continuous musical features. *Music Perception*, 21(4):561–585, 2004.

[24] B. Schuller, C. Hage, D. Schuller, and G. Rigoll. "Mister D.J., Cheer Me Up!": Musical and textual features for automatic mood classification. *J. New Music Res.*, 39(1):13–34, 2010.

[25] R. Sease and D. W. McDonald. The organization of home media. *ACM Trans. Com.-Hum. Interact.*, 18:1–20, 2011.

[26] J. A. Speck, E. M. Schmidt, B. G. Morton, and Y. E. Kim. A comparative study of collaborative vs. traditional musical mood annotation. In *ISMIR*, 2011.

[27] W. Trost, T. Ethofer, M. Zentner, and P. Vuilleumier. Mapping aesthetic musical emotions in the brain. *Cerebral Cortex*, 2011.

[28] J.-C. Wang, H.-S. Lee, H.-M. Wang, and S.-K. Jeng. Learning the similarity of audio music in bag-of-frames representation from tagged music data. In *ISMIR*, 2011.

[29] Y.-H. Yang and H. H. Chen. *Music Emotion Recognition.* CRC Press, Feb. 2011.

[30] Y.-H. Yang and H. H. Chen. Predicting the distribution of perceived emotions of a music signal for content retrieval. *IEEE Trans. Audio, Speech and Lang. Proc.*, 19(7):2184–2196, 2011.

[31] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen. A regression approach to music emotion recognition. *IEEE Trans. Audio, Speech and Lang. Proc.*, 16(2):448–457, 2008.