Affective Music Information Retrieval

Ju-Chiang Wang,* Yi-Hsuan Yang, and Hsin-Min Wang

Institute of Information Science,
Academia Sinica, Taipei, Taiwan
* Corresponding e-mail: asriver.wang@gmail.com

February 19, 2015

Abstract

Much of the appeal of music lies in its power to convey emotions/moods and to evoke them in listeners. In consequence, the past decade witnessed a growing interest in modeling emotions from musical signals in the music information retrieval (MIR) community. In this article, we present a novel generative approach to music emotion modeling, with a specific focus on the valence-arousal (VA) dimension model of emotion. The presented generative model, called *acoustic emotion Gaussians* (AEG), better accounts for the subjectivity of emotion perception by the use of probability distributions. Specifically, it learns from the emotion annotations of multiple subjects a Gaussian mixture model in the VA space with prior constraints on the corresponding acoustic features of the training music pieces. Such a computational framework is technically sound, capable of learning in an online fashion, and thus applicable to a variety of applications, including user-independent (general) and user-dependent (personalized) emotion recognition and emotion-based music retrieval. We report evaluations of the aforementioned applications of AEG on a larger-scale emotion-annotated corpora, AMG1608, to demonstrate the effectiveness of AEG and to showcase how evaluations are conducted for research on emotion-based MIR. Directions of future work are also discussed.

1 Introduction

Automatic music emotion recognition (MER) aims at modeling the association between music and emotion so as to facilitate emotion-based music organization, indexing, and retrieval. This technology has emerged in recent years as a promising solution to deal with the huge amount of music information available digitally [1,22,30,73]. It is generally believed that music cannot be composed, performed, or listened to without affection involvement [29]. The pursuit of emotional experience has also been identified as one of the primary motivations and benefits of music listening [27]. In addition to music retrieval, music emotion also finds applications in context-aware music recommendation, playlist generation, music therapy, and automatic music accompaniment for other media content, including image, video, and text, amongst others [34, 47, 62, 77].

Despite of the significant progress that has been made in recent years, MER is still considered as a challenging problem because the perception of emotion in music is usually highly subjective. A single, static ground-truth emotion label is not sufficient to describe the possible emotions different people perceive in the same piece of music [14, 23]. On the contrary, it may be more reasonable to learn a computational model from multiple responses of different listeners [43] and to present *probabilistic* (soft) rather than *deterministic* (hard) emotion assignments as the final result. In addition, the subjective nature of emotion perception suggests the need of personalization in systems for emotion-based music recommendation or retrieval [74]. Early work on MER often chose to sidestep this critical issue by either assuming that a common consensus can be achieved [22, 66], or by simply discarding music pieces for which a common consensus cannot be achieved [35].

To help address this issue, we have proposed a novel generative model referred to as *acoustic emotion Gaussians* (AEG) in our prior work [61–65]. The name of the AEG model comes from its use of multiple Gaussian distributions to model the affective content of music. The algorithmic part of AEG has been first introduced in [63], along with the preliminary evaluation of AEG for MER and emotion-based music retrieval. More details about the analysis part of the model learning of AEG can be found in a recent article [65]. Due to the parametric nature of AEG, model adaptation techniques have also been proposed to personalize an AEG model in an online, incremental fashion, rather than learning from scratch [6, 64]. The goal of this article is to position the AEG model as a theoretical framework and to provide detailed information about the model itself and its application to personalized MER and emotion-based music retrieval.

We conceptualize emotion by the valence-arousal (VA) model [45], which

has been used extensively by psychologists to study the relationship between music and emotion [13, 52]. These two dimensions are found to be the most fundamental through factor analysis of self-report of human's affective response to music stimulus. Despite differences in nomenclature, existing studies give similar interpretations of the resulting factors, most of which correspond to *valence* (or pleasantness; positive/negative affective states) and *arousal* (or activation; energy and stimulation level). For example, happiness is an emotion associated with a positive valence and a high arousal, while sadness is an emotion associated with a negative valence and a low arousal. We refer to the 2-D space spanned by valence and arousal as the *VA space* hereafter. Moreover, we are concerned with the emotion an individual *perceives* as being expressed in a piece of music, rather than the emotion the individual actually *feels* in response to the piece. This distinction is necessary [14], as we do not necessarily feel sorrow when listening to a sad tune, for example.

As the focus of this article is on *dimensional* emotion values such as valence and arousal values, we refer interested readers to [1, 20, 53] for studies and surveys on *categorical* MER research that views emotions as discrete labels such as mood tags. We also note that people have proposed approaches to model the relationship between discrete emotion labels and the dimensional VA values [46, 61], which is also beyond the scope of this article.

The article is organized as follows. We first review related work in Section 2. Then, we present the mathematical derivation of AEG and the learning algorithm in Section 3, followed by the personalization algorithm in Section 4. Sections 5 and 6 present applications of AEG to MER and emotion-based music retrieval, respectively. Finally, we conclude in Section 7.

2 Related Work on Dimensional Music Emotion Recognition

Early approaches to MER [36, 76] assumed that the perceived emotion of a music piece can be represented as a *single point* in the VA space, in which the valence and arousal values are considered as independent numerical values. The ground-truth VA values of a music piece is obtained by averaging the annotations of a number of human subjects, without considering the covariance of the annotations. To predict the VA values of a music piece, a regression model can be applied. Given N inputs $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, N$, where $\mathbf{x}_i$ is a D-dimensional feature vector of the i-th input segment, D the number of feature descriptors, and y_i the valence or arousal value, a regression model

is learned by algorithms such as support vector regression (SVR) [51] that minimize the mismatch (e.g. mean squared loss) between the predicted and the ground-truth VA values.

As emotion perception is rarely dependent on a single music factor but a combination of them [18, 28], algorithms used feature descriptors that characterize the loudness, timbre, pitch, rhythm, melody, harmony or lyrics of music [19, 40, 50, 53]. In particular, while it is usually easier to predict arousal using, for example, loudness and timbre features, the prediction of valence has been found more challenging [53,68,72]. Cross cultural aspects of emotion perception have also been studied [21]. To exploit the temporal continuity of emotion variation within a piece of music, techniques such as system identification [31], conditional random fields [24, 49], hidden Markov models [37], deep recurrent neural networks [69], or dynamic probabilistic model [60] have also been proposed. Various approaches and features for MER have been evaluated and compared using benchmarking datasets comprising over 1,000 Creative Commons licensed music pieces from the Free Music Archive, in the 2013 and 2014 MediaEval 'Emotion in Music' tasks [55, 56].

Recent years have witnessed growing attempts to model the emotion of a music piece as a probability distribution in the VA space [6, 48, 63, 71] to better account for the subjective nature of emotion perception. For instance, Figure 1 shows the VA values applied by different annotators to four music pieces. To characterize the distribution of the emotion annotations for each clip, a typical way is to use a bivariate Gaussian distribution, where the mean vector presents the most possible VA values and the covariance matrix indicates its uncertainty. For a clip with highly subjective affective content, the determinant of the covariance matrix would be larger.

Existing approaches to predicting the emotion distribution of a music clip from acoustic features fall into two categories. The *heatmap* approach [49,71] quantizes each emotion dimension by W equally spaced cells, leading to a $W \times W$ grid representation of the VA space. The approach trains W^2 regression models for predicting the emotion *intensity* of each cell. Higher intensity at a cell indicates that people are more likely to perceive the corresponding emotion from the clip. The emotion intensity over the VA space creates a heatmap-like representation of emotion distribution. However, heatmap is not a continuous representation of emotion and emotion intensity cannot be strictly considered as a probability estimate.

The *Gaussian-parameter* approach [48, 71], on the other hand, models emotion distribution of a clip as a bivariate Gaussian and trains multiple regressors, each for a parameter of the mean vector and the covariance matrix. This makes it easy to apply lessons learned from modeling the mean VA values. In addition, performance analysis of this approach is easier; one

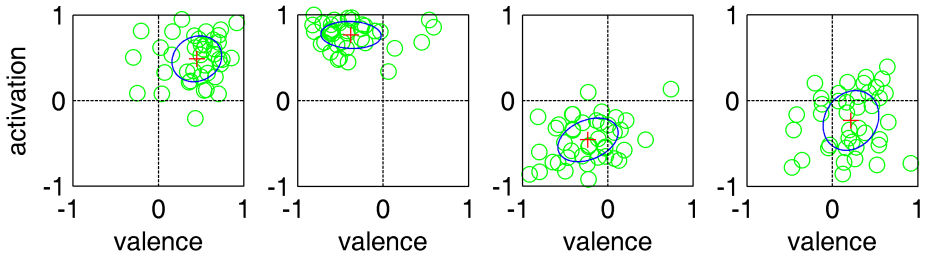


Figure 1: Subjects' annotations of the perceived emotion of four 30-second clips, which from left to right are *Dancing Queen* by ABBA, *Civil War* by Guns N' Roses, *Suzanne* by Leonard Cohen, and *All I Have To Do Is Dream* by the Everly Brothers. Each circle here corresponds to a subject's annotation, and the overall emotion for a clip can be approximated by a 2-D Gaussian distribution (the red cross and blue ellipse). Note that throughout this article we use the contour of an ellipse to outline the standard deviation of the corresponding Gaussian distribution.

can analyze the importance of different acoustic features to each Gaussian parameter individually. However, since the regression models are trained independently, the correlation between valence and arousal is not exploited. The parameter estimation of the mean and variance is disjoined as well.

A different methodology to address the subjectivity is to call for a user-dependent model trained on annotations of a specific user to personalize the emotion prediction [78–80]. In [78], two personalization methods are proposed; the first trains a *personalized* MER system for each individual specifically, whereas the second groups users according to some personal factors (e.g. gender, music experience, and personality) and then trains *group-wise* MER system for each user group. Another *two-stage* personalization scheme has also been studied [74]: the first stage estimates the general perception of a music piece, whereas the second one predicts the difference between the general perception and the personal one of the target user.

We note that none of the aforementioned approaches renders a strict probabilistic interpretation [65]. In addition, many existing work is developed on discriminative models such as multiple linear regression and SVR. Few attempts are made to develop a principled probabilistic framework that is technically sound for modeling the music emotion and that permits extending the user-independent model to a user-dependent one, preferably in an online fashion.

We also note that most existing work focuses on the *annotation* aspect of music emotion research, namely MER. Little work has been made to the

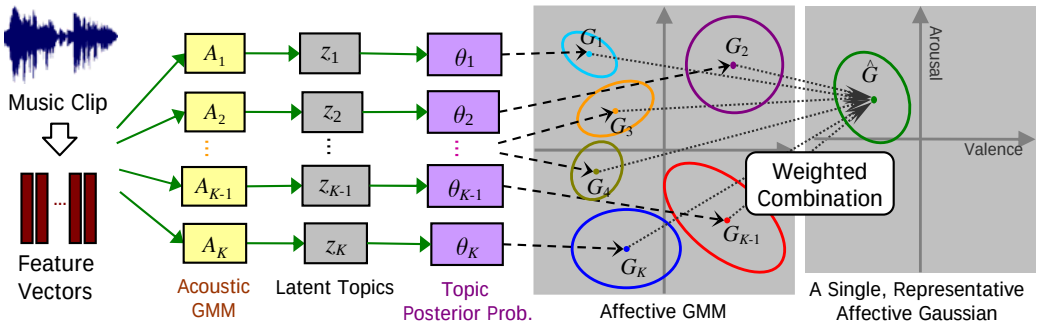


Figure 2: Illustration of the generative process of the AEG model.

retrieval aspect – the development of emotion-based music retrieval systems [73]. In what follows, we present the AEG model and its applications to the both of these two aspects.

3 Acoustic emotion Gaussians: A Generative Approach for Music Emotion Modeling

In [61–65], we proposed AEG, which is fundamentally different from the existing regression or heatmap approaches. As Figure 2 shows, AEG involves the generative process of VA emotion distributions from audio signals. While the relationship between audio and music emotion may sometimes be complicated and difficult to observe directly from an emotion-annotated corpus, AEG uses a set of clip-level *latent topics* $\{z_k\}_{k=1}^K$ to resolve this issue.

We first define the terminology and explain the basic principle of AEG. Suppose that there are K *audio descriptors* $\{A_k\}_{k=1}^K$, each is related to some acoustic feature vectors of music clips. Then, we map the associated feature vectors of A_k to a clip-level topic z_k. To implement each A_k, we use a single Gaussian distribution in the acoustic feature space. The aggregated Gaussians of $\{A_k\}_{k=1}^K$ is called an *acoustic GMM* (Gaussian mixture model). Subsequently, we map each z_k to a specific area in the VA space, which is modeled by a bivariate Gaussian distribution G_k. We refer to the aggregated Gaussians of $\{G_k\}_{k=1}^K$ as an *affective GMM*. Given a clip, its feature vectors are first used to compute the posterior distribution over the topics, termed as a *topic posterior representation* $\boldsymbol{\theta}$. In $\boldsymbol{\theta}$, the posterior probability of z_k (denoted as θ_k) is associated with A_k and will then be used to show the clip's importance to G_k. Consequently, the posterior distribution $\boldsymbol{\theta} = \{\theta_k\}_{k=1}^K$ can be incorporated into learning the affective GMM as well as making emotion prediction for a clip.

AEG-based MER follows the flow depicted in Figure 2. Based on $\boldsymbol{\theta}$ of a test clip, we obtain the weighted affective GMM $\sum_k \theta_k G_k$, which is able to generate various emotion distribution. Following this sense, if a clip's acoustic features can be completely described by the h-th topic z_h, i.e. $\theta_h = 1$, and $\theta_k = 0$, $\forall k \neq h$, then its emotion distribution would exactly follow G_h. As will be described in Section 5, we can further approximate $\sum_k \theta_k G_k$ by a single, representative affective Gaussian $\hat{G}$ for simplicity. This is illustrated in the rightmost of Figure 2.

Beyond valence and arousal, adding more dimensions (e.g. *potency*, or dominant–submissive) might help resolve the ambiguity between affective terms, such as anger and fear, which are close to one another in the second quadrant of the VA space [2, 10]. Although AEG can be easily extended to describe emotion in higher dimensions, we stay with the 2-D emotion model here again for simplicity.

3.1 Topic Posterior Representation

The topic posterior representation of a music clip is generated from its audio. We note that the temporal dynamics of audio signals is regarded as essential for human to perceive musical characteristics such as timbre, rhythm, and tonality. To capture more local temporal variation of the low-level features, we represent the acoustic features at a time instance in the segment-level, which corresponds to sufficiently long duration (e.g. 0.4 second). A segment-level feature vector $\mathbf{x}$ can be formed by, for example, concatenating the mean and standard deviation of the frame-level feature vectors within the segment. As a result, a clip is divided into multiple overlapped segments which are then represented by a sequence of vectors, $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, where T is the length of the clip.

To start the generative process of AEG, we first learn an acoustic GMM as the bases to represent a clip. This acoustic GMM can be trained using the expectation-maximization (EM) algorithm on a large set of segment-level vectors $\mathcal{F}$ extracted from existing music clips. The learned acoustic GMM defines the set of audio descriptors $\{A_k\}_{k=1}^K$, and can be expressed as follows,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k A_k(\mathbf{x} \mid \mathbf{m}_k, \mathbf{S}_k)\,, \tag{1}$$

where $A_k(\cdot)$ is the k-th component Gaussian distribution, and π_k, $\mathbf{m}_k$, and $\mathbf{S}_k$ are its corresponding prior weight, mean vector, and covariance matrix, respectively. Note that we substitute equal weight for the GMM (i.e. $\pi_k = \frac{1}{K}$, $\forall k$), because the original π_k learned from $\mathcal{F}$ does not imply the prior

distribution of the feature vectors in a clip. Such a heuristic usually results in better performance as pointed in [58].

Suppose that we have an emotion annotated corpus $\mathcal{X}$ consisting of N music clips $\{s_i\}_{i=1}^{N}$. Given a clip $s_i = \{\mathbf{x}_{i,t}\}_{t=1}^{T_i}$, we then compute the segment-level posterior probability for each feature vector in s_i based on the acoustic GMM,

$$p(A_k \mid \mathbf{x}_{i,t}) = \frac{A_k(\mathbf{x}_{i,t} \mid \mathbf{m}_k, \mathbf{S}_k)}{\sum_{h=1}^{K} A_h(\mathbf{x}_{i,t} \mid \mathbf{m}_h, \mathbf{S}_h)}\,. \tag{2}$$

Finally, the clip-level topic posterior probability $\theta_{i,k}$ of s_i can be approximated by averaging the segment-level ones,

$$\theta_{i,k} \leftarrow p(z_k \mid s_i) \approx \frac{1}{T_i} \sum_{t=1}^{T_i} p(A_k \mid \mathbf{x}_{i,t})\,. \tag{3}$$

This approximation assumes that $\theta_{i,k}$ is equally contributed by each segment of s_i and thereby capable of representing the clip's acoustic features. We use a vector $\boldsymbol{\theta}_i \in \mathbb{R}^K$, whose k-th component is $\theta_{i,k}$, as the topic posterior of s_i.

3.2 Prior Model for Emotion Annotation

To consider the subjectivity of emotional responses of a music clip, we ask multiple subjects to annotate the clip. However, as some subjects' annotations may not be reliable, we introduce a *user prior model* to quantify the contribution of each subject.

Let $\mathbf{e}_{i,j} \in \mathbb{R}^2$ (a vector including the valence and arousal values) denote one of the annotations of s_i given by the j-th subject, and let U_i denote the number of subjects who have annotated s_i. Note that $\mathbf{e}_{q,j}$ and $\mathbf{e}_{r,j}$, where $q \neq r$, may not correspond to the same subject. Then, we build the user prior model γ to describe the confidence of $\mathbf{e}_{i,j}$ in s_i using a single Gaussian distribution,

$$\gamma(\mathbf{e}_{i,j} \mid s_i) \equiv G(\mathbf{e}_{i,j} \mid \mathbf{a}_i, \mathbf{B}_i), \tag{4}$$

where $\mathbf{a}_i = \frac{1}{U_i}\sum_{j=i}^{U_i} \mathbf{e}_{i,j}$, $\mathbf{B}_i = \frac{1}{U_i}\sum_{j=1}^{U_i}(\mathbf{e}_{i,j} - \mathbf{a}_i)(\mathbf{e}_{i,j} - \mathbf{a}_i)^T$, and $G(\mathbf{e} \mid \mathbf{a}_i, \mathbf{B}_i)$ is called the *annotation Gaussian* of s_i. One can observe what $\mathbf{a}_i$ and $\mathbf{B}_i$ look like from the four example clips in Figure 1. Empirical results show that a single Gaussian performs better than a GMM for setting up $\gamma(\cdot)$ [63].

The confidence of $\mathbf{e}_{i,j}$ can be estimated based on the likelihood calculated by Eq. 4. If an annotation is far away from the mean, it gives small likelihood accordingly. In addition to Gaussian distributions, any criterion that is able to reflect the importance of a user's annotation of a clip can be applied to γ.

The probability of $\mathbf{e}_{i,j}$, referred to as the *clip-level annotation prior*, can be calculated by normalizing the likelihood of $\mathbf{e}_{i,j}$ over the cumulative likelihood of all other annotations in s_i,

$$p(\mathbf{e}_{i,j} \mid s_i) \equiv \frac{\gamma(\mathbf{e}_{i,j} \mid s_i)}{\sum_{r=1}^{U_i} \gamma(\mathbf{e}_{i,r} \mid s_i)} \,. \tag{5}$$

Based on the clip-level annotation prior, we further define the *corpus-level clip prior* to describe the importance of each clip,

$$p(s_i \mid \mathcal{X}) \equiv \frac{\sum_{j=1}^{U_i} \gamma(\mathbf{e}_{i,j} \mid s_i)}{\sum_{q=1}^{N} \sum_{r=1}^{U_q} \gamma(\mathbf{e}_{q,r} \mid s_q)} \,. \tag{6}$$

From Eqs. 5 and 6 we can make two observations. First, if a clip's annotations are consistent (i.e. $\mathbf{B}_i$ is small), it is considered less subjective. Second, if a clip is annotated by more subjects, the corresponding γ model should be more reliable. As a result, we can define the *corpus-level annotation prior* $\gamma_{i,j}$ for each $\mathbf{e}_{i,j}$ in the corpus $\mathcal{X}$ by multiplying Eqs. 5 and 6:

$$\gamma_{i,j} \leftarrow p(\mathbf{e}_{i,j} \mid \mathcal{X}) \equiv \frac{\gamma(\mathbf{e}_{i,j} \mid s_i)}{\sum_{q=1}^{N} \sum_{r=1}^{U_q} \gamma(\mathbf{e}_{q,r} \mid s_i)} \,, \tag{7}$$

which is computed beforehand and fixed in learning the affective GMM.

3.3 Learning the Affective GMM

Given a training music clip s_i in the corpus $\mathcal{X}$, we assume the emotional responses can be generated from an affective GMM weighted by its topic posterior $\boldsymbol{\theta}_i$,

$$p(\mathbf{e}_{i,j} \mid \boldsymbol{\theta}_i) = \sum_{k=1}^{K} \theta_{i,k} G_k(\mathbf{e}_{i,j} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \,, \tag{8}$$

where $G_k(\cdot)$ is the k-th affective Gaussian with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$ to be learned. Here $\theta_{i,k}$ stands for the fixed weight associated with A_k to carry the audio characteristics of s_i. We therefore call $\boldsymbol{\theta}_i$ an *acoustic prior*. Then, the objective function is in the form of the marginal likelihood function

of the annotations:

$$
\begin{aligned}
p(\mathbf{E} \mid \mathcal{X}, \mathbf{\Lambda}) &= \sum_{i=1}^{N} p(s_i \mid \mathcal{X}) \sum_{j=1}^{U_i} p(\mathbf{e}_{i,j} \mid s_i) p(\mathbf{e}_{i,j} \mid \boldsymbol{\theta}_i, \mathbf{\Lambda}) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{U_i} p(s_i \mid \mathcal{X}) p(\mathbf{e}_{i,j} \mid s_i) p(\mathbf{e}_{i,j} \mid \boldsymbol{\theta}_i, \mathbf{\Lambda}) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{U_i} p(\mathbf{e}_{i,j} \mid \mathcal{X}) \sum_{k=1}^{K} \theta_{i,k} G_k(\mathbf{e}_{i,j} \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) \,,
\end{aligned} \tag{9}
$$

where $\mathbf{E} = \{\mathbf{e}_{i,j}\}_{i=1,j=1}^{N,U_i}$, $\mathcal{X} = \{s_i, \boldsymbol{\theta}_i\}_{i=1}^{N}$, and $\mathbf{\Lambda} = \{\boldsymbol{\mu}_k, \mathbf{\Sigma}_k\}_{k=1}^{K}$ is the parameter set of the affective GMM. Taking the logarithm of Eq. 9 and replacing $p(\mathbf{e}_{i,j} \mid \mathcal{X})$ by $\gamma_{i,j}$ leads to

$$
L = \log \sum_i \sum_j \gamma_{i,j} \sum_k \theta_{i,k} G_k(\mathbf{e}_{i,j} \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) \,, \tag{10}
$$

where $\sum_i \sum_j \gamma_{i,j} = 1$. To learn the affective GMM, we can maximize the log-likelihood in Eq. 10 with respect to the Gaussian parameters. We first derive a lower bound of L according to Jensen's inequality,

$$
L \geq L_{\text{bound}} = \sum_i \sum_j \gamma_{i,j} \log \sum_k \theta_{i,k} G_k(\mathbf{e}_{i,j} \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) \,. \tag{11}
$$

Then, we treat L_{bound} as a surrogate of L and use the EM algorithm [3] to estimate the parameters of the affective GMM. In the E-step, we derive the expectation over the posterior distribution of z_k for all the training annotations,

$$
Q = \sum_i \sum_j \gamma_{i,j} \sum_k p(z_k \mid \mathbf{e}_{i,j}) \Big(\log \theta_{i,k} + \log G_k(\mathbf{e}_{i,j} \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) \Big) \,, \tag{12}
$$

where

$$
p(z_k \mid \mathbf{e}_{i,j}) = \frac{\theta_{i,k} G_k(\mathbf{e}_{i,j} \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)}{\sum_{h=1}^{K} \theta_{i,h} G_k(\mathbf{e}_{i,j} \mid \boldsymbol{\mu}_h, \mathbf{\Sigma}_h)} \,. \tag{13}
$$

In the M-step, we first set the derivative of Eq. 12 with respect to $\boldsymbol{\mu}_k$ to zero and obtain the updating form for the mean vector,

$$
\boldsymbol{\mu}_k' \leftarrow \frac{\sum_i \sum_j \gamma_{i,j} p(z_k \mid \mathbf{e}_{i,j}) \mathbf{e}_{i,j}}{\sum_i \sum_j \gamma_{i,j} p(z_k \mid \mathbf{e}_{i,j})} \,. \tag{14}
$$

Following a similar line of reasoning, we obtain the update rule for $\mathbf{\Sigma}_k$:

$$\mathbf{\Sigma}'_k \leftarrow \frac{\sum_i \sum_j \gamma_{i,j} p(z_k \mid \mathbf{e}_{i,j})(\mathbf{e}_{i,j} - \boldsymbol{\mu}'_k)(\mathbf{e}_{i,j} - \boldsymbol{\mu}'_k)^T}{\sum_i \sum_j \gamma_{i,j} p(z_k \mid \mathbf{e}_{i,j})} . \tag{15}$$

Theoretically, the EM algorithm iteratively maximizes the L_{bound} value in Eq. 11 until convergence. One can fix the number of maximal iterations or set a stopping criterion for the increasing ratio of L_{bound}.

Note that we can ignore the annotation prior by setting a uniform distribution, i.e., $\forall i, j,\ \gamma_{i,j} = 1$. This case is called "AEG Uniform" in the experiment. In contrast, the case with non-uniform annotation prior is called "AEG AnnoPrior."

3.4 Discussion

As Eqs. 14 and 15 show, the re-estimated parameters $\boldsymbol{\mu}'_k$ and $\mathbf{\Sigma}'_k$ are collectively contributed by $\mathbf{e}_{i,j}, \forall\ i,\ j$, with the weights governed by the product of $\gamma_{i,j}$ and $p(z_k \mid \mathbf{e}_{i,j})$. Consequently, the learning process seamlessly takes the annotation prior, acoustic prior, and annotation clusters over the current affective GMM into consideration. In such a way, the annotations of different clips can be shared with one another according to their corresponding prior probabilities. This can be a key factor that enables AEG to generalize the audio-to-emotion mapping.

As the affective GMM is getting fitted to the data, a small number of affective Gaussian components might overly fit to some emotion annotations, rendering the so-called *singularity* problem [3]. When this occurs, the corresponding covariance matrices would become non-positive definite (non-PD). Imagining that when a component affective Gaussian is contributed by only one or two annotations, the corresponding covariance shape will become a point or a straight line in the VA space. To tackle this issue, we can remove the component Gaussian when it happens to produce a non-PD covariance matrix during the EM iterations [65].

We note that "early stop" is a very important heuristic while learning the affective GMM. We find that setting a small number for the maximal iteration (e.g. 7 – 11) or a larger stopping threshold for the increasing ratio of L_{bound} (e.g. 0.01) empirically leads to better generalizability. It can not only prevent the aforementioned singularity problem but also avoid overly fitting to the training data. Empirical results show that the accuracy of MER improves as the iteration evolves and then degrades when the optimal iteration number has reached [65]. Moreover, AEG AnnoPrior empirically converges faster and learns smaller covariances than AEG Uniform does.

4 Personalization with AEG

The capability for personalization is a very important characteristic that completes the AEG framework, making it more applicable to real-world applications. As AEG is a probabilistic, parametric model, it can incorporate personal information of a particular user via model adaptation techniques to make custom predictions. While such personal information may include personal emotion annotation, user profile, transaction records, listening history, and relevance feedback, we focus on the use of personal emotion annotations in this article.

Because of the cognitive load for annotating music emotion, it is usually not easy to collect a sufficient amount of personal annotations at once to make the system reach an acceptable performance level. On the contrary, a user may provide annotations sporadically in different listening sessions. To this end, an online learning strategy [4] is desirable. When the annotations of a target user are scarce, a good online learning method needs to prevent over-fitting to the personal data in order to keep certain model generalizability. In other words, we cannot totally ignore the contributions of emotion perceptions from other users. Motivated by the Gaussian Mixture Model-Universal Background Model (GMM-UBM) speaker verification system [44], we first treat the affective GMM learned from broad subjects (called *background users*) as a *background (general) model*, and then employ a *maximum a posteriori* (MAP)-based method [15, 44] to update the parameters of the background model using the personal annotations in an online manner. Theoretically, the resulting *personalized model* will appropriately find a good trade-off between the target user's annotations and the background model.

4.1 Model Adaptation

In what follows, the acoustic GMM will stay fixed throughout the personalization process, since it is used as a reference model to represent the music audio. In contrast, the affective GMM is assumed to be learned on plenty of emotion annotations from quite a few subjects, so it possesses a sufficient representation (well-trained parameters) for user-independent (i.e. general) emotion perceptions. Our goal is to learn the personal perception with respect to the affective GMM $\mathbf{\Lambda}$ accordingly.

Suppose that we have a target user $u_\star$ annotating M number of music clips denoted as $\mathcal{X}_\star = \{\mathbf{e}_i, \boldsymbol{\theta}_i\}_{i=1}^{M}$, where $\mathbf{e}_i$ and $\boldsymbol{\theta}_i$ are the emotion annotation and the topic posterior of a clip, respectively. We first compute each posterior

probability over the latent topics based on the background affective GMM,

$$p(z_k \mid \mathbf{e}_i, \boldsymbol{\theta}_i) = \frac{\theta_{i,k} G_k(\mathbf{e}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{h=1}^{K} \theta_{i,h} G_k(\mathbf{e}_i \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}. \tag{16}$$

Then, we derive the expected sufficient statistics on $\mathcal{X}_\star$ over the posterior distribution of $p(z_k \mid \mathbf{e}_i, \boldsymbol{\theta}_i)$ for the mixture weight, mean, and covariance parameters:

$$\Gamma_k = \sum_{i=1}^{M} p(z_k \mid \mathbf{e}_i, \boldsymbol{\theta}_i)\,, \tag{17}$$

$$\mathbb{E}(\boldsymbol{\mu}_k) = \frac{1}{\Gamma_k} \sum_{i=1}^{M} p(z_k \mid \mathbf{e}_i, \boldsymbol{\theta}_i) \mathbf{e}_i\,, \tag{18}$$

$$\mathbb{E}(\boldsymbol{\Sigma}_k) = \frac{1}{\Gamma_k} \sum_{i=1}^{M} p(z_k \mid \mathbf{e}_i, \boldsymbol{\theta}_i) \big(\mathbf{e}_i - \mathbb{E}(\boldsymbol{\mu}_k)\big)\big(\mathbf{e}_i - \mathbb{E}(\boldsymbol{\mu}_k)\big)^T. \tag{19}$$

Finally, the new parameters of the personalized affective GMM can be obtained according to the MAP criterion [15]. The resulting update rules are the forms of interpolations between the expected sufficient statistics (i.e. $E(\boldsymbol{\mu}_k)$ and $E(\boldsymbol{\Sigma}_k)$) and the parameters of the background model (i.e. $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$) as follows:

$$\boldsymbol{\mu}'_k \leftarrow \alpha_k^{\mathrm{m}} \mathbb{E}(\boldsymbol{\mu}_k) + (1 - \alpha_k^{\mathrm{m}})\, \boldsymbol{\mu}_k\,, \tag{20}$$

$$\boldsymbol{\Sigma}'_k \leftarrow \alpha_k^{\mathrm{v}} \mathbb{E}(\boldsymbol{\Sigma}_k) + (1 - \alpha_k^{\mathrm{v}}) \left(\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T\right) - \boldsymbol{\mu}'_k (\boldsymbol{\mu}'_k)^T. \tag{21}$$

The coefficients α_k^{m} and α_k^{v} are data-dependent and are defined as

$$\alpha_k^{\mathrm{m}} = \frac{\Gamma_k}{\Gamma_k + \beta^{\mathrm{m}}}\,, \quad \alpha_k^{\mathrm{v}} = \frac{\Gamma_k}{\Gamma_k + \beta^{\mathrm{v}}}\,, \tag{22}$$

where β^{m} and β^{v} are related to the hyper parameters [15] and thus should be empirically defined by users. Note that there is no need to update the mixture weights, as they are already occupied by the fixed topic posterior weights.

4.2 Discussion

The MAP-based method is preferable in that we can determine the interpolation factor that balances the contribution between the personal annotations and the background model without loss of model generalizability, as

demonstrated by its superior effectiveness and efficiency in speaker adaptation tasks [44]. If a personal annotation $\{\mathbf{e}_m, \boldsymbol{\theta}_m\}$ is highly correlated to a latent topic z_k (i.e. $p(z_k|\mathbf{e}_m, \boldsymbol{\theta}_m)$ is large), the annotation will contribute more to the update of $\{\boldsymbol{\mu}'_k, \boldsymbol{\Sigma}'_k\}$. In contrast, if the user's annotations have nothing to do with z_h (i.e. the cumulative posterior probability $\Gamma_h = 0$), the parameters of $\{\boldsymbol{\mu}'_h, \boldsymbol{\Sigma}'_h\}$ would remain the same as those of the background model, as shown by the fact that α_k would be 0.

Another advantage of the MAP-based method is that users are free to provide personal annotations for whatever songs they like, such as the songs they are more familiar with. This can help reduce the cognitive load of the personalization process. As the AEG framework is audio-based, the annotated clips can be arbitrary and does not have to be those included in the corpus for training the background model.

Finally, we note that the model adaptation procedure only needs to be performed once, so the algorithm is fairly efficient. It only requires K times of computing the expected sufficient statistics and updating the parameters. In consequence, we can keep refining the background model whenever a small number of personal annotations are available, and readily use the updated model for personalized MER or music retrieval. The model adaptation method for GMM is not limited to the MAP method. We refer interested readers to [6, 7] for more advanced methods.

5 AEG-based Music Emotion Recognition

5.1 Algorithm

As described in Section 3, we predict the emotion distribution of an unseen clip by weighting the affective GMM using the clip's topic posterior $\hat{\boldsymbol{\theta}} = \{\hat{\theta}_k\}_{k=1}^K$ as

$$p(\mathbf{e} \mid \hat{\boldsymbol{\theta}}) = \sum_{k=1}^{K} \hat{\theta}_k G_k(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,. \tag{23}$$

In addition, we can also use a single, representative affective Gaussian $G(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ to summarize the weighted affective GMM. This can be done by solving the following optimization problem:

$$\min_{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}} \quad \sum_{k=1}^{K} \hat{\theta}_k D_{\mathrm{KL}}\big(G_k(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \;\big\|\; G(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})\big)\,, \tag{24}$$

where

$$D_{\mathrm{KL}}(G_A \parallel G_B) = \frac{1}{2}\Big(\mathrm{tr}(\boldsymbol{\Sigma}_A\boldsymbol{\Sigma}_B^{-1}) - \log \mid \boldsymbol{\Sigma}_A\boldsymbol{\Sigma}_B^{-1} \mid + (\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)^T\boldsymbol{\Sigma}_B^{-1}(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B) - 2\Big) \tag{25}$$

denotes the one-way (asymmetric) Kullback–Leibler (KL) divergence (a.k.a. relative entropy) [32] from $G_A(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A)$ to $G_B(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$. This optimization problem is strictly convex in $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$, which means that there is a unique minimizer for the two variables, respectively [11]. Let the partial derivative with respect to $\hat{\boldsymbol{\mu}}$ be 0, we have

$$\sum\nolimits_k \hat{\theta}_k(2\hat{\boldsymbol{\mu}} - 2\boldsymbol{\mu}_k) = 0\,. \tag{26}$$

Given the fact that $\sum_k \hat{\theta}_k = 1$, we derive

$$\hat{\boldsymbol{\mu}} = \sum_{k=1}^{K} \hat{\theta}_k \boldsymbol{\mu}_k\,. \tag{27}$$

Setting the partial derivative with respect to $\boldsymbol{\Sigma}_k^{-1}$ to 0,

$$\sum\nolimits_k \hat{\theta}_k \left(\boldsymbol{\Sigma}_k - \hat{\boldsymbol{\Sigma}} + (\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}})\,(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}})^T\right) = 0\,, \tag{28}$$

we obtain the optimal covariance matrix by,

$$\hat{\boldsymbol{\Sigma}} = \sum_{k=1}^{K} \hat{\theta}_k \left(\boldsymbol{\Sigma}_k + (\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}})\,(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}})^T\right)\,. \tag{29}$$

5.2 Discussion

Representing the predicted result as a single Gaussian is functionally necessary, because it is easier and more straightforward to interpret or visualize the emotion prediction to the users with only a single mean (center) and covariance (uncertainty). However, this may run counter to the theoretical arguments given in favor of a GMM that permits emotion modeling in finer granularity. For instance, it is inadequate for the excerpts whose emotional responses are by nature bi-modal. We note that in applications such as emotion-based music retrieval (cf. Section 6) and music video generation [62], one can directly use the raw weighted GMM (i.e. Eq. 23) as the emotion index of a song in response to queries given in the VA space. We will detail this aspect later in Section 6.

The computation of Eqs. 27 and 29 is quite efficient. The complexity depends mainly on K and the number of frames T of a clip: computing θ_k

requires KT operations (cf. Eq. 2), whereas computing $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ requires K vector multiplications and K matrix operations, respectively. This efficiency is important for dealing with a large-scale music database and for application such as real-time music emotion tracking on a mobile device [24,49,59,60,67].

5.3 Evaluation on General MER

5.3.1 Dataset

We use the AMG1608 dataset [8] for evaluating both general and personalized MER. The dataset contains 1,608 30-second music clips annotated by 665 subjects (345 are male; average age is 32.0±11.4) recruited mostly from the crowdsourcing platform Mechanical Turk [41]. The subjects were asked to rate the VA values that best describe their general (instead of moment-to-moment) emotion perception of each clip via the internet. The VA values, which are real values ranging in between [–1, 1], are entered by clicking on the emotion space on a square interface panel. The subjects were instructed to rate the perceived rather than felt emotion. Each music clip was annotated by 15–32 subjects. Each subject annotated 12–924 clips, and 46 out of the 665 subjects annotated more than 150 music clips, making the dataset a useful corpus for research on MER personalization. The average Krippendorff's α across the music clips is 0.31 for valence and 0.46 for arousal, which are both in the range of fair agreement. Please refer to [8] for more details about this dataset.

5.3.2 Acoustic Features

As different emotion perceptions are usually associated with different patterns of features [17], we use two toolboxes, MIRtoolbox [33] and YAAFE [39], to extract four sets of frame-based features from audio signals, including MFCC-related features, tonal features, spectral features, and temporal features, as listed in Table 1. We down-sample all the audio clips in AMG1608 at 22,050 Hz and normalize them to the same volume level. All the frame-based features are extracted with the same frame size of 50ms and 50% hop size. Each dimension in the frame-based feature vectors is normalized to zero mean and unit standard deviation. We concatenate all the four sets of features for each frame, as this leads to better performance in acoustic modeling in our pilot study [7]. As a result, a frame-level feature vector contains 72 dimensions of features.

However, it does not make sense to analyze and predict the music emotion on a specific frame. Instead of bag-of-frames approach [57,58], we adopt the

Table 1: Frame-based acoustic features used in the evaluation.

Feature	Dim.	Description
MFCCs	40	20 Mel-frequency cepstral coefficients and the first-order time differences [12].
Tonal	17	Octave band signal intensity using a triangular octave filter bank and the ratio of these intensity values [39].
Spectral	11	Linear predictor coefficients that capture the spectral envelope of the audio signal [38], spectral flux, [39] and spectral shape descriptors [42].
Temporal	4	Shape and statistics (centroid, spread, skewness, and kurtosis) [16].
All	72	Concatenation of all the four types of features mentioned above.

bag-of-segments approach for the topic posterior representation, because a segment is able to capture more local temporal variation of the low-level features. Our preliminary result has also confirmed this hypothesis. To generate a segment-level feature vector representing a basic term in the bag-of-segments approach, we concatenate the mean and standard deviation of 16 consecutive frame-level feature vectors, leading to a 144-dimensional vector for a segment. The hop size for a segment is 4 frames. Given the acoustic GMM (cf. Eq. 1), we then follow Eqs. 2 and 3 addressed in Section 3.1 to compute the topic posterior vector of a music clip.

5.3.3 Evaluation Metrics

The accuracy of general MER is evaluated using 3 performance metrics: two-way KL divergence (KL2) [32], Euclidean distance, and R^2 (also known as the coefficient of determination) [54]. The first two measure the distance between the prediction and the ground truth. The lower the value is, the better the performance. KL2 considers the performance with respect to the bivariate Gaussian distribution of a chip, while the Euclidean distance is concerned with the VA mean only. R^2 is also concerned with the VA mean only. In contrast to the distance measure, a high R^2 value is preferred. Moreover, R^2 is computed separately for valence and arousal.

Specifically, we are given the distribution of the ground truth annotations $\mathcal{N}_i = G(\mathbf{a}_i, \mathbf{B}_i)$ (cf. Section 3.2) and the predicted distribution of each test clip $\hat{\mathcal{N}}_i = G(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i)$, both of which are modeled as a bivariate Gaussian distribution, where $i \in \{1, \ldots, N\}$ denotes the index of a clip in the test set. Instead of one-way KL divergence (cf. Eq. 25) for determining the representative Gaussian, we evaluate the performance of emotion distribution

prediction based on the KL2 divergence defined by

$$D_{\mathrm{KL2}}(G_A, G_B) \equiv \frac{1}{2}\Big(D_{\mathrm{KL}}(G_A \parallel G_B) + D_{\mathrm{KL}}(G_B \parallel G_A)\Big). \tag{30}$$

The average KL2 divergence (AKL), which measures the symmetric distance between the predicted emotion distribution and the ground truth one, is computed by $\frac{1}{N}\sum_{i=1}^{N} D_{\mathrm{KL2}}(\mathcal{N}_i, \hat{\mathcal{N}}_i)$. Using the l_2 norm, we can compute the average Euclidean distance (AED) between the mean vectors of two Gaussian distributions by $\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{a}_i - \hat{\boldsymbol{\mu}}\|_2$. The R^2 statistics is a standard way to measure the fitness of regression models [54]. It is used to evaluate the prediction accuracy as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{e}_i - e_i)^2}{\sum_{i=1}^{N}(e_i - \bar{e})^2}, \tag{31}$$

where $\hat{e}_i$ and e_i denote the predicted (either valence or arousal) value and the ground truth one of a clip, respectively, and $\bar{e}$ is the average ground truth value over the test set. When the predictive model perfectly fits the ground truth values, R^2 is equal to 1. If the predictive model does not fit the ground truth well, R^2 may become negative.

We perform three-fold cross-validation to evaluate the performance of general MER. Specifically, the AMG1608 dataset is randomly partitioned into three folds, and an MER model is trained on two of them and tested on the other one. Each round of validation generates the predicted result of one-third of the complete dataset. After three rounds, we will have the predicted result of each clip in the complete dataset. Then, AKL, AED, and the R^2 for valence and arousal are computed over the complete dataset, instead of computing the performance over each one-third of the dataset. This strategy gives an unbiased estimate for R^2.

5.3.4 Result

We compare the performance of AEG with two baseline methods. The first one, referred to as the *base-rate* method, uses a reference affective Gaussian whose mean and covariance are set using the global mean and covariance of the training annotations without taking into account the acoustic features. In other words, the prediction for every test clip would be the same for the base-rate method. The performance of this base-rate method can be considered as a lower bound in this task accordingly. Moreover, we compare the performance of AEG with SVR [51], a representative regression-based approach for predicting emotion values or distributions, using the same type

Table 2: Performance evaluation on general MER (↓ stands for smaller-better and ↑ larger-better).

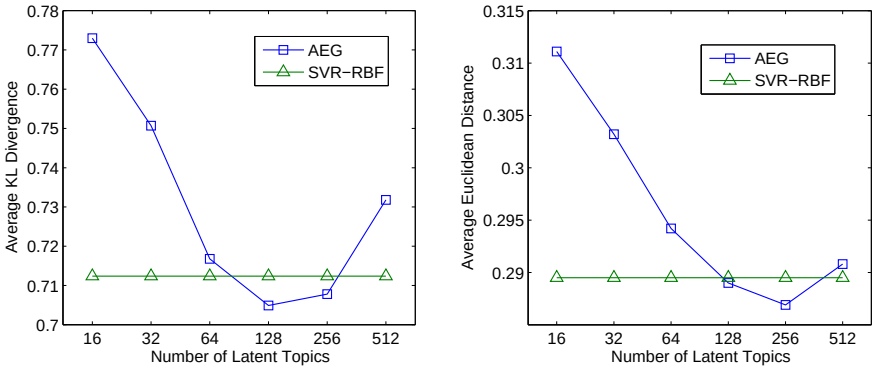
Method	AKL ↓	AED ↓	R^2 Valence ↑	R^2 Arousal ↑
Base-rate	1.2228	0.4052	–0.0009	0.0000
SVR-RBF	0.7124	0.2895	0.1409	0.6613
AEG ($K = 128$)	**0.7049**	0.2890	**0.1601**	0.6554
AEG ($K = 256$)	0.7078	**0.2869**	0.1579	**0.6686**

of acoustic features. Specifically, the feature vector of a clip is formed by concatenating the mean and standard deviation of all the frame-level feature vectors within a clip, yielding a 144-dimensional vector. We use the radial basis function (RBF) kernel SVR implemented by the libSVM library [5], with parameters optimized by grid search with three-fold cross-validation on the training set. We further use a heuristic favorable for SVR to regularize every invalid predicted covariance parameter [65]. This heuristic significantly improves the AKL performance of SVR.

Our pilot study empirically shows that AEG Uniform gives better emotion prediction in AED, compared to AEG AnnoPrior, possibly because the introduction of the annotation prior (cf. Eq. 7) may bias the estimation of the mean parameters in the EM learning. In contrast, AEG AnnoPrior leads to better result in AKL, indicating its capability of estimating a more proper covariance for a learned affective GMM. In light of this, we use a following *hybrid* method to take advantage of both AEG AnnoPrior and AEG Uniform in optimizing the affective GMM. Suppose that we have learned two affective GMMs, one for AEG AnnoPrior and the other for AEG Uniform. To generate a combined affective GMM, for its k-th component Gaussian, we take the mean from the k-th Gaussian of AEG Uniform and the covariance from the k-th Gaussian of AEG AnnoPrior. This combined affective GMM is eventually used to predict the emotion for a test clip with Eqs. 27 and 29 in this evaluation.

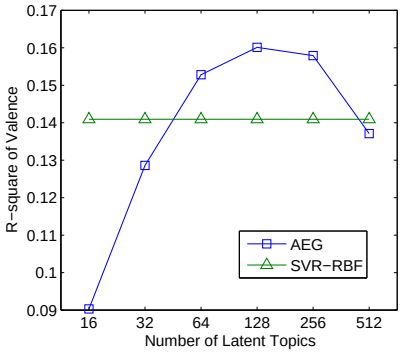
Table 2 compares the performance of AEG with the two baseline methods. It can be seen that both SVR and AEG outperform the base-rate method by a great margin, and that AEG can outperform SVR. For AEG, we can obtain better AKL and better R^2 for valence when $K = 128$, but better AED and better R^2 for arousal when $K = 256$. The best R^2 achieved for valence and arousal are 0.1601 and 0.6686. In particular, the superior performance of AEG in R^2 for valence is remarkable. Such observation suggests AEG a promising approach, as it is typically more difficult to model the valence perception from audio signals [70].

Figure 3 presents the result of AEG when we vary the value of K (i.e.

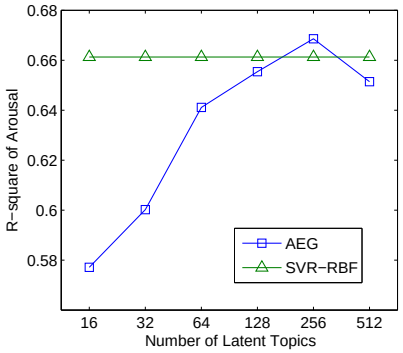


(a) AKL, smaller-better.

(b) AED, smaller-better.



(c) R^2 of valence, larger-better.



(d) R^2 of arousal, larger-better.

Figure 3: Performance evaluation on general MER, using different numbers of latent topics in AEG.

the number of latent topics). It can be seen that the performance of AEG improves as a function of K when K is smaller than 256, but starts to decrease when K is sufficient larger. The best result is obtained when K is set to 128 or 256. As the parameters of SVR-RBF has also been optimized, this result shows that, if the optimal case of AEG is not attained (e.g., $K = 64$ or 512), AEG is still on par with the state-of-the-art SVR approach to general MER.

5.4 Evaluation on Personalized MER

5.4.1 Evaluation Setup

The trade-off between the number of personal annotations (feedbacks) and the performance of personalization is important for personalized MER. On one hand, we hope to have more personal annotations to more accurately

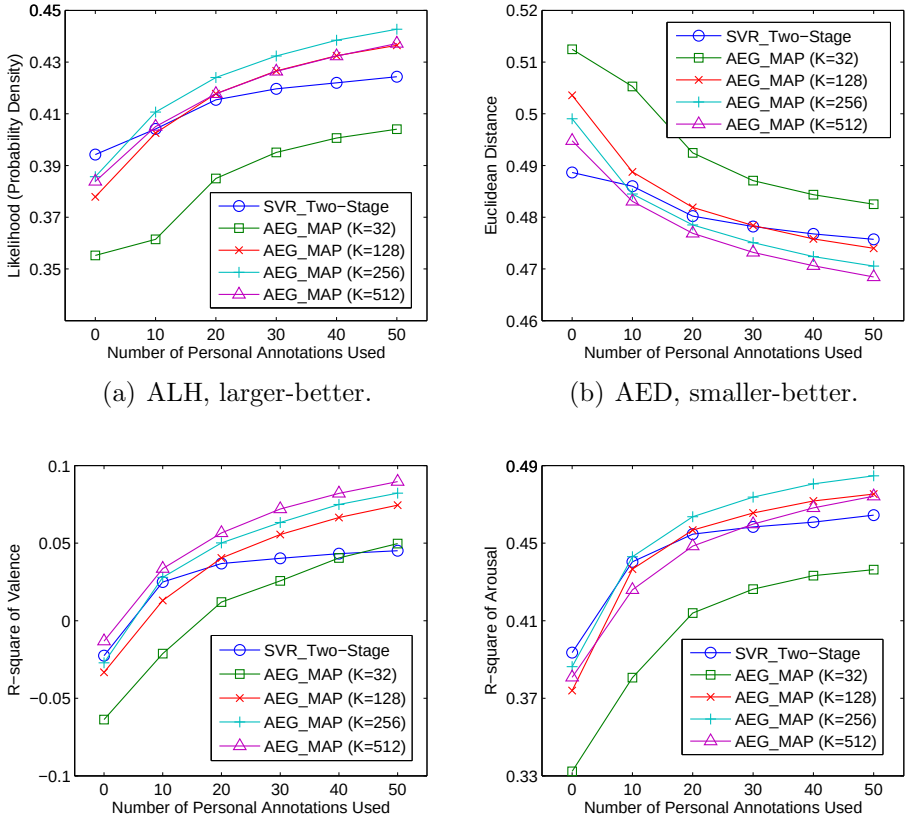
model the emotion perception of a particular user. On the other hand, we want to restrict the number of personal annotations so as to relieve the burden on the user. To reflect this, evaluation on the performance of personalized MER is conducted by fixing the test set for each user, but varying the number of available emotion annotations from the particular user to test how the performance improves as personal data amasses.

We consider 41 users who have annotated more than 150 clips in this evaluation. We use the data of 6 of them for parameter tuning, and the data of the remaining 35 in the evaluation and report the average result for these 35 test users. One hundred annotations of each test user are randomly selected as the personalized training set for personalization for the user. Once the model is created, another 50 clips annotated by the same user are randomly selected. Specifically, for each test user, a general MER model is trained with 600 clips randomly selected from the original AMG1608, excluding those annotated by the test user under consideration and those self-inconsistent annotations. Then, the general model is incrementally personalized five times using different numbers of clips selected from the personalized training set. We use 10, 20, 30, 40, and 50 clips iteratively, with the preceding clips being a subset of the current ones each time. The process is repeated 10 times for each user.

We use the following evaluation metrics here: the AED, the R^2, and the average likelihood (ALH) of generating the ground-truth annotation (a single VA point) $\mathbf{e}_\star$ of the test user using the predicted affective Gaussian, i.e. $p(\mathbf{e}_\star \mid \hat{\boldsymbol{\mu}}_\star, \hat{\boldsymbol{\Sigma}}_\star)$. Larger ALH corresponds to better accuracy. We do not report KL divergence here because each clip in the dataset is annotated by a user at most once, which does not constitute a probability distribution.

5.4.2 Result

We compare the MAP-based personalization method of AEG with the two-stage personalization method of SVR proposed in [78]. In the two-stage SVR method, the first stage creates a general SVR model for general emotion prediction, whereas the second stage creates a personalized SVR that is trained solely on a user's annotations. The final prediction is obtained by linearly combining the predictions from the general SVR and the personalized SVR with weights 0.7 and 0.3, respectively. The weights are derived empirically according to our pilot study. As for AEG, we only update the mean parameters with $\beta^{\mathrm{m}} = 0.01$, because our pilot study shows that updating the covariance empirically does not lead to better performance. This observation is also in line with the findings in speaker adaptation [44]. We train the background model with AEG Uniform for simplicity.

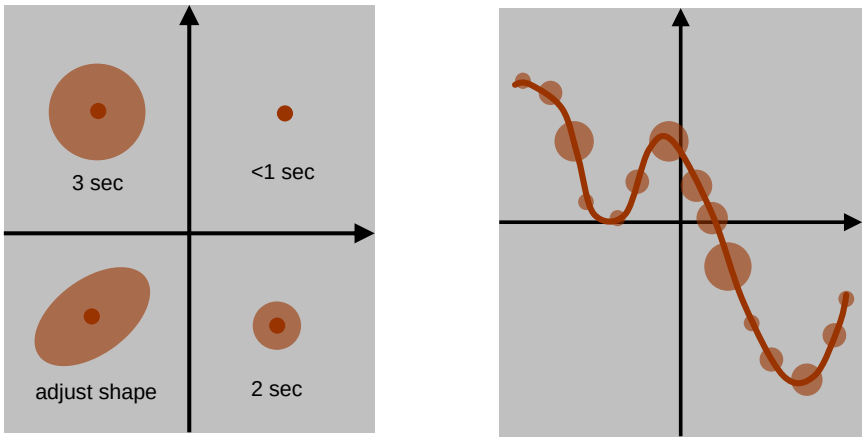


(a) ALH, larger-better. (b) AED, smaller-better.

(c) R^2 of valence, larger-better. (d) R^2 of arousal, larger-better.

Figure 4: Performance evaluation on personalized MER, with varying numbers of personal data.

Figure 4 compares the result of different personalized MER methods, when we vary the number of available personal annotations. The starting point of each curve is the result given by the general MER model trained on partial users of the AMG1608 dataset. We can see that the result of the general model is inferior to those reported in Figure 3, showing that a general MER model is less effective when it is used to predict the emotion perception of individual users, compared to the case of predicting the *average* emotion perception of users. We can also see that the result of the considered personalized methods generally grows as the number of personal annotations increases. When the value of K is sufficiently large, AEG-based personalization methods can outperform the SVR method. Moreover, while the result of SVR starts to saturate when the number of personal annotations is larger than 20, AEG has the potential of keeping on improving the perfor-



(a) Point-based (b) Trajectory-based

Figure 5: The stress-sensitive user interface for emotion-based music retrieval. Users can (a) specify a point or (b) draw a trajectory, while specifying the variance with different levels of duration.

mance by exploiting more personal annotations. We also note that there is no significant performance difference for AEG when K is large enough (e.g. ≥ 128).

Although our evaluation shows that personalization methods can improve the result of personalized emotion prediction, the low values in the R^2 statistics for valence and arousal still show that the problem is fairly challenging. Future work is still needed to improve either the quality of the emotion annotation data or the feature extraction or machine learning algorithms for modeling emotion perception.

6 Emotion-based Music Retrieval

6.1 The VA-oriented Query Interface

The VA space offers a ready canvas for music retrieval through the specification of a point in the emotion space [75]. Users can retrieve music pieces of certain emotions without specifying the titles. Users can also draw a trajectory to indicate the desired emotion changes across a list of songs (e.g. from angry to tender).

In addition to the above point-based query, one can also issue a Gaussian-based query to an AEG-based retrieval system. As Figure 5 shows, users can specify the desired variances (or the confidence level at the center point) of

Table 3: The two approaches of the emotion-based music retrieval system

Approach	Indexing phase	Indexed type	Matching phase
Emotion Prediction	full procedure of MER by AEG	an affective GMM (Eq. 23) or a 2-dim Gaussian $\{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\}$	likelihood (for point query) or distance (for Gaussian query)
Folding-In	compute only the topic posterior	K-dim vector $\hat{\boldsymbol{\theta}}$	cosine similarity of pseudo song (K-dim vector $\boldsymbol{\lambda}$)

emotion by pressing a point in the VA space with different levels of duration or strength. The variance of the Gaussian gets smaller as one increases the duration or strength of pressing, as Figure 5 (a) shows. Larger variances indicate less specific emotion around the center point. After specifying the size of a circular variance shape, one can even pinch fingers to adjust the variance shape. For a trajectory-based query input, similarly, the corresponding variances are determined according to the dynamic speed when drawing the trajectory, as Figure 5 (b) shows. Fast speed corresponds to a less specific query and the system will return pieces whose variances of emotion are larger. If songs with more specific emotions are desirable, one can slow down the speed when drawing the trajectory. The queries inputted by such a *stress-sensitive interface* can be handled by AEG for emotion-based music retrieval.

6.2 Overview of the Emotion-based Music Retrieval System

As Figure 6 shows, the content-based retrieval system can be divided into two phases. In the *feature indexing* phase, we index each music clip in an unlabeled music database by one of the following two approaches: The *emotion prediction* approach indexes a clip with the *predicted emotion distribution* (an affective GMM or a single 2-D Gaussian) given by MER, whereas the *folding-in* approach indexes a clip with the *topic posterior* (a K-dimensional vector). In the later *music retrieval* phase, given an arbitrary emotion-oriented query the system returns a list of music clips ranked according to one of the following two approaches: *likelihood/distance-based matching* and *pseudo song-based matching*. These two ranking approaches correspond to one of the two indexing approaches, respectively, as summarized in Table 3. We present the details of the two approaches in the following subsections.

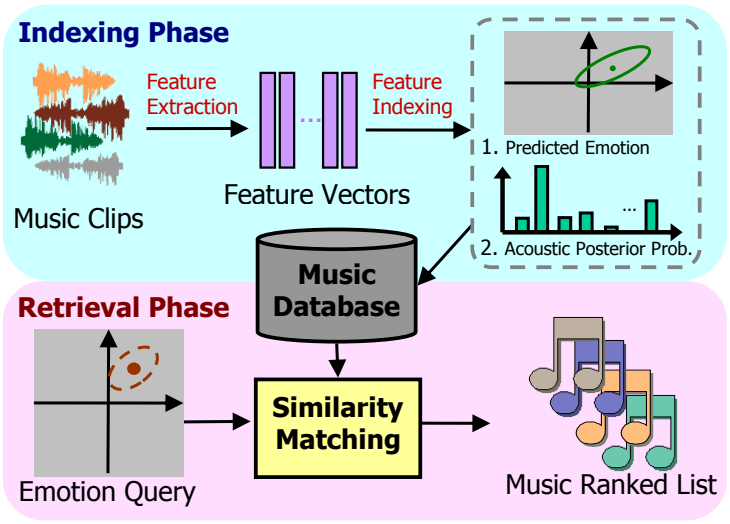


Figure 6: The diagram of the content-based music retrieval system using an emotion query.

6.3 The Emotion Prediction-based Approach

This approach indexes each clip as a single, representative Gaussian distribution or an affective GMM in the offline MER procedure. The query is then used to compare with the predicted emotion distribution of each clip in the database. The system ranks all the clips based on the likelihoods or distances in response to the query. Clips with larger likelihood or smaller distance should be placed in the higher order.

Given a point query $\tilde{\mathbf{e}}$, the corresponding likelihood of the indexed emotion distribution of a clip $\hat{\boldsymbol{\theta}}_i$ is generated by a single Gaussian $p(\tilde{\mathbf{e}} \mid \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i)$ or an affective GMM $p(\tilde{\mathbf{e}} \mid \hat{\boldsymbol{\theta}}_i)$ (cf. Eq. 23), where $\{\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i\}$ is the predicted parameters of the representation Gaussian for $\hat{\boldsymbol{\theta}}_i$, and $\hat{\theta}_{i,k}$ is the k-th component of $\hat{\boldsymbol{\theta}}_i$. Note that here we use the topic posterior vector to represent a clip in the database.

When it comes to a Gaussian-based query $\tilde{G} = G(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, the approach generates the ranking scores based on the KL2 divergence. In the case of indexing with a single Gaussian, we use Eq. 30 to compute $D_{\mathrm{KL2}}\big(\tilde{G}, G(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i)\big)$ between the query and a clip. On the other hand, in the case of indexing with an affective GMM, we compute the weighted KL2 divergence by

$$D_{\mathrm{KL2}}\big(\tilde{G}, p(\mathbf{e} \mid \hat{\boldsymbol{\theta}}_i)\big) = \sum_{k=1}^{K} \hat{\theta}_{i,k} D_{\mathrm{KL2}}\big(\tilde{G}, G_k(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\big)\,. \tag{32}$$

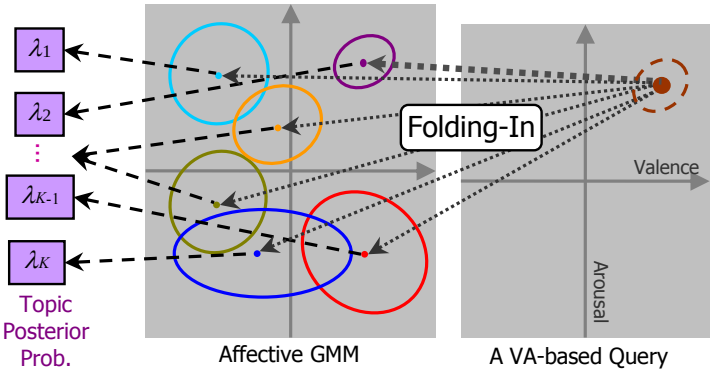


Figure 7: Illustration of the Folding-In process of emotion-based music retrieval by AEG.

6.4 The Folding-In-based Approach

As Figure 7 shows, this approach estimates the probability distribution $\boldsymbol{\lambda} = \{\lambda_k\}_{k=1}^{K}$, subject to $\sum_k \lambda_k = 1$, for an input VA-oriented query in an online manner. Each estimated λ_k corresponds to the relevance of a query to the k-th latent topic z_k, so we can treat the distribution of $\boldsymbol{\lambda}$ as the topic posterior of the query and call it a *pseudo song*. In the case of Figure 7, for example, we show a query that is very likely to be represented by the 2-nd affective Gaussian component. The folding-in process is likely to assign a dominative weight $\lambda_2 = 1$ for z_2, and $\lambda_h = 0$, $\forall h \neq 2$. This implies that the query is highly related to the song whose topic posterior is dominated by θ_2. Therefore, the pseudo song can be used to match with the topic posterior vector $\hat{\boldsymbol{\theta}}_i$ of each clip in the database.

Given a point query $\tilde{\mathbf{e}}$, we start the folding-in process by first generating the pseudo song via maximizing the query likelihood of the $\boldsymbol{\lambda}$-weighted affective GMM with respective to $\boldsymbol{\lambda}$. By taking the logarithm of Eq. 23, we obtain the following objective function,

$$\max_{\boldsymbol{\lambda}} \quad \log \sum_{k=1}^{K} \lambda_k \; G_k(\tilde{\mathbf{e}} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,, \tag{33}$$

where λ_k is the k-th component of the vector $\boldsymbol{\lambda}$. In some sense, a good $\boldsymbol{\lambda}$ will make the corresponding $\boldsymbol{\lambda}$-weighted affective GMM well generate the query $\tilde{\mathbf{e}}$. The problem in Eq. 33 can be solved by the EM algorithm. In the E-step, the posterior probability of z_k is computed by

$$p(z_k \mid \tilde{\mathbf{e}}) = \frac{\lambda_k G_k(\tilde{\mathbf{e}} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{h=1}^{K} \lambda_h G_h(\tilde{\mathbf{e}} \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}\,. \tag{34}$$

In the M-step, we then only update λ_k by

$$\lambda'_k \leftarrow p(z_k \mid \tilde{\mathbf{e}})\,. \tag{35}$$

As for a Gaussian-based query $\tilde{G}$, we fold in the query into the learned affective GMM to estimate a pseudo song as well. This time, we maximize the following log-likelihood function,

$$\max_{\boldsymbol{\lambda}} \quad \log \sum_{k=1}^{K} \lambda_k \; p(\tilde{G} \mid G_k)\,, \tag{36}$$

where $p(\tilde{G} \mid G_k)$ is the likelihood function based on KL2 (cf. Eq. 30):

$$p(\tilde{G} \mid G_k) = \exp\big(-D_{\mathrm{KL2}}(\tilde{G}, G_k)\big)\,. \tag{37}$$

Again, Eq. 36 can be solved by the EM algorithm, with the following update,

$$\lambda'_k \leftarrow p(z_k \mid \tilde{G}) = \frac{\lambda_k p(\tilde{G} \mid G_k)}{\sum_{h=1}^{K} \lambda_h p(\tilde{G} \mid G_h)}\,. \tag{38}$$

The EM processes for both point- and Gaussian-based queries stop early after few iterations (e.g. 3), because the pseudo song estimation is sensitive to over-fitting. Several initialization settings can be used, such as a random, uniform, or prior distribution. Considering the stability and the reproducibility of the experimental result, we opt for using a uniform distribution for initialization. Note that random initialization may introduce discrepant results among different trials even with identical experimental settings, whereas initializing with a prior distribution may render biased results in favor of songs that predominates the training data [63]. Finally, the retrieval system ranks all the clips in descending order of the following cosine similarities in response to the pseudo song:

$$\Phi(\boldsymbol{\lambda}, \boldsymbol{\theta}_i) = \frac{\boldsymbol{\lambda}^T \boldsymbol{\theta}_i}{\|\boldsymbol{\lambda}\| \|\boldsymbol{\theta}_i\|}\,. \tag{39}$$

6.5 Discussion

The Emotion Prediction approach is straightforward, as the purpose of MER is to automatically index unseen music pieces in the database. In contrast, the Folding-In approach goes one step further to embed a VA-based query into the space of music clips. Although the folding-in process requires an additional step of estimating the pseudo song, it is in fact more flexible. In

a personalized music retrieval context, for example, a personalized affective GMM can readily produce a personalized pseudo song for comparing with the original topic posterior vectors of all the pieces in the database, without the need to predict the emotion again with the personalized model.

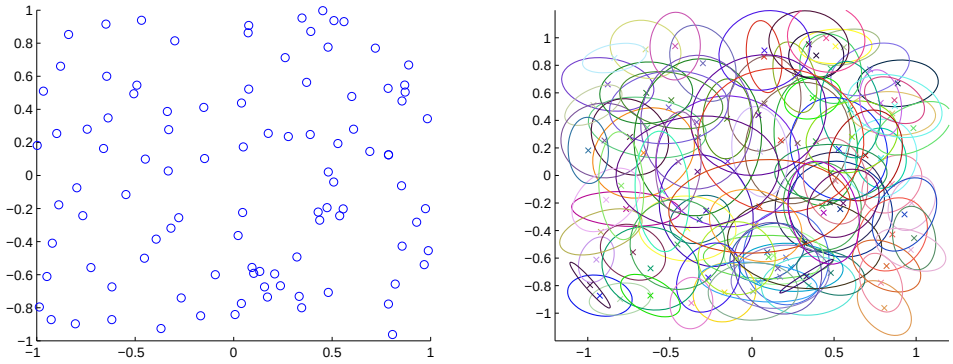
The complexity of the Emotion Prediction approach mainly comes from computing the likelihood of a point query on each music clip's emotion distribution or the KL divergence between the Gaussian query and the emotion distribution of each clip. Therefore, the matching process needs to compute N (the number of clips in the database) times the likelihood or the KL divergence. In the Folding-In approach, the complexity comes from estimating the pseudo song (with the EM algorithm) and computing the cosine similarity between the pseudo song and each clip. EM needs to compute $K \times ITER$ times the likelihood of a component affective Gaussian or the Gaussian KL divergence, where $ITER$ is the number of EM iterations. Then, the matching process computes N times the cosine similarity. Obviously, computing the likelihood on an emotion distribution (i.e. a single Gaussian or a GMM) is computationally more expensive than computing the cosine similarity (as K is usually not large). Therefore, when N is large (e.g. $N \gg K \times ITER$), the Folding-In approach is considered as a more feasible one in practice.

6.6 Evaluation for Emotion-based Music Retrieval

6.6.1 Evaluation Setup

The AMG1608 dataset is again adopted in this music retrieval evaluation. We consider two emotion-based music retrieval scenarios: query-by-point and query-by-Gaussian. For each scenario, we create a set of synthetic queries and use the learned AEG model to respond to each test query and return a ranked list of music clips from an unlabeled music database. The generation of the test query set for query-by-point is simple. As Figure 8 (a) shows, we uniformly sample 100 2-D query points within $\left[[-1,-1]^T, [1,1]^T\right]$ in the VA space. The test query set for query-by-Gaussian is then based on this set of points. Specifically, we convert a point query to a Gaussian query by associating with the point a 2-by-2 covariance matrix, as Figure 8 (b) shows. Motivated by our empirical observation from data, the covariance of a Gaussian query is set in inverse proportion to the distance between the mean of the Gaussian query (determined by the corresponding point query) and the origin of the VA space. That is, if a given point query is far from the origin (with large emotion magnitude), the user may want to retrieve songs with a specific emotion (with smaller covariance ellipse).

The performance is evaluated by aggregating the ground truth *relevance*



(a) (b)

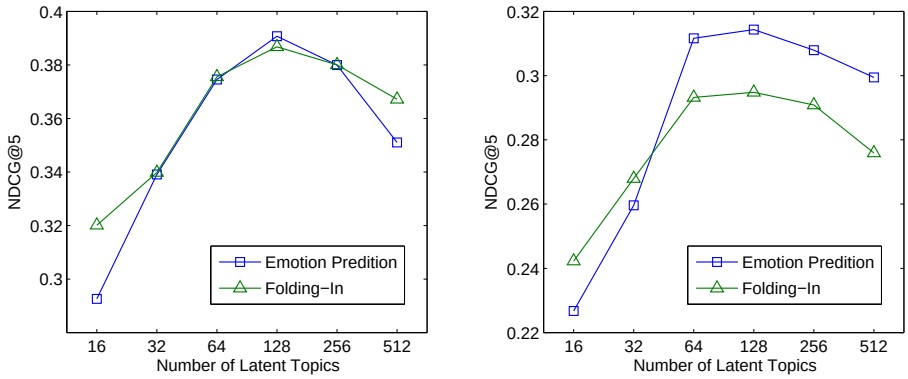
Figure 8: Test queries used in evaluating emotion-based music retrieval: (a) 100 points generated uniformly in between [–1, 1]. (b) 100 Gaussians generated based on the previous 100 points.

scores of the retrieved music clips according to the normalized discounted cumulative gain (NDCG), a widely used performance measure for ranking problems [25]. The NDCG@P, which measures the relevance of the top P retrieved clips for a query, is computed by

$$\text{NDCG@}P = \frac{1}{Z_P}\left\{R(1) + \sum_{i=2}^{P}\frac{R(i)}{\log_2 i}\right\}, \tag{40}$$

where $R(i)$ is the ground truth relevance score of the rank-i clip, $i = 1, \ldots, Q$, where $Q \geq P$ is the number of clips in the music database, and Z_P is the normalization term that ensures the ideal NDCG@P equal 1. Let $\mathcal{N}_i$ (with parameters $\{\mathbf{a}_i, \mathbf{B}_i\}$) denote the ground-truth annotation Gaussian of the rank-i clip. For a point query $\tilde{\mathbf{e}}$, $R(i)$ is obtained by $p(\tilde{\mathbf{e}} \mid \mathbf{a}_i, \mathbf{B}_i)$, the likelihood of the query point. For a Gaussian query $\tilde{\mathcal{N}}$, $R(i)$ is given by $p(\tilde{\mathcal{N}} \mid \mathcal{N}_i)$ defined by Eq. 37. From Eq. 40, we see that if the system ranks the clips in similar order as the descending order obtained on $\{R(i)\}_{i=1}^{Q}$, we obtain a larger NDCG. We report the average NDCG computed over the test query set. Note that we do not adopt evaluation metrics, such as the mean average precision (MAP) and the area under the ROC curve (AUC), because currently it is not trivial to set a threshold to binarize $R(i)$.

We perform three-fold cross-validation as that used in evaluating general MER. In each round, the test fold (with 536 clips) serves as the unlabeled music database.



(a) Point-based query, larger-better. (b) Gaussian-based query, larger-better.

Figure 9: Evaluation result of emotion-based music retrieval.

6.6.2 Result

We implement a Random approach to reflect the lower bound performance by using a random permutation for each test query, without taking into consideration any ranking approach. We further implement an *Ensemble* approach that averages the rankings of a test query given by Emotion Prediction and Folding-In. Specifically, both approaches assign an ordinal number to a clip according to their respective rankings. Then, we average the two ordinal numbers of a clip as a new score, and re-rank all the clips in ascending order of their new scores.

Note that we only consider AEG Uniform for simplicity in the result presentation. Our preliminary study reveals that AEG Uniform in general perform slightly better than AEG AnnoPrior and the hybrid method mentioned in Section 5.3.4 in the retrieval task. Moreover, for the Folding-In approach, early stop is not only important to the folding-in process, but also necessary to learning the affective GMM. According to our pilot study, setting $ITER = 2 - 4$ for learning affective GMM and $ITER = 3$ for learning the pseudo song lead to the optimal performance.

Figures. 9 (a) and (b) compare the NDCG@5 of the Emotion Prediction and Folding-In approaches to emotion-based music retrieval using either point-based or Gaussian-based queries. We are interested in how the result changes as we vary the number of latent topics. It can be found that the two approaches perform very similarly for point-based query when K is in between 64 and 256. Moreover, we see that Emotion Prediction can outperform Folding-In for Gaussian-based query when K is sufficiently large ($K \geq 64$). The optimal model is attained when $K = 128$ in all cases. Similar to the

Table 4: The query-by-point retrieval performance in terms of NDCG@5, 10, 20, and 30.

Method	$P = 5$	$P = 10$	$P = 20$	$P = 30$
Random	0.1398	0.1504	0.1666	0.1804
Emotion Prediction	0.3907	0.4027	0.4288	0.4490
Folding-In	0.3868	0.4067	0.4333	0.4533
Ensemble	0.3954	0.4129	0.4398	0.4601

Table 5: The query-by-Gaussian retrieval performance in terms of NDCG@5, 10, 20, and 30.

Method	$P = 5$	$P = 10$	$P = 20$	$P = 30$
Random	0.1032	0.1090	0.1185	0.1272
Emotion Prediction	0.3143	0.3306	0.3481	0.3658
Folding-In	0.2932	0.3147	0.3383	0.3532
Ensemble	0.3204	0.3368	0.3601	0.3783

result in General MER, it seems that setting K either too large or too small would lead to sub-optimal result.

Tables 4 and 5 present the result of NDCG@5, 10, 20, and 30 for different retrieval methods, including the random baseline, Emotion Prediction, Folding-In, and the Ensemble approaches. The latter three use AEG Uniform with $K = 128$. It is obvious that the latter three can significantly outperform the random baseline, demonstrating the effectiveness of AEG in emotion-based music retrieval. It can also be found that the Ensemble approach leads to the best result.

A closer comparison between Emotion Prediction and Folding-In for point-based query shows nip and tuck, whereas the former performs consistently better regardless of the value of P for Gaussian-based query. Moreover, the NDCG measure seems more favorable for point-based query than Gaussian-based one. Our observation indicates that the standard deviation of the ground truth relevance scores (i.e. $\{R(i)\}_{i=1}^{Q}$) for Gaussian-based query is much larger, resulting in a more challenging measurement basis than that for point-based query. However, the relative performance difference between the two methods is similar for point-based and Gaussian-based queries.

7 Conclusion

AEG is a principled probabilistic framework that nicely unifies the computation processes for MER and emotion-based music retrieval for dimensional

emotion representations such as valence and arousal. Moreover, AEG better takes into account the subjective nature of music emotional responses through the use of probabilistic inference and model adaptation, further making it possible to personalize an emotion-based MIR system. The codes for implementing AEG can be retrieved from the link below: `http://slam.iis.sinica.edu.tw/demo/AEG/`.

Despite that AEG is a powerful approach, there remains a number of challenges for MER, including:

- Is it the best way to consider the valence-arousal space as a coordinate space (with two orthogonal axes)?
- How do we define the "intensity" of emotion? Does the magnitude of a point in the emotion space implies intensity? Would it be possible to train regressors that treat the emotion space as a polar coordinate?
- What are the features that are more important for modeling emotion?
- Cross genre generazability [13].
- Cross culture generazability [21].
- How to incorporate lyrics features for MER?
- How to model the effect of the singing voice in emotion perception?
- How do findings in MER help emotion-based music synthesis or manipulation?

We note that AEG is only suitable for an emotion-based MIR system when we characterize emotions in terms of valence and arousal. It does not apply to systems that use categorical mood tags to describe emotion. A corresponding probabilistic model for categorical MER is yet to be developed. More research efforts are also needed for the personalization and retrieval aspects for categorical MER.

The AEG model itself can also be improved in a number of directions. For example, there are several alternative methods that one can adopt to enhance the latent acoustic descriptors (i.e. $\{A_k\}_{k=1}^K$ in Section 3) for clip-level topic poster representation, such as deep learning [50] or sparse representations [57]. One can also perform discriminative training to reduce the prediction error by using the same corpus with respect to the selection of Gaussian components or parameter refinement over the affective GMM. For example, a stacked discriminative learning on the parameters initialized by a EM-learned generative model has been studied for years in speech recognition [9, 26].

Following this research line, it may help improve AEG as well. Finally, the AEG framework can be easily extended to include multi-modal content such as lyrics, review comments, album cover, and music video. For instance, given a silent video sequence, one can accompany it with a piece of music based on music emotion [62].

8 Acknowledgement

This work was supported by the Academia Sinica–UCSD Postdoctoral Fellowship to Ju-Chiang Wang, the Academia Sinica Career Development Program to Yi-Hsuan Yang, and the Ministry of Science and Technology of Taiwan under Grants NSC 101-2221-E- 001-019-MY3, NSC 102-2221-E-001-004-MY3, and NSC 102-2221-E-001-008-MY3.

References

[1] M. Barthet, G. Fazekas, and M. Sandler. Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models. In *Proc. Int. Symp. Computer Music Modeling and Retrieval*, pages 492–507, 2012.

[2] E. Bigand, S. Vieillard, F. Madurell, J. Marozeau, and A. Dacquet. Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition and Emotion*, 19(8):1113–1139, 2005.

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

[4] L. Bottou. Online algorithms and stochastic approximations. In D. Saad, editor, *Online Learning and Neural Networks*. Cambridge University Press, Cambridge, UK, 1998.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intelligent System and Technology*, 2(3):27:1–27:39, 2011.

[6] 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. Acoustics, Speech, and Signal Processing*, pages 2149–2153, 2014.

[7] Y.-A. Chen, J.-C. Wang, Y.-H. Yang, and H. H. Chen. Personalization of music emotion recognition by mixture model adaptation. *IEEE Trans. Audio, Speech, and Language Processing*, 2015. Submitted.

[8] Y.-A. Chen, Y.-H. Yang, J.-C. Wang, and H. H. Chen. The AMG1608 dataset for music emotion recognition. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2015. [Online] `http://amg1608.blogspot.tw/`.

[9] W. Chou. Minimum classification error approach in pattern recognition. In W. Chou and B.-H. Juang, editors, *Pattern Recognition in Speech and Language Processing*. CRC Press, 2003.

[10] G. Collier. Beyond valence and activity in the emotional connotations of music. *Psychology of Music*, 35(1):110–131, 2007.

[11] J. V. Davis and I. S. Dhillon. Differential entropic clustering of multivariate Gaussians. In *Advances in Neural Information Processing Systems*, volume 19, pages 337–344, 2007.

[12] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

[13] T. Eerola. Modelling emotions in music: Advances in conceptual, contextual and validity issues. In *Proc. AES Int. Conf.*, 2014.

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

[15] J. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chains. *IEEE Trans. Speech and Audio Processing*, 2:291–298, 1994.

[16] O. Gillet and G. Richard. Automatic transcription of drum loops. In *Proc. IEEE Int. Conf. Acoutstics, Speech, and Signal Processing*, pages 269–272, 2004.

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

[18] K. Hevner. Expression in music: A discussion of experimental studies and theories. *Psychological Review*, 48(2):186–204, 1935.

[19] X. Hu and J. S. Downie. When lyrics outperform audio for music mood classification: A feature analysis. In *Proc. Int. Soc. Music Information Retrieval Conf.*, pages 619–624, 2010.

[20] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann. The 2007 MIREX audio mood classification task: Lessons learned. In *Proc. Int. Soc. Music Information Retrieval Conf.*, pages 462–467, 2008.

[21] X. Hu and Y.-H. Yang. A study on cross-cultural and cross-dataset generalizability of music mood regression models. In *Proc. Sound and Music Computing Conf.*, 2014.

[22] A. Huq, J. P. Bello, and R. Rowe. Automated music emotion recognition: A systematic evaluation. *J. New Music Research*, 39(3):227–244, 2010.

[23] D. Huron. *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press, Cambridge, Massachusetts, 2006.

[24] V. Imbrasaite, T. Baltrusaitis, and P. Robinson. Emotion tracking in music using continuous conditional random fields and relative feature representation. In *Proc. Int. Works. Affective Analysis in Multimedia*, 2013.

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

[26] B.-H. Juang, W. Chou, and C.-H. Lee. Minimum classification error rate methods for speech recognition. *IEEE Trans. Speech and Audio Processing*, 5(3):257–265, 1997.

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

[28] P. N. Juslin. Cue utilization in communication of emotion in music performance: Relating performance to perception. *J. Experimental Psychology: Human Perception and Performance*, 16(6):1797–1813, 2000.

[29] P. N. Juslin and J. A. Sloboda. *Music and Emotion: Theory and Research*. Oxford University Press, New York, 2001.

[30] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. J. Scott, J. A. Speck, and D. Turnbull. Music emotion recognition: A state of the art review. In *Proc. Int. Soc. Music Information Retrieval Conf.*, pages 255–266, 2010.

[31] 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.

[32] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22(1):79–86, 1951.

[33] O. Lartillot and P. Toiviainen. A matlab toolbox for musical feature extraction from audio. In *Proc. Int. Conf. Digital Audio Effects*, pages 237–244, 2007.

[34] A. J. Lonsdale and A. C. North. Why do we listen to music? a uses and gratifications analysis. *British Journal of Psychology*, 102:108–134, 2011.

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

[36] 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 Research*, 36(4):281–299, 2007.

[37] J. Madsen, B. S. Jensen, and J. Larsen. Modeling temporal structure in music for emotion prediction using pairwise comparisons. In *Proc. Int. Soc. Music Information Retrieval Conf.*, pages 319–324, 2014.

[38] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of IEEE*, 63(4):561–580, 1975.

[39] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard. YAAFE, an easy to use and efficient audio feature extraction software. In *Proc. Int. Soc. Music Inform. Retrieval Conf.*, pages 441–446, 2010.

[40] R. Panda, B. Rocha, and R. P. Paiva. Dimensional music emotion recognition: Combining standard and melodic audio features. In *Proc. Int. Symp. Computer Music Modeling and Retrieval*, 2013.

[41] G. Paolacci, J. Chandler, and P. Ipeirotis. Running experiments on Amazon Mechanical Turk. *Judgment Decision Making*, 5(5):411–419, 2010.

[42] G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Technical report, IRCAM, Paris, France, 2004.

[43] 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.

[44] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

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

[46] P. Saari and T. Eerola. Semantic computing of moods based on tags in social media of music. *IEEE Trans. Knowledge and Data Engineering*, 26(10):2548–2560, 2014.

[47] P. Saari, T. Eerola, G. Fazekasy, M. Barthet, O. Lartillot, and M. Sandler. The role of audio and tags in music mood prediction: A study using semantic layer projection. In *Proc. Int. Soc. Music Information Retrieval Conf.*, pages 201–206, 2013.

[48] E. M. Schmidt and Y. E. Kim. Prediction of time-varying musical mood distributions from audio. In *Proc. Int. Soc. Music Information Retrieval Conf.*, pages 465–470, 2010.

[49] E. M. Schmidt and Y. E. Kim. Modeling musical emotion dynamics with conditional random fields. In *Proc. Int. Soc. Music Information Retrieval Conf.*, pages 777–782, 2011.

[50] E. M. Schmidt and Y. E. Kim. Learning rhythm and melody features with deep belief networks. In *Proc. Int. Soc. Music Information Retrieval Conf.*, pages 21–26, 2013.

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

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

[53] 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 Research*, 39(1):13–34, 2010.

[54] A. Sen and M. S. Srivastava. *Regression Analysis: Theory, Methods, and Applications*. Springer Science & Business Media, 1990.

[55] M. Soleymani, A. Aljanaki, Y.-H. Yang, M. N. Caro, F. Eyben, K. Markov, B. Schuller, R. Veltkamp, F. Weninger, and F. Wiering. Emotional analysis of music: A comparison of methods. In *Proc. ACM Multimedia*, pages 1161–1164, 2014.

[56] M. Soleymani, M. N. Caro, E. Schmidt, C.-Y. Sha, and Y.-H. Yang. 1000 songs for emotional analysis of music. In *Proc. Int. Work. Crowdsourcing for Multimedia*, pages 1–6, 2013.

[57] L. Su, C.-C. M. Yeh, J.-Y. Liu, J.-C. Wang, and Y.-H. Yang. A systematic evaluation of the bag-of-frames representation for music information retrieval. *IEEE Trans. Multimedia*, 16(5):1188–1200, 2014.

[58] 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 *Proc. Int. Soc. Music Information Retrieval Conf.*, pages 85–90, 2011.

[59] J.-C. Wang, H.-M. Wang, and S.-K. Jeng. Playing with tagging: A real-time tagging music player. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pages 77–80, 2012.

[60] J.-C. Wang, H.-M. Wang, and G. Lanckriet. A histogram density modeling approach to music emotion recognition. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2015.

[61] J.-C. Wang, Y.-H. Yang, K. Chang, H.-M. Wang, and S.-K. Jeng. Exploring the relationship between categorical and dimensional emotion semantics of music. In *Proc. ACM Int. Works. Music information retrieval with user-centered and multimodal strategies*, pages 63–68, 2012.

[62] J.-C. Wang, Y.-H. Yang, I. Jhuo, Y.-Y. Lin, and H.-M. Wang. The acousticvisual emotion Gaussians model for automatic generation of music video. In *Proc. ACM Multimedia*, pages 1379–1380, 2012.

[63] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng. The acoustic emotion Gaussians model for emotion-based music annotation and retrieval. In *Proc. ACM Multimedia*, pages 89–98, 2012.

[64] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng. Personalized music emotion recognition via model adaptation. In *Proc. APSIPA Annual Summit & Conference*, 2012.

[65] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng. Modeling the affective content of music with a Gaussian mixture model. *IEEE Trans. Affective Computing*, 2015. in press.

[66] M.-Y. Wang, N.-Y. Zhang, and H.-C. Zhu. User-adaptive music emotion recognition. In *Proc. IEEE Int. Conf. Signal Processing*, pages 1352–1355, 2004.

[67] S.-Y. Wang, J.-C. Wang, Y.-H. Yang, and H.-M. Wang. Towards time-varying music auto-tagging based on CAL500 expansion. In *Proc. IEEE Int. Conf. Multimedia and Expo*, pages 1–6, 2014.

[68] X. Wang, Y. Wu, X. Chen, and D. Yang. A two-layer model for music pleasure regression. In *Proc. Int. Works. Affective Analysis in Multimedia*, 2013.

[69] F. Weninger, F. Eyben, and B. Schuller. On-line continuous-time music mood regression with deep recurrent neural networks. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pages 5449–5453, 2014.

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

[71] 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 Language Processing*, 19(7):2184–2196, 2011.

[72] Y.-H. Yang and H. H. Chen. Ranking-based emotion recognition for music organization and retrieval. *IEEE Trans. Audio, Speech, and Language Processing*, 19(4):762–774, 2011.

[73] Y.-H. Yang and H.-H. Chen. Machine recognition of music emotion: A review. *ACM Trans. Intelligent Systems and Technology*, 3(4), 2012.

[74] Y.-H. Yang, Y.-C. Lin, and H. H. Chen. Personalized music emotion recognition. In *Proc. ACM SIGIR Int. Conf. Research and Development in Information Retrieval*, pages 748–749, 2009.

[75] Y.-H. Yang, Y.-C. Lin, H.-T. Cheng, and H. H. Chen. Mr. Emo: Music retrieval in the emotion plane. In *Proc. ACM Multimedia*, pages 1003–1004, 2008.

[76] 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 Language Processing*, 16(2):448–457, 2008.

[77] Y.-H. Yang and J.-Y. Liu. Quantitative study of music listening behavior in a social and affective context. *IEEE Trans. Multimedia*, 15(6):1304–1315, 2013.

[78] Y.-H. Yang, Y.-F. Su, Y.-C. Lin, and H. H. Chen. Music emotion recognition: The role of individuality. In *Proc. ACM Int. Work. Human-Centered Multimedia*, pages 13–21, 2007.

[79] C.-C. Yeh, S.-S. Tseng, P.-C. Tsai, and J.-F. Weng. Building a personalized music emotion prediction system. In *Advances in Multimedia Information Processing-PCM 2006*, pages 730–739. Springer, 2006.

[80] B. Zhu and T. Liu. Research on emotional vocabulary-driven personalized music retrieval. In *Edutainment*, pages 252–261, 2008.