Modeling the Affective Content of Music with a Gaussian Mixture Model

Ju-Chiang Wang, Yi-Hsuan Yang, *Member, IEEE*, Hsin-Min Wang, *Senior Member, IEEE*, and Shyh-Kang Jeng, *Senior Member, IEEE*

Abstract—Modeling the association between music and emotion has been considered important for music information retrieval and affective human computer interaction. This paper presents a novel generative model called acoustic emotion Gaussians (AEG) for computational modeling of emotion. Instead of assigning a music excerpt with a deterministic (hard) emotion label, AEG treats the affective content of music as a (soft) probability distribution in the valence-arousal space and parameterizes it with a Gaussian mixture model (GMM). In this way, the subjective nature of emotion perception is explicitly modeled. Specifically, AEG employs two GMMs to characterize the audio and emotion data. The fitting algorithm of the GMM parameters makes the model learning process transparent and interpretable. Based on AEG, a probabilistic graphical structure for predicting the emotion distribution from music audio data is also developed. A comprehensive performance study over two emotion-labeled datasets demonstrates that AEG offers new insights into the relationship between music and emotion (e.g., to assess the "affective diversity" of a corpus) and represents an effective means of emotion modeling. Readers can easily implement AEG via the publicly available codes. As the AEG model is generic, it holds the promise of analyzing any signal that carries affective or other highly subjective information.

Index Terms—Music information retrieval, music emotion recognition, valence, arousal, Gaussian mixture model, subjectivity

1 Introduction

Music has been widely used for mood and emotion regulation in our daily life, either for negative mood management, positive mood maintenance, or diversion from boredom [1], [2], [3]. In the digital age, emotion/mood is also considered as an important criterion used by people in organizing and navigating music libraries, according to music information behavior studies [4], [5], [6]. In light of this, a great deal of research work has been undertaken in the music information retrieval (MIR) community to computationally model the relationship between music and emotion [7]-[35]. In addition to data management purposes, such computational models also find applications in context-aware recommendation (e.g., that takes the emotion state of the listener into account) [36], [37], [38], affective human-computer interaction [39], [40], [41], [42], [43], [44], and music therapy [45], [46], [47], amongst others .

A fundamental issue in modeling emotion is that emotion perception is by nature subjective and highly dependent on the listener and the situational context [48], [49], [50]. It is possible for people to disagree on the affective content of the same music excerpt. Therefore, a straightforward, *deterministic* approach that associates each music excerpt with a single emotion label might not work well in practice [13], [14], [30]. In contrast, it would be better if the computational model takes the subjectivity issue into consideration and is able to accommodate individual and situational differences using for instance personalization or model adaptation techniques [21], [22], [23], [24], [31], [51], [52], [53].[1]

To illustrate this issue, in Fig. 1 we show the emotion annotations we collected for 60 excerpts of pop songs from human listeners in a previous work [22]. Each block shows the labels of 40 human listeners for an excerpt in the *valence-arousal* (VA) emotion space [54] using the user interface shown in Fig. 2.[2] We see that the emotion annotations (represented by cross marks) are concentrated for some excerpts but are fairly diverse for some others. The Krippendorff's α for measuring the inter-user agreement [43], [56] is 0.704 for arousal and 0.387 for valence, indicating that the perception for valence is in particular user-dependent. Therefore, many details of the affective content of an excerpt would be lost if we simply take the average VA ratings across the human listeners as the "genuine" (ground-truth) label for the excerpt.

- *J.-C. Wang and H.-M. Wang are with the Institute of Information Science, Academia Sinica, Taipei 115, Taiwan. E-mail: {asriver, whm}@iis.sinica.edu.tw.*
- *Y.-H. Yang is with the Research Center for Information Technology and Innovation, Academia Sinica, Taipei 115, Taiwan. E-mail: yang@citi.sinica.edu.tw.*
- *S.-K. Jeng is with the Department of Electrical Engineering, National Taiwan University, Taipei 106, Taiwan. E-mail: skjeng@cc.ee.ntu.edu.tw.*

Manuscript received 19 Feb. 2014; revised 3 Dec. 2014; accepted 16 Jan. 2015. Date of publication 1 Feb. 2015; date of current version 3 Mar. 2015.
Recommended for acceptance by A. Hanjalic.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TAFFC.2015.2397457

[1] Although the terms 'mood' and 'emotion' have different meanings in psychology [49], we use them interchangeably in this paper.

[2] In this example, and throughout this work, the affective content of music is described in the continuous space spanned by valence (affect appraisals: positive/negative) and arousal (or activation; energy and stimulation level) and that every listener can specify a point in the space to summarize the emotion of each short excerpt (30 seconds here). Depending on the experiment design, the emotion might refer to the one *perceived* as being expressed in a excerpt (this is the case for this example), or the emotion the listener actually *felt* in response to the stimulus [55].

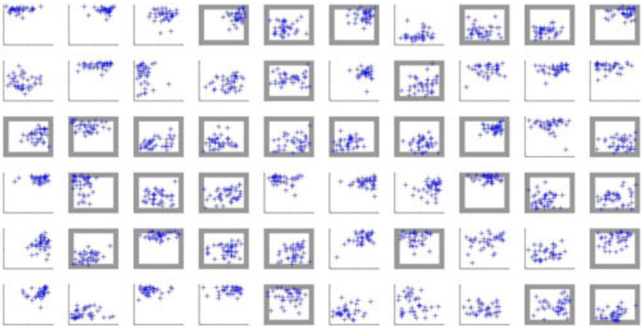


Fig. 1. The emotion annotation in a [−1,1]×[−1,1] VA space for the MER60 dataset [22].[3] Each block shows the labels made by 40 listeners (represented by cross marks) for a 30-second excerpt. We see that for some excerpts the inter-listener disagreement is high. A gray bounding box is drawn for an excerpt whose distribution of emotion annotation can be assumed to be Gaussian according to the Mardia multivariate normality test [57].

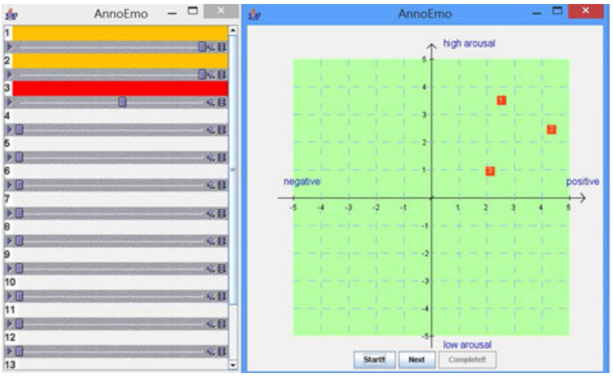


Fig. 2. A user interface [22] for rating the valence and arousal (VA) values of music (cf. Section 4.1). In this example, a user has listened to three excerpts and labeled all of them in the first quadrant of the VA space; different VA values were assigned according to the emotions the user perceived in these excerpts.

Despite that the subjectivity issue has been increasingly acknowledged [27], [31], few attempts have been made to develop a principled probabilistic framework for modeling emotion. Existing solutions to deal with subjectivity are most developed upon discriminative models that, for example, model the difference between the ‘general’ response (i.e., the average VA ratings for an excerpt) and an individual’s response [23]. Although such approaches might work better than the deterministic approach in practice, they do not provide a theoretical framework for understanding the relationship between emotion and music.

The present study aims at addressing this issue by using a generative model that is interpretable and theoretically sound. To this end, we propose modeling the affective content of music as a parametric probability distribution (i.e., a soft assignment) instead of a (hard) single point. This approach better accounts for the subjectivity, as it assumes that the emotion ratings from each human subject can be generated from the model. A novel probabilistic graphical model is then proposed to infer the emotion distribution from the acoustic features of music excerpts. In particular, we employ two sets of Gaussian mixture models (GMMs) to characterize the audio and emotion data, respectively. Accordingly, the generative model is referred to as the *acoustic emotion Gaussians* (AEG) model.

Compared to existing discriminative-based models [21], [22], [23], the proposed generative model is characterized by the transparency of its model learning process. Moreover, it provides a unified framework for *music emotion recognition* (i.e., automatic annotation) and *emotion-based music retrieval*, since one can map a music excerpt to the emotion space as well as an emotion-related query [58] to the acoustic feature space based on the generative process. Additional information such as user feedback can also be easily incorporated through the probabilistic model adaptation [59]. From a theoretical point of view, the proposed model is generic and hopefully lays a new foundation for emotion modeling of not only music but any real-world signal that carries affective information.

While the GMM assumption on acoustic data has been widely adopted [59], [60], [61], [62], [63], the GMM assumption on the emotion data is rare and needs justification. To this end, we want to verify the adequacy of using a bivariate Gaussian to model the emotion distribution of an excerpt. In statistics, the Mardia multivariate normality test [57] can be used to determine whether a dataset is well-modeled by a Gaussian distribution. Empirically, we found that the annotations of only 32 out of 60 excerpts in the MER60 dataset [22] can be assumed to be bivariate Gaussian under the Mardia’s test at significance level 5 percent [57], as Fig. 1 illustrates. For the remaining excerpts, a single Gaussian is not enough. Therefore, while it is computationally convenient to use a single Gaussian to model an emotion distribution, a GMM could be a better choice as it represents a finer granularity of emotion modeling.

The main ideas and the model learning algorithm of AEG have been introduced in [29], with application to both music emotion recognition and emotion-based retrieval. This paper extends and complements the prior one in the following aspects:

- We discuss the rationale and model assumptions of AEG in detail in this paper (Section 3).
- We qualitatively analyze the learning process of AEG and compare the models learned from two different corpora to offer insights into the probabilistic model (Section 4.2).
- We propose a novel way of measuring the *affective diversity* of an emotion-annotated corpus based on AEG (Section 4.3). This measure also provides insights into the accuracy that can be expected for music emotion recognition.
- We systematically evaluate the performance of AEG for music emotion recognition with more discussions on the parameter settings.

For reproductivity, the codes for implementing and evaluating AEG have been made publicly available to the research community.[4]

[3] Please note that for all the figures in this paper that show the VA space, the horizontal and vertical axes correspond to the valence and arousal dimensions, respectively.

[4] Available online, http://slam.iis.sinica.edu.tw/demo/AEG/

2 Related Work

Computational modeling of the relationship between music and emotion has been studied for years and many excellent reviews have been available [26], [27], [28]. The majority of existing work deals with the *annotation* aspect of emotion modeling, aiming at automatically annotating music excerpts with emotion. This task has also been referred to as 'music emotion recognition' [6], 'mood classification' [12] or 'emotion classification' [53], amongst others. Many approaches follow a typical pattern recognition paradigm and train classification or regression models [64] (depending on whether emotions are described in terms of discrete classes or continuous dimensions) trying to "reproduce" the ground-truth labels obtained from a listening test [65]. In contrast, the *retrieval* aspect—retrieving a set of music excerpts given an emotion-related query—has received relatively little attention. Although the proposed model can be applied to retrieval as well (as demonstrated in [29]), we focus on the annotation aspect in this paper. In particular, to avoid redundancy with recent surveys, our review here is centered around the subjectivity issue of modeling valence and arousal.

Valence and arousal have been widely recognized as the two most fundamental dimensions of emotion [54], [66]. As there is still no consensus on the "best" taxonomy for classifying emotions [27], the VA model suggests a simple yet powerful way of organizing emotions with two standard dimensions [67]. Moreover, representing emotions by VA values instead of discrete classes avoids the semantic ambiguity and possible overlaps of affective terms [27]. It also provides a continuous space that might correspond to the internal human representations of emotion [55], making it easier to track the dynamic emotion variation within a longer piece of music [8], [24], [25], [34], a case which is outside of the scope of this paper.

Early approaches to modeling valence and arousal (e.g., [15], [16]) assumed that the affective content of a music excerpt can be represented as a *single point* in the VA space. The ground-truth VA values of a music excerpt is obtained by averaging the annotations of a number of listeners, without considering the covariance of the annotations. Moreover, for simplicity, the two dimensions are usually assumed to be independent, so that one can model each dimension by fitting a regression model [64] that minimizes the error (e.g., mean squared difference) between the predicted and the ground-truth values. We refer to such methods collectively as the *VA-point* approach.

To better account for the subjective nature of emotion perception, approaches that consider emotion as a *distribution* rather than a point in the VA space have been proposed recently. Existing approaches generally fall into two categories: the *heatmap* approach and the *Gaussian-parameter* approach. The former quantizes each emotion dimension by τ equally spaced cells, leading to a $\tau \times \tau$ grid representation of the VA space [21], [24]. This approach then trains τ^2 regression models for predicting the emotion *intensity* at each cell. Higher intensity at a cell indicates that people are more likely to perceive the corresponding emotion from the excerpt. The heatmap can be considered as a non-parametric way of modeling an emotion distribution, as there is no assumption of the underlying distribution. However, it is less straightforward to build a generative model for the heatmap as dedicated methods have to be developed to model the correlation among adjacent cells.

TABLE 1
Approaches for Modeling Valence and Arousal

Approach	Type	Main idea
VA-point [15], [16], [19], [22], [23], [35]	Discriminative	Predict the mean VA values using regression
Heatmap [21], [24]	Discriminative	Predict the emotion density over a number of cells that quantizes the emotion space
Gaussian-parameter [21], [25]	Discriminative	Predict the parameters of the VA mean and covariance using independent regressors
Proposed [29]	Generative	Fit a Gaussian using a probabilistic graphical model

In contrast, the *Gaussian-parameter* approach [21], [25] parameterizes emotion distribution as a Gaussian and uses regression models to predict from acoustic features the Gaussian parameters (i.e., mean and covariance) of a music excerpt. This approach stands as an intuitive extension of the VA-point approach; one can apply lessons learned from previous work to construct effective models for the mean VA values. However, similar to the heatmap approach, the Gaussian-parameter approach is mostly discriminative and does not offer a strict probabilistic interpretation. The correlation among the Gaussian parameters also remains unmodeled. Because the parameters of covariance are modeled independently, the predicted covariance is not guaranteed to be a positive definite matrix. As a result, heuristic may be applied to adjust the predicted parameters to produce a valid covariance.

Table 1 summarizes the approaches discussed above. We note that few attempts have been made to build a generative framework for music emotion modeling. In Section 4.3, we will compare the performance of the proposed method with a Gaussian-parameter based approach for music emotion recognition.

Personalization is also an important aspect of dealing with the subjectivity issue. A number of personalization methods have been proposed for the VA-point approach [22], [23]. For example, the two-stage approach described in [23] trains a first-layer regression to model the general response and a second-layer one to model the difference between the general response and the individual response of the target user. In [51], we have presented some preliminary result of using model adaptation techniques developed in speech signal processing [59] to personalize AEG. The method uses *maximum a posteriori* to refine the model parameters of AEG given the feedback from a listener. Experiments conducted on MER60 showed that a personalized model can be built with a reasonable amount of user feedback.

3 Probabilistic Model

We make the following model assumptions involving a set of random variables: the audio data of a music excerpt $\mathbf{X}$, which contains a sequence of acoustic feature vectors computed over short-time frame instances $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, $\mathbf{x}_t \in \mathbb{R}^M$, the position of the excerpt on the continuous valence-arousal emotion space $\mathbf{y} \in \mathbb{R}^2$, and the associated discrete latent topic $\mathbf{z} \in \{1, 2, , K\}$.

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

$$p(\mathbf{x} \,|\, \mathbf{z} = k) \sim \mathcal{N}(\mathbf{m}_k, \mathbf{S}_k)\,. \tag{1}$$

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

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \,|\, \mathbf{m}_k, \mathbf{S}_k)\,, \tag{2}$$

where π_k, $\mathbf{m}_k$, and $\mathbf{S}_k$ are the model parameters associated with the kth latent topic.[5] Then, given an observed frame $\mathbf{x}_t$, the posterior probability of a topic is computed by

$$p(\mathbf{z} = k \,|\, \mathbf{x}_t) = \frac{\pi_k \mathcal{N}(\mathbf{x}_t \,|\, \mathbf{m}_k, \mathbf{S}_k)}{\sum_{h=1}^{K} \pi_h \mathcal{N}(\mathbf{x}_t \,|\, \mathbf{m}_h, \mathbf{S}_h)}\,. \tag{3}$$

3) The excerpt-level posterior probability of $\mathbf{z} = k$ given $\mathbf{X}$ can be approximated by averaging the frame-level posterior probabilities,[6]

$$p(\mathbf{z} = k \,|\, \mathbf{X}) \approx \frac{1}{T} \sum_{t=1}^{T} p(\mathbf{z} = k \,|\, \mathbf{x}_t)\,. \tag{4}$$

In other words, it is assumed that every frame of the excerpt has equal contribution.
4) The distribution of $\mathbf{y}$ given a topic $\mathbf{z} = k$ is Gaussian

$$p(\mathbf{y} \,|\, \mathbf{z} = k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,, \tag{5}$$

where the parameters $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ are associated with the kth latent topic as well. Accordingly, the marginal distribution of $\mathbf{y}$ is

$$\begin{aligned} p(\mathbf{y} \,|\, \mathbf{X}) &= \sum_k p(\mathbf{y} \,|\, \mathbf{z} = k) p(\mathbf{z} = k \,|\, \mathbf{X}) \\ &= \sum_k \mathcal{N}(\mathbf{y} \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) p(\mathbf{z} = k \,|\, \mathbf{X})\,. \end{aligned} \tag{6}$$

We make the following observations.

- The first assumption suggests that $\mathbf{z}$ uses K discrete latent topics to connect the acoustic feature space and the emotion space. Introducing such a hidden layer helps model the complicated relationship between the input and output data, as demonstrated by existing latent topic models such as probabilistic latent topic analysis (pLSA) [60] and latent Dirichlet allocation (LDA) [69].
- The Gaussian assumption on the acoustic feature space in assumption 2 is usually employed in audio signal processing [59], [60], [61], [62], [63]. We refer to the model $\{\pi_k, \mathbf{m}_k, \mathbf{S}_k\}_{k=1}^{K}$ as the *acoustic GMM*. The total number of parameters is $K{+}MK{+}M^2K$, which can be reduced to $2MK$ by assuming each $\pi_k = \frac{1}{K}$,[7] and each $\mathbf{S}_k$ to be diagonal (i.e., no correlation among the features) [70].
- Assumptions 2 and 3 in combination suggest that an audio encoding approach similar to *bag-of-frames* (BoF) [63], [71], [72], [73] is adopted. The BoF approach uses a *codebook* of size K to quantize a frame-level input $\mathbf{x}_t$ as a codeword and assumes that the excerpt-level information can be represented by the histogram over the *codewords* of the codebook. When the acoustic GMM is employed, the frame-level encoding result is computed as a probability $p(\mathbf{z}|\mathbf{x}_t)$, which has been proven to be more effective than the conventional BoF approach with smaller K [63], [74]. One can easily extend the frame-level features to block-level ones for representing $\mathbf{x}_t$, so that more local temporal characteristics can be captured [75]. We leave this for our future study.
- The Gaussian model of the emotion space in assumption 4 is intended to parameterize the emotion distribution owing to the subjective nature of emotion perception. The K topics collectively define a GMM for the emotion data. Therefore, we refer to the model $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ as the *affective GMM*. Unlike the acoustic GMM, we use a full covariance matrix (2×2) for each $\boldsymbol{\Sigma}_k$ to model the correlation between valence and arousal. We will illustrate several parameters of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ empirically learned from data in Section 4.2.
- The four assumptions as a whole suggest that music excerpts sharing similar distributions in $p(\mathbf{z}|\mathbf{X})$ would also have similar distributions in the emotion space, so that the annotations of different excerpts can blend with one another.
- The parameters $\boldsymbol{\Theta} \equiv \{\pi_k, \mathbf{m}_k, \mathbf{S}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ are statistical and can be estimated from data using maximum likelihood estimation

$$\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} \sum_{i=1}^{N} \log p(\mathbf{Y}^{(i)} | \mathbf{X}^{(i)}, \boldsymbol{\Theta})\,, \tag{7}$$

where N denotes the number of available excerpts for training, and the superscript (i) denotes the ith training excerpt. We will elaborate on the learning algorithm in Section 3.1.
- We can make personalized model adaptation inference over $\boldsymbol{\Theta}$ by $\max p(\boldsymbol{\Theta}|\mathbf{y}_u, \mathbf{X})$ [51], [76], where $\mathbf{y}_u$ denotes the information of a target user u, although this extension is not studied here.

[5] Because the association between the latent topics and audio data is modeled in the short-time level, one can model the time-varying emotion variation across the excerpt [8], [24], [25], [34]. Our recent work has realized this concern [68].

[6] We have found this formulation led to better empirical performance than using $(\prod_{t=1}^{T} p(\mathbf{z} = k \,|\, \mathbf{x}_t))^{1/T}$ in a pilot study.

[7] This uniform prior assumption has been shown adequate for musical acoustic data modeling [63].

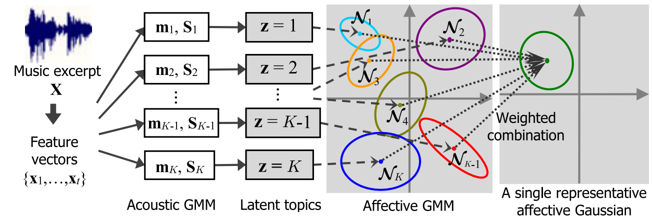

Music excerpt
X
Feature vectors
$\{\mathbf{x}_1,\ldots,\mathbf{x}_t\}$
$\mathbf{m}_1, \mathbf{S}_1$
$\mathbf{m}_2, \mathbf{S}_2$
$\mathbf{m}_{K\text{-}1}, \mathbf{S}_{K\text{-}1}$
$\mathbf{m}_K, \mathbf{S}_K$
$\mathbf{z} = 1$
$\mathbf{z} = 2$
$\mathbf{z} = K\text{-}1$
$\mathbf{z} = K$
$\mathcal{N}_1$
$\mathcal{N}_2$
$\mathcal{N}_3$
$\mathcal{N}_4$
$\mathcal{N}_{K\text{-}1}$
$\mathcal{N}_K$
Weighted combination
Acoustic GMM
Latent topics
Affective GMM
A single representative affective Gaussian

Fig. 3. Illustration of the generative process of the proposed acoustic emotion Gaussians model.

- The model is also applicable to describe emotions in more than two dimensions (e.g., by including the third important dimension *potency*, or dominant–submissive [77], [78]), although we focus on the VA model here for simplicity.

As Fig. 3 depicts, AEG involves a generative process for the affective content of music. For a given excerpt, we first extract the frame-level feature vectors, compute $p(\mathbf{z}|\mathbf{X})$ according to Eqs. (3) and (4), and then compute $p(\mathbf{y}|\mathbf{X})$ according to Eq. (6). If an excerpt's acoustic content $\mathbf{X}$ can be completely described by a single latent topic $\mathbf{z} = k$, i.e., $p(\mathbf{z} = k|\mathbf{X}) = 1$ and $p(\mathbf{z} = h|\mathbf{X}) = 0, \forall h \neq k$, its emotion distribution would exactly follow $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Otherwise, the emotion distribution would be a weighted combination of $\{\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}$ using $p(\mathbf{z}|\mathbf{X})$ as the weights.

3.1 Fitting the Model Parameters

To simplify the model fitting process, we divide the parameter set to $\{\pi_k, \mathbf{m}_k, \mathbf{S}_k\}_{k=1}^{K}$ (acoustic GMM) and $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ (affective GMM) and fit them separately.[8]

The acoustic GMM is conceptually similar to a codebook [72], [73] that is applicable to describe the acoustic feature bases for any music excerpt, but it is more general due to its probabilistic treatment. As the representative for a topic, moreover, the use of Gaussian distribution in fact conveys more semantic meanings than a single codeword does. An acoustic GMM can be learned from a collection of unlabeled frame-level acoustic feature vectors $\mathcal{U}$, whose size can be arbitrarily large since no human annotation is needed. This is a typical problem of learning a *universal background model* in the speech processing field and can be tackled by the expectation-maximization (EM) algorithm [59]. As aforementioned, we fix $\pi_k = \frac{1}{K}$ and assume each $\mathbf{S}_k$ to be diagonal for simplicity [63]. Once the parameters $\{\mathbf{m}_k, \mathbf{S}_k\}_{k=1}^{K}$ are learned, the resulting acoustic GMM can be used to compute $p(\mathbf{z}|\mathbf{X})$ (cf. Eq. (5)) for a music excerpt.

Fitting the affective GMM parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$, on the contrary, requires a labeled dataset $\mathcal{L} = \{\mathbf{X}^{(i)}, \mathbf{Y}^{(i)}\}_{i=1}^{N}$, where $\mathbf{Y}^{(i)} = [\mathbf{y}_1^{(i)}, \ldots, \mathbf{y}_{U^{(i)}}^{(i)}]$ denotes the set of VA values entered by listeners, $\mathbf{y}_j^{(i)} \in \mathbb{R}^2$ the individual annotation of the jth listener (e.g., each cross mark in Fig. 1), and $U^{(i)}$ the number of annotations available for the ith excerpt. Based on the acoustic GMM, for each training excerpt we compute $p(\mathbf{z}|\mathbf{X}^{(i)})$, which is called the *acoustic prior* and will stay fixed in the learning process of affective GMM. Then, the data log-likelihood can be derived by

$$\begin{aligned} L &= \log \prod_{i=1}^{N} \prod_{j=1}^{U^{(i)}} p(\mathbf{y}_j^{(i)} \mid \mathbf{X}^{(i)}) \\ &= \sum_{i,j} \log \sum_{k} \mathcal{N}(\mathbf{y}_j^{(i)} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) p(\mathbf{z} = k \mid \mathbf{X}^{(i)}) \,. \end{aligned} \tag{8}$$

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

$$\hat{L} = \sum_{i,j} \gamma_j^{(i)} \log \sum_{k} \mathcal{N}(\mathbf{y}_j^{(i)} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) p(\mathbf{z} = k \mid \mathbf{X}^{(i)}) \,, \tag{9}$$

where $0 \leq \gamma_j^{(i)} \leq 1$ and $\sum_{i,j} \gamma_j^{(i)} = 1$. This equation reduces to Eq. (8) when a uniform setting $\gamma_j^{(i)} = \frac{1}{\sum_h U^{(h)}}$ is adopted. We will describe a model for $\gamma_j^{(i)}$ in Section 3.2.1. One can observe from Eq. (9) a very important attribute of AEG, i.e., the affective GMM is learned based on the raw emotion annotations from each listener instead of the aggregated ones across subjects. This scheme directly takes the subjectivity into account, making AEG fundamentally different from the Gaussian-parameter approach (cf. Section 2).

Although maximizing $\hat{L}$ is intractable, we can employ the EM algorithm to find an approximated solution [79]. In the E-step, we compute the posterior probability of $\mathbf{z} = k$ given $\mathbf{y}_j^{(i)}$,

$$p(\mathbf{z} = k|\mathbf{y}_j^{(i)}) = \frac{p(\mathbf{z} = k|\mathbf{X}^{(i)}) \mathcal{N}(\mathbf{y}_j^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_h p(\mathbf{z} = k|\mathbf{X}^{(i)}) \mathcal{N}(\mathbf{y}_j^{(i)}|\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)} \,. \tag{10}$$

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

$$\boldsymbol{\mu}'_k \leftarrow \frac{\sum_{i,j} \gamma_j^{(i)} p(\mathbf{z} = k \,|\, |\mathbf{y}_j^{(i)}) \mathbf{y}_j^{(i)}}{\sum_{i,j} \gamma_j^{(i)} p(\mathbf{z} = k \mid \mathbf{y}_j^{(i)})} \,, \tag{11}$$

$$\boldsymbol{\Sigma}'_k \leftarrow \frac{\sum_{i,j} \gamma_j^{(i)} p(\mathbf{z} = k \mid \mathbf{y}_j^{(i)}) (\mathbf{y}_j^{(i)} - \boldsymbol{\mu}'_k)(\mathbf{y}_j^{(i)} - \boldsymbol{\mu}'_k)^T}{\sum_{i,j} \gamma_j^{(i)} p(\mathbf{z} = k \mid \mathbf{y}_j^{(i)})} \,. \tag{12}$$

The EM algorithm iteratively maximizes the value of $\hat{L}$ defined in Eq. (9) until convergence. One can fix the number of maximal iterations or set a stopping criterion according to the relative increase in $\hat{L}$.

As Eqs. (11) and (12) show, the parameter update is collectively determined by $\mathbf{y}_j^{(i)}, \gamma_j^{(i)}$ and $p(\mathbf{z}|\mathbf{y}_j^{(i)}), \forall i, j$. In this way, the learning process jointly takes the data likelihood, annotation prior and acoustic prior over the current affective GMM into consideration, so that the annotations of different excerpts can share with one another according to their corresponding probabilities. This is another unique attribute of AEG.

Algorithm 1 summarizes the learning process of affective GMM. The initialization of the parameters $\{\boldsymbol{\mu}_k^0, \boldsymbol{\Sigma}_k^0\}_{k=1}^{K}$ can be obtained by, for example, using the sample mean vector and covariance matrix $\boldsymbol{\mu}_{\mathcal{L}}, \boldsymbol{\Sigma}_{\mathcal{L}}$ computed over the whole data set $\mathcal{L}$.

[8] Another option is to jointly learn all the parameters $\{\pi_k, \mathbf{m}_k, \mathbf{S}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$; this is left as a future work.

Algorithm 1. Fitting the affective GMM

```
Input: Acoustic prior $\{p(\mathbf{z} \mid \mathbf{X}^{(i)})\}_{i=1}^{N}$ ;
       annotation prior $\{\gamma_j^{(i)}\}_{i=1,j=1}^{N,U^{(i)}}$ ;
       initial model $\{\boldsymbol{\mu}_k^0 = \boldsymbol{\mu}_{\mathcal{L}}, \boldsymbol{\Sigma}_k^0 = \boldsymbol{\Sigma}_{\mathcal{L}}\}_{k=1}^{K}$ ;
       maximal number of iterations $R$ or
       threshold of stopping ratio $\Gamma$ ;
Output: Model parameters $\{\boldsymbol{\mu}_k', \boldsymbol{\Sigma}_k'\}_{k=1}^{K}$
Initialize $r \leftarrow 0$ and $L_0$ using Eq. (9);
repeat
    Compute the posterior probability using Eq. (10) with $\{\boldsymbol{\mu}_k^r, \boldsymbol{\Sigma}_k^r\}_{k=1}^{K}$;
    $r \leftarrow r + 1$ ;
    Update $\{\boldsymbol{\mu}_k^r, \boldsymbol{\Sigma}_k^r\}_{k=1}^{K}$ using Eqs. (11) and (12) ;
    Compute $L_r$ using Eq. (9) ;
until $r = R$ or $(L_r - L_{r-1})/|L_{r-1}| < \Gamma$ ;
Let $\boldsymbol{\mu}_k' \leftarrow \boldsymbol{\mu}_k^r$ and $\boldsymbol{\Sigma}_k' \leftarrow \boldsymbol{\Sigma}_k^r$ ;
```

3.2 Optimizing the Learning Algorithm

3.2.1 Prior Model for Emotion Annotation

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

$$\gamma_j^{(i)} \leftarrow \frac{\mathcal{N}\big(\mathbf{y}_j^{(i)} | \mathbf{a}^{(i)}, \mathbf{B}^{(i)}\big)}{\sum_h \mathcal{N}\big(\mathbf{y}_h^{(i)} | \mathbf{a}^{(i)}, \mathbf{B}^{(i)}\big)} \,, \tag{13}$$

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

$$\gamma_j^{(i)} \leftarrow (1 - \lambda) \cdot 1 + \lambda \cdot \gamma_j^{(i)} \,. \tag{14}$$

3.2.2 Singularity Issue in Learning the Affective GMM

In practice, as the affective GMM is getting fitted to the data, a small number of affective Gaussian components might overly fit to some emotion annotations, leading to *singularity* [79]. When this occurs, some covariance matrices become non-positive definite (non-PD), making the corresponding affective Gaussians ill-defined. For instance, if a component affective Gaussian is contributed by only one or two annotations, the shape of its covariance becomes a point or a straight line. It is particularly important to avoid the singularity issue when the size of training examples is too small, when there are prevalent outliers in the set of annotations, or when the value of K is set to an overly large value. A straightforward approach to circumvent this issue is to perform *early stop* in Algorithm 1 by setting a smaller R (e.g., 8) or a larger Γ (e.g., 0.01), or to stop the EM update whenever a non-PD covariance matrix appears. However, both methods might lead to an insufficient model. Alternatively, one can i) regularize the covariance matrices by adding small values to the diagonal, or ii) remove that ill-conditioned Gaussian component, which in effect dynamically decreases the value of K. We adopt the latter approach here.

3.3 Discussion

With the generative process expressed in Eq. (6) (cf. assumption 4), we gain insights into potential extensions for the AEG model. If we can represent a type of musical features of a song (e.g., lyrics, melody, rhythm, music structure, or implicit music similarity) into a probability vector $p(\mathbf{z}|\mathbf{X})$, we can incorporate the feature into learning an AEG model. Such a process would need to first construct a set of (latent or visible) reference classes and then compute the posterior probability of a song over the classes. For example, we can predefine five rhythm classes and estimate for each music excerpt the posterior probability of each class based on a tempo detection algorithm. However, defining a set of effective reference classes for music emotion should consider diverse feature types. At the initial stage of developing AEG, we opt to use the low-level audio features and learn the reference classes (latent topics) via the acoustic GMM. It is believed that the learned topics can involve information such as rhythm and melody, which are usually estimated with specific algorithms from low-level features in MIR.[9]

4 Analysis

We demonstrate in this section how AEG can be employed to empirically analyze the relationship between music and emotion. We first present a qualitative study that investigates the parameter fitting process of AEG, and then an application of AEG to measure the quality of an emotion-annotated corpus. Two real-world music corpora are considered in this study: the MER60 [22] and DEAP [41] datasets.[10]

4.1 Music Corpora and Acoustic Features

The MER60 dataset consists of 60 pieces of 30-second clips selected from the chorus parts of contemporary Western pop songs [21], [22]. A total number of 99 participants (46

[9] We refer interested readers to [80] that demonstrates the latent topics of an acoustic GMM based on tags. For instance, we found that the first topic (cf. Table 2 in [80]) is associated with rhythm tags such as 'slow.'

[10] Available at http://mac.iis.sinica.edu.tw/ ~yang/MER/NTUMIR-60/and http://www.eecs.qmul.ac.uk/mmv/datasets/deap/

males and 53 females) were recruited for emotion annotation in a silent computer lab. Different participants annotated different numbers of songs, but we ensured that each song was annotated by exactly 40 different participants. The VA values, which are real values ranging in between [−1, 1], were entered by left-clicking on the continuous VA space displayed by the user interface shown in Fig. 2. The interface uses a small rectangle to indicate the annotation for a piece and permits instant playback of the piece by right-clicking on the rectangle. This encourages the listener to make careful comparison between the ratings of different pieces and revise pervious annotations if needed, which in turn improves the quality of annotation [22]. Prior to annotation, the listeners were instructed with the purpose of emotion modeling, the meaning of valence and arousal, and the difference between perceived and felt emotion [55]. The MER60 dataset is concerned with the *perceived* emotion.

The DEAP dataset [41] contains 120 pieces of one-minute music video clips of Western pop music collected from YouTube. The one-minute segment was intended to be the one with the "maximum emotional content" as estimated by an affective highlighting algorithm [10]. Each clip was annotated by 14-16 listeners (50 percent female), who were asked to rate the valence, arousal and dominance (i.e., potency) on a *discrete* nine-point scale from 1 to 9 using a web-based self-assessment tool [41]. Unlike MER60, for DEAP the listeners were not allowed to modify previous annotation and were asked to annotate the *felt* emotion.

For the acoustic feature representation, a hybrid set of frame-level energy, timbre and harmonic descriptors were computed by using the MIRToolbox [19], [81] with a frame size of 50 ms and 50 percent overlap. The features include root-mean-square energy, zero-crossing rate, spectral flux, centroid, spread, skewness, kurtosis, entropy, flatness, 85 percent-rolloff, 95 percent-rolloff, brightness, roughness, irregularity, 13-dimensional MFCCs, delta MFCCs, delta-delta MFCCs, key clarity, musical mode, harmonic changes likelihood, 12-bin chroma vector, chroma peak, and chroma centroid, leading to a 70-dimensional feature vector for a frame. Please refer to [20], [21] for details of the features. Please note that the use of the delta and delta-delta MFCC features is able to capture local temporal timbre patterns [82]. Each feature dimension was normalized to zero mean and unit variance. Such a hybrid set of feature descriptors are employed in most previous work on computationally modeling music (e.g., [11], [18], [20]), due to the complicated acoustic patterns involved in emotion induction and the difficulty of finding a universal feature representation that well characterizes every emotion [83], [84].[11]

To learn the acoustic GMM, we generated 235 K frame-level feature vectors from an in-house music collection (using the same 70-dimensional features by the MIRToolbox) to constitute $\mathcal{U}$. Based on an identical acoustic GMM, two affective GMMs were learned on the annotations of MER60 and DEAP, respectively.

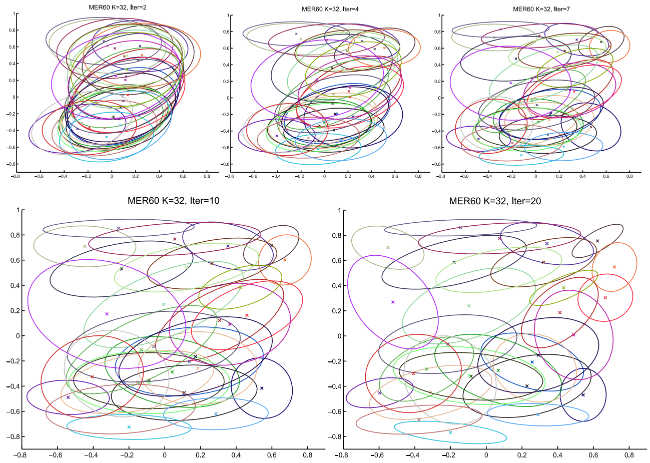


Fig. 4. The affective GMMs ($K = 32$) learned from the MER60 dataset [22] at iterations 2, 4, 7, 10, and 20 of Algorithm 1. Each Gaussian component is represented by an ellipse with a unique color.

4.2 Analyzing the Parameters of Affective GMM

To offer insights into the probabilistic model, we fit the parameters of affective GMM on the two datasets and examine the learned affective GMMs at each iteration of Algorithm 1. The number of latent topics K is set to 32 in this analysis; we will analyze the effect of K later in Section 5.2.

The affective GMMs learned from MER60 are depicted in Fig. 4, from which the following two trends are observed. First, while being close together in the beginning, the Gaussians gradually separate from one another as the iteration proceeds. Second, the size of the ellipse (i.e., the covariance) of each Gaussian gets increasingly smaller until convergence; at the 20th iteration, the Gaussians collectively cover different areas in the VA space, making it possible to approximate all kinds of emotion distribution by combining the learned affective GMM with different weights, which are set according to acoustic prior $p(\mathbf{z}|\mathbf{X})$ for each music excerpt individually.

We note that the specific area covered by a Gaussian suggests the mapping from the acoustic space to the emotion space governed by a specific latent topic. For example, the latent topics with affective Gaussians distributed in the first quadrant might represent happiness-related emotions. As another example, we see that there are many Gaussians with horizontally elongated ellipses, suggesting that it is more difficult to discriminate positive/negative valence, comparing to high/low arousal. This observation is in line with the empirical performance for modeling valence and arousal reported in the literature (e.g., [16], [24]). Discriminating emotions in the third and fourth quadrants (e.g., sadness and tenderness) appears to be even more challenging, as suggested by the large number of horizontally elongated ellipses in those areas. These examples well illustrate the insights that AEG can offer into computational emotion modeling. Such insights cannot be obtained if emotions are considered as points rather than distributions.

On the other hand, Fig. 5 shows the affective GMM learned from DEAP. By comparing Figs. 4 and 5, the following observations can be made.

[11] As the focus of this paper is on the computational model itself, we only consider conventional acoustic features that are usually adopted in related work.

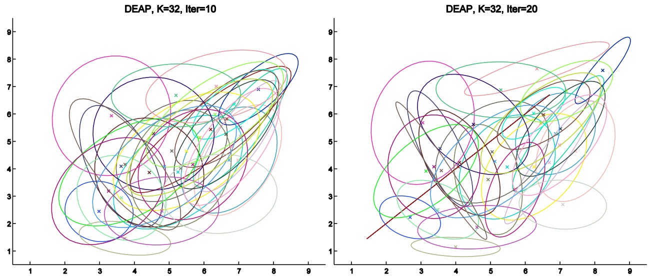


Fig. 5. The affective GMMs ($K = 32$) learned from DEAP at iterations 10 and 20.

- The *global emotion distribution* of the ground-truth annotations $\{\mathbf{Y}^{(i)}\}_{i=1}^{N}$ of a corpus can be approximately outlined by the affective GMM. We would expect that the model learned from MER60 is more generalizable, as the model covers almost the full VA space. In contrast, the model learned from DEAP might not be able to perform well in the corners of the second and fourth quadrants.
- One could assess the inter-user agreement of a corpus from the learned affective GMM. In particular, we would consider that the inter-user agreement is higher for MER60 than for DEAP, since the former contains smaller and diverse Gaussians, suggesting that the association between music and emotion is clearer for the listeners.
- Some of the latent topics might be too vague to be useful for emotion modeling, such as those topics with overly large, overly elongated, or largely overlapped component affective Gaussians. For better performance, some future study can be done to remove such latent topics according to the size and shape of the corresponding component affective Gaussians.
- The singularity issue (cf. Section 3.2.2) seems to be dataset-dependent. Possibly because the two datasets were developed in different ways (e.g., DEAP uses integer ratings whereas MER60 uses a continuous scale), we observe that a Gaussian for DEAP has degenerated to a straight line at iteration 20 (see Fig. 5), but no such case for MER60, even for more iterations or larger K (up to 256 according to our empirical observation).

4.3 Measuring Affective Diversity of a Corpus

Describing emotion as a probability distribution also makes it possible to assess the *affective diversity* of a corpus. This can be done by measuring the dissimilarity between the emotion distributions of different excerpts within the corpus. To this end, we can summarize the annotations $\mathbf{Y}^{(i)}$ for each song of an emotion-labeled corpus $\mathcal{L}$ by a single Gaussian $\mathcal{G}^{(i)}$, and then measure the difference between any $\mathcal{G}^{(i)}$ and $\mathcal{G}^{(j)}$, where $i \neq j$, by the information-theoretic Kullback-Leibler (KL) divergence [85]

$$D_{\mathrm{KL}}(\mathcal{G}_A||\mathcal{G}_B) = \frac{1}{2}\big(\mathrm{tr}(\boldsymbol{\Sigma}_A\boldsymbol{\Sigma}_B^{-1}) - \log|\boldsymbol{\Sigma}_A\boldsymbol{\Sigma}_B^{-1}| + (\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)^T\boldsymbol{\Sigma}_B^{-1}(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B) - d\big)\,, \tag{15}$$

TABLE 2
The PWKLs of the Two Datasets and the Accuracy of Music Emotion Recognition in Terms of AKL and AED

Dataset	PWKL	Method	AKL	AED
MER60	5.095	base-rate	4.179 ± 4.536	0.531 ± 0.165
		SVR [21]	2.052 ± 1.679	0.411 ± 0.216
		AEG	1.193 ± 1.387	0.342 ± 0.157
DEAP	1.194	base-rate	0.759 ± 0.744	1.405 ± 0.612
		SVR [21]	0.530 ± 0.502	1.212 ± 0.587
		AEG	0.453 ± 0.456	1.145 ± 0.516

where $\mathcal{G}_A \equiv \mathcal{N}(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A)$, $\mathcal{G}_B \equiv \mathcal{N}(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$ and $d = 2$ because the valence-arousal space is 2-dimensional. Accordingly, the *pairwise KL divergence* (PWKL) of $\mathcal{L}$ can be defined as

$$PWKL(\mathcal{L}) = \frac{1}{N_{PW}}\sum_{i\neq j} D_{\mathrm{KL}}(\mathcal{G}^{(i)}||\mathcal{G}^{(j)})\,, \tag{16}$$

where $N_{PW} = \frac{N(N-1)}{2}$ denotes the number of possible pairs in $\mathcal{L}$. Intuitively, larger PWKL indicates higher diversity of the annotations of a corpus.

It should be noted that the affective diversity of a corpus is dependent on two major factors:

- *Acoustic diversity*, whether the corpus contains music excerpts of different acoustic variations. This is related to the selection of excerpts and the length of excerpts to be annotated [86].
- *Annotation diversity*, whether the recruited listeners provided diverse set of emotion annotations. This is related to the selection of listeners, the tool, user interface and environment for annotations, and how the listeners are instructed [35].

Intuitively, if a corpus is high in affective diversity, a computational model learned from the corpus is more likely to be generalizable to any music excerpt. Whereas, for a corpus with small diversity, it would be possible to obtain high accuracy in emotion recognition by using a Gaussian with the sample mean and covariance of the corpus (i.e., $\mathcal{N}(\boldsymbol{\mu}_{\mathcal{L}}, \boldsymbol{\Sigma}_{\mathcal{L}})$). Therefore, we would consider the evaluation result of a diverse corpus as more useful.

The second column of Table 2 compares the PWKL of the two datasets. We see that the PWKL of MER60 (5.095) is much larger than that of DEAP (1.194). Because the two datasets are different in many ways (e.g., music vs. music video, perceived emotion vs. felt emotion, and continuous graphical interface in the lab vs. online ordinal rating), it is difficult to tell whether the difference in PWKL stems from acoustic diversity or annotation diversity. However, combined with the findings in Section 4.2, it should be safe to presume that the annotation consistency of MER60 is relatively higher. Accordingly, the system trained on MER60 would be more effective, because it might be less likely to give a predicted affective Gaussian that is close to the origin with a large covariance. We will further explain this in Section 5.2.

5 Music Emotion Recognition

In what follows, we systematically evaluate the performance of AEG for music emotion recognition on MER60 and DEAP. The purpose of this performance study is to validate the effectiveness of AEG and to investigate the effect of some parameter settings.

5.1 Algorithm

Given the model parameters $\{\pi_k, \mathbf{m}_k, \mathbf{S}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$, AEG can predict the emotion of an unlabeled music excerpt $\hat{\mathbf{X}}$ by following the generative process from the acoustic feature space to the emotion space. Specifically, it first computes $p(\mathbf{z}|\hat{\mathbf{X}})$ with Eqs. (3) and (4) from the acoustic features and then generates the emotion distribution as a weighted GMM:

$$p(\mathbf{y} \,|\, \hat{\mathbf{X}}) = \sum_{k=1}^{K} p(\mathbf{z} = k \,|\, \hat{\mathbf{X}}) \mathcal{N}(\mathbf{y} \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \,. \qquad (17)$$

In addition to $p(\mathbf{y}|\hat{\mathbf{X}})$, we can also use a single, representative affective Gaussian $\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ to summarize the weighted GMM, as illustrated in the rightmost part of Fig. 3. The representative Gaussian can be approximated by

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \sum_k p(\mathbf{z} = k|\hat{\mathbf{X}}) \boldsymbol{\mu}_k \,, \\ \hat{\boldsymbol{\Sigma}} &= \sum_k p(\mathbf{z} = k|\hat{\mathbf{X}}) \big(\boldsymbol{\Sigma}_k + (\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}})^T \big). \end{aligned} \qquad (18)$$

Interested readers are referred to [87] for the derivation of the above equations.

Representing the predicted result as a single Gaussian is functionally necessary owing to the following two reasons. First, 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). Second, to compare the performance of AEG with that of a Gaussian-parameter approach [21], [25], we have to follow the conventional setting that outputs the prediction as a single Gaussian and then measures the error between the predicted one and the ground-truth one.

However, using a single Gaussian may run counter to the theoretical arguments given in favor of a GMM that permits emotion modeling in a finer granularity. For instance, it is inadequate for the excerpts whose emotional responses are by nature bi-modal. We note that, in some applications such as emotion-based music retrieval [58] and music video generation [52] (those do not need to present the prediction to the users), one can directly use the raw weighted GMM (i.e., Eq. (17)) as the emotion index of a song in response to queries that can be represented in the VA space. In the case of music retrieval, for example, the ranking score can be obtained by feeding a VA point query into the affective GMM of a song [29]. As the retrieval perspective is beyond the scope of this paper, we leave the further study for our future work.

The computation of Eq. (18) is fairly efficient. The complexity depends mainly on K and the number of frames T of an excerpt: computing θ_k requires KT operations (cf. Eq. (3)), 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 database and for applications such as real-time music emotion tracking on a mobile device [34], [68], [88].

5.2 Evaluation

We performed leave-one-out validation (i.e., holding one excerpt for test in turn and the remaining for training) [21], [79] since both MER60 and DEAP are small in scale. Following [24], we evaluated the accuracy in terms of i) the one-way (asymmetric) KL divergence (cf. Eq. (15)) of the predicted Gaussian over the ground-truth one and ii) the euclidean distance between their mean vectors. The *average KL divergence* and *average euclidean distance* are termed as AKL and AED, respectively. Smaller AKL and AED indicate better performance. We consider AKL as the major performance indicator, because it takes both mean and covariance into consideration. Moreover, the value of AED is sensitive to the numerical range of the emotion space, whereas AKL is not.

We compared AEG with a base-rate predictor as well as the Gaussian-parameter approach [21], [25], which can be considered as a state-of-the-art method for modeling the emotion distribution from music signals. The idea of the base-rate predictor is to use a fixed prior Gaussian distribution as the predicted result for every test excerpt, without taking into account the acoustic features. The mean and covariance of this prior Gaussian is computed from the joint (global) annotations of the excerpts in the training set.

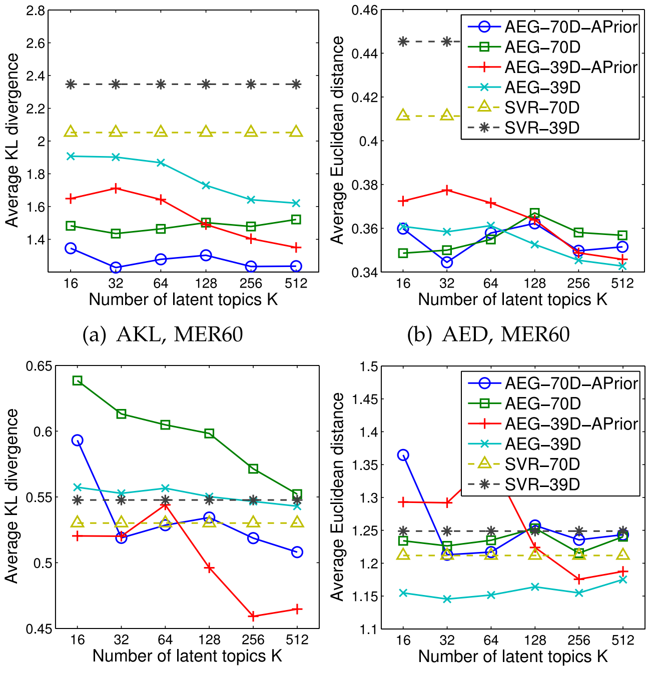
We employed support vector regression (SVR) [64] for the Gaussian-parameter approach to train five independent regression models for the mean and covariance parameters of the emotion Gaussian of a song. Our implementation of this 'SVR' approach was based on the free library LIBSVM [89],[12] along with the radial-basis function (RBF) kernel and a grid parameter search to optimize the parameters for each Gaussian parameter using inner cross-validation. We used the heuristic described in Algorithm 2 to make the predicted parameters of covariance valid.

Algorithm 2. Covariance regularization heuristic

```
Input: Covariance parameters $\{\sigma_{ij}\}_{i,j=1}^{d,d}$
for each diagonal element $\sigma_{ii}$ do
    if $\sigma_{ii} \leq 0$ then $\sigma_{ii} \leftarrow 0.01$ ;
end
for each off-diagonal element $\sigma_{ij}$ $(i \neq j)$ do
    if $\sigma_{ij} > \sqrt{\sigma_{ii}\sigma_{jj}}$ then $\sigma_{ij} \leftarrow \sqrt{\sigma_{ii}\sigma_{jj}}$ ;
    if $\sigma_{ij} < -\sqrt{\sigma_{ii}\sigma_{jj}}$ then $\sigma_{ij} \leftarrow -\sqrt{\sigma_{ii}\sigma_{jj}}$ ;
end
```

As for AEG, the following parameter settings were further evaluated: i) frame-level acoustic features (either the 70-dimensional features or the 39-dimensional subset with only MFCC-related features), ii) the number of latent topics K (ranging from 16 to 512), iii) whether or not to employ the annotation prior ($\lambda = 0$ or 1 defined in Section 3.2.1) when learning the affective GMM, and iv) the fixed maximal number of iterations R (cf. Section 3.2.2). For instance, "AEG-70D-APrior" denotes the model trained with the 70-dimensional features and the annotation prior ($\lambda = 1$), and "AEG-39D" denotes the model trained with MFCC features but without the annotation prior ($\lambda = 0$). For MER60, we set $R = 10$. As for DEAP, we set $R = 10$ when $K \leq 128$ and $R = 7$ otherwise.

[12] http://www.csie.ntu.edu.tw/~cjlin/libsvm



(a) AKL, MER60

(b) AED, MER60

(c) AKL, DEAP

(d) AED, DEAP

Fig. 6. The AKL and AED (both the lower the better) for music emotion recognition on MER60 and DEAP.

Figs. 6a and 6b show the evaluation result for MER60. It can be observed that AEG consistently outperforms SVR regardless of the value of K and the acoustic features. In particular, AEG-70D-APrior ($K = 32$) significantly outperforms SVR-70D in both AKL and AED (p-value $< 1\%$ under the two-tailed t-test). Similar observations can be made from Figs. 6c and 6d, which display the results for DEAP. For DEAP, the performance difference between AEG-39D-APrior ($K = 256$) and SVR-70D is also significant (p-value $<$ 5% under the two-tailed t-test). This result, which demonstrates the superiority of AEG over SVR in modeling emotion, is not surprising, as SVR does not learn the Gaussian parameters in a probabilistic way. Instead, AEG is by nature designed to model the parametric distribution of emotion annotations.

By comparing Figs. 6a and 6c, we see that the AKL of MER60 is generally much larger than that of DEAP (1.0–2.5 vs. 0.45–0.65). This is possibly due to lower affective diversity of DEAP (which might lead to an overly optimistic result for emotion recognition), as discussed in Section 4.3.[13]

Moreover, the following observations can be made regarding the parameter settings of AEG.

- *Acoustic features*. The 70-D feature vector leads to better AKL and AED for the diverse dataset, MER60, but the MFCC-related one performs slightly better for DEAP. For MER60, the performance of AEG-70D-APrior is competitive even with small K, whereas the case with AEG-39D-APrior requires larger K to achieve better performance. Therefore, the performance of different features appears to be dataset-dependent.
- *Number of latent topics*. The performance of AEG generally grows along with K until it saturates. This is expected since with higher K one can obtain finer acoustic feature modeling [59], [60], [61], [62], [63] at the cost of increasing model complexity as well as computational cost.
- *Annotation prior*. For both datasets, using the annotation prior consistently improves AKL but not AED. AKL is improved possibly because the annotation prior adds information regarding the annotation covariance of each training excerpt to the update of the model parameters (i.e., Eqs. (11) and (12)). However, this may introduce bias to the original annotation mean of a training excerpt and slightly harm AED. Therefore, the benefit of the annotation prior is not conclusive.
- *Early stop*. Smaller values of R not only avoid the singularity problem (cf. Section 3.2.2) but also prevent overfitting to the training data. Because of this early stop, AEG remains effective even when K is large. However, we also observe that overly large K (e.g., 512) would produce many redundant (overlapped) component affective Gaussians. If a subset of component affective Gaussians are about overlapped, their corresponding topics in the acoustic GMM are found to be very similar. Therefore, using large K may not empirically harm the accuracy of AEG, but it degrades the efficiency.

In Table 2, we summarize the PWKL for the two datasets, and the AKL and AED for base-rate predictor, SVR, and AEG (the latter two are with the best parameter settings). The base-rate predictor can be considered as a "non-effective" (baseline) model, where its prior Gaussian typically holds a large covariance and a mean close to the origin of the VA space. If a computational model is uncertain of the emotion of an excerpt, it tends to give a conservative estimate similar to a prior Gaussian, instead of a Gaussian with random parameters; this is especially the case for regression models [16]. Due to this property, one can also regard the performance of the base-rate predictor as additional measures of affective diversity for an emotion-labeled dataset. For example, we see that the AKL of base-rate is correlated with PWKL.

Table 2 shows that AEG outperforms the two competing methods greatly. The AKL obtained by AEG reaches 1.193 and 0.453 for MER60 and DEAP, respectively, which stands for 71.5 and 40.3 percent improvements over the base-rate predictor. There is also a pronounced performance difference between the base-rate predictor and SVR, confirming that the prediction of either SVR or AEG is effective.

Finally, in Fig. 7 we present the ground-truth emotion annotations from the listeners and their corresponding Gaussians (top row), and the predicted Gaussians by AEG-70D-Aprior ($K = 32$) (medium row) and SVR-70D [21] (bottom row) for six randomly selected clips of MER60. By comparing the three rows, it can be seen that the estimate of AEG has smaller size (i.e., covariance) and relatively more accurate position (mean) than that of SVR. This further validates the effectiveness of AEG.

[13] In contrast, Figs. 6b and 6d are not comparable because the two datasets use different scales (i.e., $[-1, 1]$ vs. $\{1, , 9\}$).

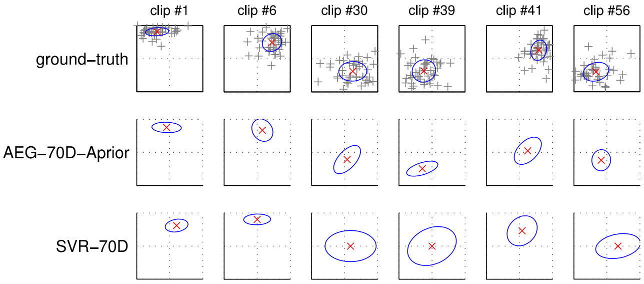


Fig. 7. The ground-truth emotion annotations and Gaussians of six clips selected from the MER60 dataset are presented in the top row. The corresponding predicted Gaussians by AEG and SVR are shown in the middle and bottom rows, respectively.

6 Conclusions and Future Work

In this paper, we have presented a generative approach for music emotion modeling, coined as acoustic emotion Gaussians. We have also provided theoretical discussions and empirical evidences showing that AEG represents a principled and effective probabilistic approach for research on emotion. Much more insights into the associations between music and emotion and the subjective nature of emotion perception can be gained by employing this model. The model can also be employed to qualitatively assess the inter-user agreement and the affective diversity of an emotion-labeled corpus.

Through the performance study on two well-known datasets, MER60 and DEAP [22], [41], we have also provided insights into the model learning process of AEG and the effect of different parameter settings. The effectiveness of AEG has also been validated in the context of automatic music emotion recognition by comparing against an existing Gaussian-parameter based method [21] and a base-rate predictor that uses a fixed annotation Gaussian.

For future work, we are interested in designing innovative personalization methods for emotion recognition and retrieval based on AEG [76], exploring other acoustic or lyrics features, incorporating temporal dynamics of acoustic signals for time-varying emotion tracking, making alignment between categorical and dimensional emotion semantics [90], and applying the model to other tasks such as context-aware recommendation [37], [38], implicit tagging [42], and computer-generating emotional music [47]. Another important future work would be to account for the effect of the listener mood while collecting emotion annotations and while evaluating the performance of the system [35], [91]. With the availability of the AEG source codes, we are looking forward to future endeavors that adopt AEG as a theoretical framework for studying emotions in related fields such as psychology or neurobiology.

Acknowledgments

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. Ju-Chiang Wang is the corresponding author.

References

[1] A. J. Lonsdale and A. C. North, "Why do we listen to music? a uses and gratifications analysis," *Brit. J. Psychol.*, vol. 102, pp. 108–134, 2011.

[2] M. Zentner, D. Grandjean, and K. R. Scherer, "Emotions evoked by the sound of music: Characterization, classification, and measurement," *Emotion*, vol. 8, no. 4, pp. 494–521, 2008.

[3] P. Juslin and P. Laukka, "Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening," *J. New Music Res.*, vol. 33, no. 3, pp. 217–238, 2004.

[4] J. H. Lee and J. S. Downie, "Survey of music information needs, uses, and seeking behaviours: Preliminary findings," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2004, pp. 441–446.

[5] M. Kamalzadeh, D. Baur, and T. Möller, "A survey on music listening and management behaviours," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2012, pp. 373–378.

[6] Y.-H. Yang and H. H. Chen, *Music Emotion Recognition*. Boca Raton, FL, USA: CRC Press, 2011.

[7] H. Katayose, M. Imai, and S. Inokuchi, "Sentiment extraction in music," in *Proc. Int. Conf. Pattern Recog.*, 1998, pp. 1083–1087.

[8] E. Schubert, "Modeling perceived emotion with continuous musical features," *Music Perception*, vol. 21, no. 4, pp. 561–585, 2004.

[9] M. Leman, V. Vermeulen, L. D. Voogdt, D. Moelants, and M. Lesaffre, "Prediction of musical affect using a combination of acoustic structural cues," *J. New Music Res.*, vol. 34, no. 1, pp. 39–67, 2005.

[10] A. Hanjalic and L. Xu, "Affective video content representation and modeling," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[11] L. Lu, D. Liu, and H. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 1, pp. 5–18, Jan. 2006.

[12] 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 Inf. Retrieval Conf.*, 2008, pp. 462–467.

[13] Y.-H. Yang, C. C. Liu, and H. H. Chen, "Music emotion classification: A fuzzy approach," in *Proc. ACM Multimedia*, 2006, pp. 81–84.

[14] C.-C. Yeh, S.-S. Tseng, P.-C. Tsai, and J.-F. Weng, "Building a personalized music emotion prediction system," in *Prof. Pacific Rim Conf. Multimedia*, 2006, pp. 730–739.

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

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

[17] X. Hu and J. S. Downie, "When lyrics outperform audio for music mood classification: A feature analysis," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2010, pp. 619–624.

[18] 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.*, vol. 39, no. 1, pp. 13–34, 2010.

[19] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2009, pp. 621–626.

[20] Y.-H. Yang and H. H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 4, pp. 762–774, May 2011.

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

[22] Y.-H. Yang, Y.-F. Su, Y.-C. Lin, and H. H. Chen, "Music emotion recognition: The role of individuality," in *Proc. ACM Int. Workshop Human-Centered Multimedia*, 2007, pp. 13–21.

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

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

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

[26] 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 Inf. Retrieval Conf.*, 2010, pp. 255–266.

[27] Y.-H. Yang and H.-H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, p. 40, 2012.

[28] M. Barthet, G. Fazekas, and M. Sandler, "Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models," in *Proc. Int. Symp. Comput. Music Model. Retrieval*, 2012, pp. 492–507.

[29] 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*, 2012, pp. 89–98.

[30] T.-L. Wu and S.-K. Jeng, "Probabilistic estimation of a novel music emotion model," in *Proc. Int. Multimedia Model. Conf.*, 2008, pp. 487–497.

[31] A. C. Mostafavi, Z. W. Raś, and A. Wieczorkowska, "Developing personalized classifiers for retrieving music by mood," in *Proc. Int. Workshop New Frontiers Mining Complex Patterns*, 2013.

[32] 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. 14th Int. Soc. Music Inf. Retrieval Conf.*, 2013, pp. 201–206.

[33] R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. P. Paiva, "Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis," in *Proc. Int. Symp. Comput. Music Model. Retrieval*, 2013, pp. 1–13.

[34] V. Imbrasaitė, T. Baltrušaitis, and P. Robinson, "Emotion tracking in music using continuous conditional random fields and baseline feature representation," in *Proc. Int. Workshop Affective Anal. Multimeda*, 2013, pp. 1–6.

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

[36] S. J. Breckler, R. B. Allen, and V. J. Konečni, "Mood-optimizing strategies in aesthetic-choice behavior," *Music Perception*, vol. 2, no. 4, pp. 459–470, 1985.

[37] N. Kummer, D. Kadish, A. Dulic, and H. Najjaran, "The empathy machine," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2012, pp. 2265–2271.

[38] Y.-H. Yang and J.-Y. Liu, "Quantitative study of music listening behavior in a social and affective context," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1304–1315, 2013.

[39] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[40] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affective Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[41] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 18–31, Jan.–Mar. 2012.

[42] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 42–55, Jan.–Mar. 2012.

[43] I. Siegert, R. Böck, and A. Wendemuth, "Inter-rater reliability for emotion annotation in human-computer interaction: Comparison and methodological improvements," *J. Multimodal User Interfaces*, vol. 8, no. 1, pp. 17–28, 2013.

[44] H. Gunesa and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image Vis. Comput.*, vol. 31, no. 2, pp. 120–136, 2013.

[45] C. L. Pelletier, "The effect of music on decreasing arousal due to stress: A meta-analysis," *J. Music Therapy*, vol. 41, pp. 192–214, 2004.

[46] P. Laukka, "Uses of music and psychological well-being among the elderly," *J. Happiness Stud.*, vol. 8, no. 2, pp. 215–241, 2007.

[47] I. Wallis, T. Ingalls, and E. Campana, "Computer-generating emotional music: The design of an affective music algorithm," in *Proc. Int. Conf. Digital Audio Effects*, 2008, pp. 1–6.

[48] M. Zentner and K. R. Scherer, "Emotional effects of music: production rules," in *Music and Emotion: Theory and Research*, P. N. Juslin and J. A. Sloboda, Eds. New York, NY, USA: Oxford Univ. Press, 2001.

[49] K. R. Scherer, "Which emotions can be induced by music? what are the underlying mechanisms? and how can we measure them," *J. New Music Res.*, vol. 33, no. 5, pp. 239–251, 2004.

[50] J. Hanratty, "Individual and situational differences in emotional expression," Ph.D. dissertation, School of Psychol., Queen's Univ. Belfast, Belfast, U.K., 2010.

[51] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng, "Personalized music emotion recognition via model adaptation," in *Proc. APSIPA Annu. Summit Conf.*, 2012, pp. 1–7.

[52] 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*, 2012, pp. 1379–1380.

[53] D. Su and P. Fung, "Personalized music emotion classification via active learning," in *Proc. ACM Int. Workshop Music Inf. Retrieval User-Centered Multimodal Strategies*, 2012, pp. 51–62.

[54] J. A. Russell, "A circumplex model of affect," *J. Personality Soc. Sci.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[55] A. Gabrielsson, "Emotion perceived and emotion felt: Same or different?" *Musicae Sci.*, vol. 5, pp. 123–147, 2002.

[56] K. Krippendorff, *Content Analysis: An Introduction to its Methodology*. Thousand Oaks, CA, USA: Sage, 2013.

[57] K. V. Mardia, "Measures of multivariate skewness and kurtosis with applications," *Biometrika*, vol. 57, no. 3, pp. 519–530, 1970.

[58] 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*, 2008, pp. 1003–1004.

[59] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2000.

[60] E. Hörster, R. Lienhart, and M. Slaney, "Continuous visual vocabulary models for pLSA-based scene recognition," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2008, pp. 319–328.

[61] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. Okuno, "An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 2, pp. 435–447, 2008.

[62] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[63] 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 Inf. Retrieval Conf.*, 2011, pp. 85–90.

[64] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Comput.*, vol. 12, pp. 1207–1245, 2000.

[65] B. L. Sturm, "Evaluating music emotion recognition: Lessons from music genre recognition," in *Proc. Int. Workshop Affective Anal. Multimeda*, 2013, pp. 1–6.

[66] R. E. Thayer, *The Biopsychology of Mood and Arousal*. New York, NY, USA: Oxford Univ. Press, 1989.

[67] J. A. Sloboda and P. N. Juslin, "At the interface between the inner and outer world: Psychological perspectives," in *Handbook of Music and Emotion: Theory, Research, Applications*, P. N. Juslin and J. A. Sloboda, Eds. New York, NY, USA: Oxford Univ. Press, 2010.

[68] J.-C. Wang, H.-M. Wang, and G. Lanckriet, "A histogram density modeling approach to music emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, accepted for publication, 2015.

[69] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[70] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*. Englewood Cliffs, NJ, USA: Prentice Hall PTR, 2001.

[71] M. Hoffman, D. Blei, and P. Cook, "Easy as CBA: A simple probabilistic model for tagging music," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2009, pp. 369–374.

[72] 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*, vol. 16, no. 5, pp. 1188–1200, Aug. 2014.

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

[74] J.-C. Wang, H.-S. Lee, S.-K. Jeng, and H.-M. Wang, "Posterior weighted Bernoulli mixture model for music tag annotation and retrieval," in *Proc. APSIPA Annu. Summit Conf.*, 2010, pp. 247–252.

[75] K. Seyerlehner, "Content-based music recommender systems: Beyond simple frame-level audio similarity," Ph.D. dissertation, Johannes Kepler Univ. Linz, Linz, Austria, 2010.

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

[77] 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 Emotion*, vol. 19, no. 8, pp. 1113–1139, 2005.

[78] G. Collier, "Beyond valence and activity in the emotional connotations of music," *Psychol. Music*, vol. 35, no. 1, pp. 110–131, 2007.

[79] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2006.

[80] J.-C. Wang, M.-S. Wu, H.-M. Wang, and S.-K. Jeng, "Query by multi-tags with multi-level preferences for content-based music retrieval," in *Proc. Int. Conf. Multimedia Expo*, 2011, pp. 1–6.

[81] O. Lartillot and P. Toiviainen, "A Matlab toolbox for musical feature extraction from audio," in *Proc. Int. Conf. Digital Audio Effects*, 2007, pp. 237–244.

[82] S. Furui, "Speaker-independent isolated word recognition based on emphasized spectral dynamics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1986, vol. 11, pp. 1991–1994.

[83] P. N. Juslin, "Cue utilization in communication of emotion in music performance: Relating performance to perception," *J. Exp. Psychol.: Human Perception Perform.*, vol. 16, no. 6, pp. 1797–1813, 2000.

[84] A. Gabrielsson and E. Lindström, "The influence of musical structure on emotional expression," in *Music and Emotion: Theory and Research*, P. N. Juslin and J. A. Sloboda, Eds. New York, NY, USA: Oxford Univ. Press, 2001.

[85] J. Hershey and P. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2007, vol. 4, pp. 317–320.

[86] Y.-H. Yang and X. Hu, "Cross-cultural music moood classification: A comparison on English and Chinese songs," in *Proc. 8th Int. Soc. Music Inf. Retrieval Conf.*, 2012, pp. 19–24.

[87] J.-C. Wang, Y.-H. Yang, and H.-M. Wang, "Affective music information retrieval," in *Emotions and Personality in Personalized Services*, M. Tkalcic et al., Eds., in press, 2015.

[88] J.-C. Wang, H.-M. Wang, and S.-K. Jeng, "Playing with tagging: A real-time tagging music player," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 77–80.

[89] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 27:1–27:27, 2011.

[90] 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. Int. ACM Workshop Music Inf. Retrieval User-Centered Multimodal Strategies*, 2012, pp. 63–68.

[91] M. Quirin, M. Kazén, and J. Kuhl, "When nonsense sounds happy or helpless: The implicit positive and negative affect test (ipanat)," *J. Personality Soc. Psychol.*, vol. 97, no. 3, p. 500, 2009.

Ju-Chiang Wang received the PhD degree in electrical engineering from National Taiwan University, Taiwan, in 2013. He has been with the Institute of Information Science, Academia Sinica, since 2007, and is currently a visiting postdoctoral fellow in the Department of Electrical and Computer Engineering, UC San Diego. His research interests mainly encompass music information retrieval (MIR), machine learning, and audio signal processing. He received the First Prize of Multimedia Grand Challenge in ACM Multimedia 2012, the Silver Medal from 2013 Merry Electroacoustic Thesis Award, Taiwan, and the 2013 Best PhD Dissertation Award from Taiwanese Association for Artificial Intelligence (TAAI). Apart from a computer science researcher, he is a semi-professional musician playing and producing music in some indie-rock bands.

Yi-Hsuan Yang (M'11) received the PhD degree in communication engineering from National Taiwan University, Taiwan, in 2010. Since 2011, he has been affiliated with the Academia Sinica Research Center for Information Technology Innovation as an assistant research fellow. He is also an adjunct assistant professor with the National Cheng Kung University. His research interests include music information retrieval, machine learning, and affective computing. He received the 2011 IEEE Signal Processing Society (SPS) Young Author Best Paper Award, the 2012 ACM Multimedia Grand Challenge First Prize, the 2012 Academia Sinica Career Development Award, and the 2013 Project for Excellent Junior Research Investigators by the National Science Council of Taiwan. He is a coauthor of the book *Music Emotion Recognition* (CRC Press 2011) and a tutorial speaker on music affect recognition in the International Society for Music Information Retrieval Conference (ISMIR 2012). In 2014, he will serve as a technical program cochair of ISMIR, a coorganizer of the International Workshop on Affective Analysis in Multimedia and MediaEval 'Emotion in Music' Task, and a guest editor of the *IEEE Transactions on Affective Computing*. He is a member of the IEEE.

Hsin-Min Wang (S'92–M'95–SM'04) received the BS and PhD degrees in electrical engineering from National Taiwan University in 1989 and 1995, respectively. In October 1995, he joined the Institute of Information Science, Academia Sinica, where he is currently a research fellow and deputy director. He also holds a joint appointment as professor in the Department of Computer Science and Information Engineering at National Cheng Kung University. He currently serves as the president of the Association for Computational Linguistics and Chinese Language Processing (ACLCLP), a managing editor of *Journal of Information Science and Engineering*, and an editorial board member of *APSIPA Transactions on Signal and Information Processing* and *International Journal of Computational Linguistics and Chinese Language Processing*. His major research interests include spoken language processing, natural language processing, multimedia information retrieval, and pattern recognition. He received the Chinese Institute of Engineers (CIE) Technical Paper Award in 1995 and the ACM Multimedia Grand Challenge First Prize in 2012. He is an APSIPA distinguished lecturer for 2014-2015. He is a member of the International Speech Communication Association (ISCA) and ACM. He is a senior member of the IEEE.

Shyh-Kang Jeng (M'86-SM'98) received the BSEE and the PhD degrees from National Taiwan University, Taipei, Taiwan, in 1979 and 1983, respectively. In 1981, he joined the faculty of the Department of Electrical Engineering, National Taiwan University, where he is currently a professor. Between 1985 and 1993, he was a visiting research associate professor and a visiting research professor with the University of Illinois, Urbana-Champaign. In 1999, he visited the Center for Computer Research in Music and Acoustics, Stanford University, Stanford, CA, for six months. His research interest includes time-domain electromagnetic field computation techniques, antenna design, multimedia signal processing, computational neuroscience, and computational cognitive neuroscience. He received the 1998 Outstanding Research Award of National Science Council and the 2004 Outstanding Teaching Award of National Taiwan University. He is a senior member of the IEEE.