Exploring the Relationship between Categorical and Dimensional Emotion Semantics of Music*

Ju-Chiang Wang[1,2], Yi-Hsuan Yang[1], Kaichun Chang[3], Hsin-Min Wang[1], Shyh-Kang Jeng[2]
[1]Institute of Information Science, Academia Sinica, Taiwan
[2]Department of Electrical Engineering, National Taiwan University, Taiwan
[3]Department of Computer Science, King's College London, UK
[1]{asriver, yang, whm}@iis.sinica.edu.tw; [2]skjeng@cc.ee.ntu.edu.tw; [3]ken.chang@kcl.ac.uk

ABSTRACT

Computational modeling of music emotion has been addressed primarily by two approaches: the *categorical* approach that categorizes emotions into mood classes and the *dimensional* approach that regards emotions as numerical values over a few dimensions such as valence and activation. Being two extreme scenarios (discrete/continuous), the two approaches actually share a unified goal of understanding the emotion semantics of music. This paper presents the first computational model that unifies the two semantic modalities under a probabilistic framework, which makes it possible to explore the relationship between them in a computational way. With the proposed framework, mood labels can be mapped into the emotion space in an unsupervised and content-based manner, without any training ground truth annotations for the semantic mapping. Such a function can be applied to automatically generate a semantically structured tag cloud in the emotion space. To demonstrate the effectiveness of the proposed framework, we qualitatively evaluate the mood tag clouds generated from two emotion-annotated corpora, and quantitatively evaluate the accuracy of the categorical-dimensional mapping by comparing the results with those created by psychologists, including the one proposed by Whissell & Plutchik and the one defined in the Affective Norms for English Words (ANEW).

Categories and Subject Descriptors

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

Keywords

Computational music emotion analysis, multi-modal emotion semantic modeling, automatic layout of tag cloud.

1. INTRODUCTION

A central issue of music emotion research is the conceptualization of emotion and the associated emotion taxonomy. There is still no consensus on which emotion model or how many emotion categories should be used [1]. Different viewpoints on this issue have led to the proposal of different ways of data annotation, model training, and result presentation. Existing work can be primarily classified into two approaches: the *categorical* approach that views emotions as discrete mood classes (such as happy, angry, sad, and relaxed) [2, 3] and the *dimensional* approach that regards emotions as numerical values over dimensions such as valence (positive/negative affective states) and activation (or arousal; energy level) [4–6], the two most fundamental emotion dimensions discovered by psychologists [7]. The valence-activation space is often referred to as the VA space in the literature.

From the application point of view, the two approaches actually offer complementary advantages: the former offers an atomic description of music that is easy to be incorporated into conventional text-based retrieval systems, while the latter offers a simple means for constructing a 2-D user interface [4]. It is therefore interesting to investigate the relationship between them and combine their advantages to enhance the performance of emotion-based music retrieval systems. For example, as a novice user may be unfamiliar with the essence of the valence and activation dimensions, it would be helpful to display mood tags in the emotion space to give the user some cues. This can be achieved if we have the mapping between the mood tag space and the VA space.

The visualization of mood tags over the VA space is interesting in its own right. As there could be hundreds of mood tags [8], a mood tag cloud with clear semantic structure would greatly facilitate applications such as tag-based music search and browsing. In addition to simple methods such as organizing mood tags as a sortable long list, dimension reduction techniques, such as Self-Organizing Map (SOM), have been adopted to determine the proximity of tags in the tag cloud according to some similarity functions [3, 9]. However, a major problem with dimension reduction is that the resulting dimensions may not strictly conform to valence and activation, sometimes even do not have strong semantic meaning. For example, Figure 1 shows the tag cloud generated with SOM by Laurier *et al.* [3]. Although similar tags are close to one another in this layout, the two dimensions do not correspond to valence and activation. The ambiguity of semantic interpretation hinders the practical usage of the tag cloud. Moreover, from such tag cloud it is difficult to infer the *emotion intensity* of the mood tags. In contrast, one can regard the distance from the origin as a measure of emotion intensity in the VA space.

The valence and activation values of some affective terms can be found in the psychology literature [10, 11], but most of them concern general emotions that are experienced in daily life, rather than emotions that are elicited from music

*This work was supported by the National Science Council of Taiwan under Grants: NSC-101-2221-E-001-019-MY3, NSC-101-2631-H-001-007, and NSC-100-2218-E-001-009.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MIRUM'12, November 2, 2012, Nara, Japan.
Copyright 2012 ACM 978-1-4503-1089-5/12/10 ...$15.00.

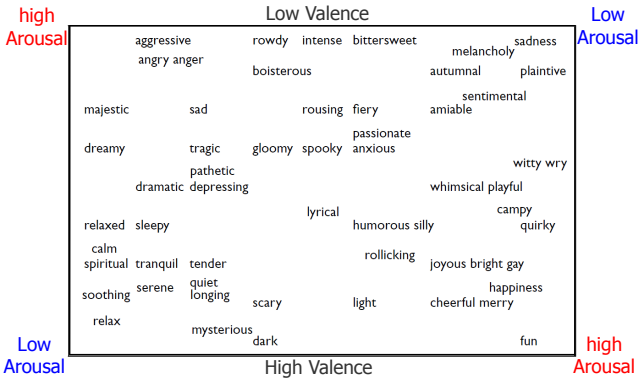


Figure 1: A mood tag cloud generated by SOM [3].[2]

listening [12]. Therefore, the valence and activation values may not be available for affective terms that are specific to music perception. Furthermore, affective terms are not cross lingual and not always have exact translations in different languages; the mapping for the affective terms can be cultural-dependent or even corpus-dependent [13]. Therefore, it would be desirable if we can learn the categorical-dimensional mapping automatically from data.

In this paper, we explore the relationship between the categorical and dimensional emotion semantics from a *machine learning* perspective. Specifically, we establish a probabilistic framework consisting of two component models, the *Acoustic Tag Bernoullis* (ATB) model and the *Acoustic Emotion Gaussians* (AEG) model, to computationally model the generative processes from acoustic features (e.g., timbre and harmony) to the perceptions of a mood tag and a pair of valence-activation values, respectively. The ATB and AEG models are learned from an emotion-tagged dataset and a VA-annotated dataset, respectively. In addition, a set of *latent feature classes* are introduced as a bridge between the two spaces, so that the ATB and AEG models can share and transit the semantic information to each other. The latent feature classes are learned directly from acoustic features, and thus the mood-tagged dataset and the VA-annotated dataset can be totally separate, relieving the requirement for a jointly-annotated dataset for the two emotion modalities. In other words, the categorical-dimensional mapping is learned *without any training ground truth VA values of the mood tags*. Once the ATB and AEG models are learned from data and combined by the latent feature classes, it is easy to align one emotion modality to the other (e.g., to strictly align mood tags to the VA space).

The proposed framework has three additional advantages. First, considering that emotion perception is inherently subjective, it computes a "soft" assignment of the VA values for each tag [14], which can be utilized as additional visual aids for the tag cloud. Second, as a generative model, the model learning and semantic mapping processes are transparent and easy to be observed and interpreted. For instance, one can analyze the learned model to understand why a tag is mapped into a certain area in the VA space, or verify whether an arbitrary tag is mood-related or not. Third, this work represents one of the first attempts that establish a joint model for discrete mood tags and continuous emotion dimensions, which have been addressed separately in the literature. It thereby offers new opportunities and insights to the research on affective computing as a whole [2, 5, 6].

The rest of the paper is organized as follows. Section 2 describes the methodology. Section 3 presents the implementation and evaluations. Section 4 discusses the results and implications. Section 5 concludes the paper.

[2]While the vertical axis appears to correspond to valence, there is no strong semantic meaning of the horizontal axis; we can find high-activation terms in the top-left and bottom-right corners, yet low-activation terms in the top-right and bottom-left corners.

2. METHODOLOGY

This section presents the methodology of the proposed framework, as illustrated in Figure 2. We first introduce the generative processes of the tags and VA emotions along with the mathematical details, respectively. Then, we describe the mapping procedure from a mood tag to a representative emotion Gaussian in the VA space.

2.1 Acoustic GMM Posterior Representation

There are two component models in the framework, namely the acoustic tag Bernoullis (ATB) model and the acoustic emotion Gaussians (AEG) model, which model the tag and VA spaces respectively from the same acoustic feature space. To start the generative processes of ATB and AEG, we apply the idea of the bag-of-frames modeling [16] that utilizes a universal acoustic Gaussian mixture model (GMM) as an acoustic feature reference to map a set of frame-based acoustic feature vectors of a music clip into a fixed dimensional probabilistic representation. The acoustic GMM, which is pre-learned using the EM algorithm [15] on a universal set of frame-based feature vectors, $\mathcal{F}$, is defined as,

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

where $\mathbf{x}$ is a frame-based feature vector; π_k, $\mathbf{m}_k$, and $\mathbf{S}_k$ are the prior weight, mean vector, and covariance matrix of the k-th component acoustic Gaussian, which is denoted by a *latent feature classes* z_k (cf. Figure 2). Accordingly, each z_k, which is derived by the acoustic GMM learning, represents a certain kind of acoustic pattern.

For an arbitrary music clip s, the acoustic posterior probability of z_k for its t-th frame vector $\mathbf{x}_t$ is given by

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

In the implementation, we replace π_k in Eq. 2 with $\frac{1}{K}$ because π_k was not useful in our previous work [16].

Then, the *clip-level* acoustic GMM posterior probabilities $\{\theta_k\}_{k=1}^{K}$ (cf. Figure 2) can be summarized by

$$\theta_k \leftarrow p(z_k|s) = \frac{1}{T} \sum_{t=1}^{T} p(z_k|\mathbf{x}_t), \tag{3}$$

where T is the number of frames in s. On can see from Eq. 3 that a component acoustic posterior probability θ_k actually is obtained based on the whole set of frames of the music clip. This statistical representation is able to capture the long-term acoustic characteristics of every music clip in a K-dimensional probabilistic space, and thus it should be sufficient for music emotion modeling. The acoustic GMM posterior of a music clip s_i in the corpus is then represented by vector $\boldsymbol{\theta}_i$, whose k-th component is θ_{ik}.

2.2 Learning the ATB Model

The left hand side of Figure 2 shows the ATB model. Suppose that we have a tagged music corpus $\mathcal{X}_\mathcal{T}$ with N clips, each denoted as s_i, and that the binary relevance between a

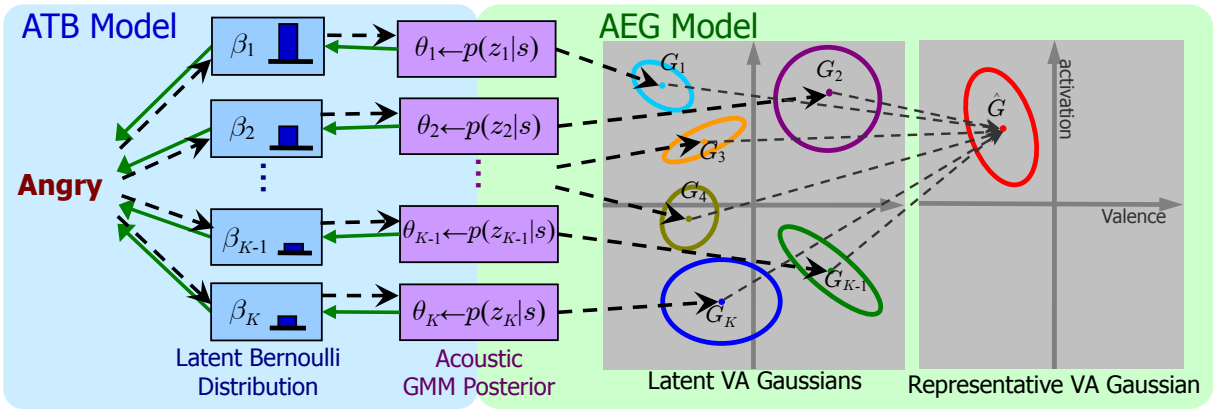


Figure 2: Illustration of the generation flow between tag-based and VA-based emotion semantics of music. Two component models, namely ATB and AEG, are shown in the left and right panels, respectively. The representative VA Gaussian of a tag can be generated by following the black dashed arrows.[4]

tag (for example "angry") and z_k can be modeled by a latent Bernoulli distribution with parameter β_k as follows,

$$p(y_i|z_k, \beta_k) = \beta_k^{y_i}(1-\beta_k)^{(1-y_i)}, \tag{4}$$

where $y_i \in \{0, 1\}$ is the binary tag label of s_i, and β_k describes the probability that z_k is associated to "angry." Then, the likelihood of "angry" for s_i is generated from the acoustic GMM posterior of s_i (green arrows) as follows,

$$p(y_i|\boldsymbol{\theta}_i, \boldsymbol{\beta}) = \sum_{k=1}^{K} \theta_{ik} p(y_i|z_k, \beta_k), \tag{5}$$

where $\boldsymbol{\beta}$ represents $\{\beta_k\}_{k=1}^K$. By taking the logarithm of Eq. 5 and summing over all clips in $\mathcal{X}_\mathcal{T}$, we have the cumulative log-likelihood as

$$\log p(\mathbf{Y}|\mathcal{X}_\mathcal{T}, \boldsymbol{\beta}) = \sum_{i=1}^{N} \log \sum_k \theta_{ik} p(y_i|z_k, \beta_k), \tag{6}$$

where $\mathbf{Y}$ denotes $\{y_i\}_{i=1}^N$ for "angry." We can maximize Eq. 6 using the EM algorithm with respect to $\boldsymbol{\beta}$ [15]. In the E-step, the posterior probability of z_k for s_i is

$$p(z_k|y_i, \boldsymbol{\theta}_i, \boldsymbol{\beta}) = \frac{\theta_{ik}\beta_k^{y_i}(1-\beta_k)^{(1-y_i)}}{\sum_{h=1}^{K} \theta_{ih}\beta_h^{y_i}(1-\beta_h)^{(1-y_i)}}. \tag{7}$$

In the M-step, the update rule for $\boldsymbol{\beta}$ is

$$\beta_k' \leftarrow \frac{\sum_{i=1}^{N} p(z_k|y_i, \boldsymbol{\theta}_i, \boldsymbol{\beta}) y_i}{\sum_{i=1}^{N} p(z_k|y_i, \boldsymbol{\theta}_i, \boldsymbol{\beta})}. \tag{8}$$

The idea of the ATB model has been introduced in [17], which is known as the *CBA* model. In our implementation, we replace the codeword histogram in CBA with the acoustic GMM posterior to represent a clip's acoustic features.

2.3 Learning the AEG Model

The right hand side of Figure 2 shows the AEG model [18], in which we assume that each z_k has its corresponding effective area G_k in the VA space, and G_k can be modeled by a bivariate latent Gaussian distribution. We denote the mixture of latent VA Gaussians as the *VA GMM* hereafter. To cover the emotion perception of different subjects, typically a music clip is annotated by multiple subjects. Suppose we have an emotion annotated music corpus $\mathcal{X}_\mathcal{E}$ with M music clips,[5] in which the multiple emotion annotations of s_i are denoted as $\mathbf{e}_{ij}, j = 1, \ldots, U_i$, representing the annotation given by the j-th subject u_{ij}, and U_i is the number of subjects who have annotated s_i.

Similar to ATB, the likelihood of $\mathbf{e}_{ij}$ of s_i can be generated from the acoustic GMM posterior of s_i (black dashed arrows in Figure 2) by

$$p(\mathbf{e}_{ij}|u_{ij}, s_i, \boldsymbol{\theta}_i) = \sum_{k=1}^{K} \theta_{ik} \mathcal{N}(\mathbf{e}_{ij}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{9}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ denote the mean vector and covariance matrix of G_k (cf. Figure 2) to be learned.

The total log-likelihood over $\mathcal{X}_\mathcal{E}$ is derived by first summing the log-likelihood of each subject u_{ij} for s_i, and then summing the log-likelihood of each clip s_i,

$$L = \sum_{i=1}^{M} \sum_{j=1}^{U_i} \log \sum_k \theta_{ik} \mathcal{N}(\mathbf{e}_{ij}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{10}$$

Again, we apply the EM algorithm to maximize L in Eq. 10 with respect to $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ [15]. In the E-step, the posterior probability of z_k given $\mathbf{e}_{ij}$ for s_i is

$$p(z_k|\mathbf{e}_{ij}, \boldsymbol{\theta}_i) = \frac{\theta_{ik}\mathcal{N}(\mathbf{e}_{ij}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{h=1}^{K} \theta_{ih}\mathcal{N}(\mathbf{e}_{ij}|\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}. \tag{11}$$

In the M-step, we derive the following update rules,

$$\boldsymbol{\mu}_k' \leftarrow \frac{\sum_{i,j} p(z_k|\mathbf{e}_{ij}, \boldsymbol{\theta}_i)\mathbf{e}_{ij}}{\sum_{i,j} p(z_k|\mathbf{e}_{ij}, \boldsymbol{\theta}_i)}, \tag{12}$$

$$\boldsymbol{\Sigma}_k' \leftarrow \frac{\sum_{i,j} p(z_k|\mathbf{e}_{ij}, \boldsymbol{\theta}_i)(\mathbf{e}_{ij} - \boldsymbol{\mu}_k')(\mathbf{e}_{ij} - \boldsymbol{\mu}_k')^T}{\sum_{i,j} p(z_k|\mathbf{e}_{ij}, \boldsymbol{\theta}_i)}. \tag{13}$$

2.4 Combining the ATB and AEG Models

With the learned ATB and AEG models, which are respectively parameterized by $\{\beta_k\}_{k=1}^K$ and $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$, we can interpret the generative process from a tag to a representative VA Gaussian, as shown in Figure 2.

A mapped acoustic GMM posterior $\{\hat{\theta}_k\}_{k=1}^K$, i.e., $\hat{\boldsymbol{\theta}}$, for "angry" is first generated (black dashed arrows) using the set of latent Bernoulli distributions $\{\beta_k\}_{k=1}^K$ learned for the tag. The intuition is that, for example, if z_1 dominates the latent Bernoulli distributions, i.e., $\beta_1 = 1$, and $\{\beta_h = 0\}_{h=2}^K$, this will lead to $\hat{\theta}_1 = 1$, and $\{\hat{\theta}_h = 0\}_{h=2}^K$. The process of

[4]Note that in this paper, the horizontal and vertical axes of the figures that present the VA space correspond to valence and activation, respectively.

[5]Note that $\mathcal{X}_\mathcal{T}$ and $\mathcal{X}_\mathcal{E}$ can be separate corpora consisting of different music clips.

estimating $\hat{\boldsymbol{\theta}}$ is called the "Fold-In" process, which is widely used in text information retrieval. Such a Fold-In process can be derived by maximizing the following log-likelihood,

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p(y = 1|\boldsymbol{\theta}, \boldsymbol{\beta}), \tag{14}$$

where $p(y = 1|\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{k=1}^{K} \theta_k \beta_k$. Eq. 14 can be solved using the EM algorithm. The E-step computes the posterior probability of z_k by

$$p(z_k|y = 1, \boldsymbol{\beta}) = \frac{\theta_k \beta_k}{\sum_{h=1}^{K} \theta_h \beta_h}. \tag{15}$$

The update rule leads to $\hat{\theta}_k \leftarrow p(z_k|y = 1, \boldsymbol{\beta})$ in the M-step for the next iteration. The estimated $\hat{\boldsymbol{\theta}}$ can be regarded as a summarized representation of "angry" on $\mathcal{X}_{\mathcal{T}}$ in the space of acoustic GMM posterior, and serves as the *bridge* between the tag and VA semantic modalities of music emotion.

When a proper $\hat{\boldsymbol{\theta}}$ has been estimated for "angry," to map "angry" into the VA space, we need to derive a single representative VA Gaussian $\hat{G}$ for $\sum_k \hat{\theta}_k G_k$, i.e., the weighted VA GMM, as illustrated in the rightmost part of Figure 2. This can be resorted to the information theory to estimate the mean vector $\hat{\boldsymbol{\mu}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$ of $\hat{G}$ by solving the following optimization problem,

$$\mathcal{N}(\mathbf{e}|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \underset{\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}}{\operatorname{argmin}} \sum_k \hat{\theta}_k D_{\mathrm{KL}}(\mathcal{N}(\mathbf{e}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \| \mathcal{N}(\mathbf{e}|\boldsymbol{\mu}, \boldsymbol{\Sigma})), \tag{16}$$

where $D_{\mathrm{KL}}(\mathcal{N}_A \| \mathcal{N}_B)$ denotes the one-way relative entropy (i.e., the KL divergence) from $\mathcal{N}_A$ to $\mathcal{N}_B$. The optimal mean vector and covariance matrix for Eq. 16 are [19]

$$\hat{\boldsymbol{\mu}} = \sum_{k=1}^{K} \hat{\theta}_k \boldsymbol{\mu}_k, \tag{17}$$

$$\hat{\boldsymbol{\Sigma}} = \sum_{k=1}^{K} \hat{\theta}_k \left(\boldsymbol{\Sigma}_k + (\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}})^T \right). \tag{18}$$

The estimated mean in Eq. 17 can be used for allocating the tag "angry" in the VA space and the estimated covariance in Eq. 18 implies the confidence for the tag location.

2.4.1 Remark

As the proposed framework is in nature content-based, there are several advantages. First, in principle, any tag (not limited to emotion-related ones) can be projected to the VA space, as long as there are some exemplar songs for that tag. Second, music examples with similar $\boldsymbol{\theta}$s can share their label information during the model learning, and this would facilitate the generalization of model learning. Third, the framework can be utilized to study the importance of different acoustic features for the tripartite relationship among music signal, tag, and emotion dimensions.

3. EVALUATION

In this section, we first introduce the two annotated music corpora and acoustic features used in this work. Then, we describe how we implement the framework, and demonstrate the mapped VA Gaussians for a number of tags and the generated tag clouds. Finally, we compare the generated tag clouds against those created by psychologists [20, 21].

3.1 Music Corpora

The first corpus, MER60, consists of 60 pieces of 30-second clips that come with both VA annotations and mood tag labels [14]. Each clip was selected from the chorus section of a unique English pop song, and annotated by 40 subjects for VA values. For the mood tag labels, we queried Last.fm with artist names and song titles for all the tags that are assigned to the 60 clips by online users. We then filtered the results and kept 50 mood tags that appear in the All Music Guide (AMG)'s mood taxonomy, which has been defined by professional music editors. Although MER60 is viewed as a jointly annotated corpus in the following evaluation setting, the mood tags and VA annotations of each clip actually originate from different sets of users.

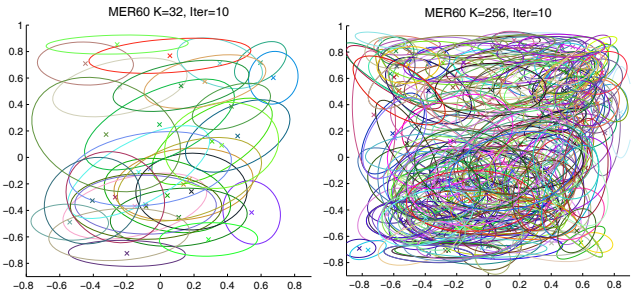


Figure 3: Demonstration of two VA GMMs learned on MER60 at iteration 10. Each component Gaussian is represented by an ellipse with a unique color.

The second corpus, AMG1644, is compiled by crawling the "top songs" listed in AMG for 33 mood classes, most of which constitute the mood taxonomy adopted by the mood classification task of MIREX [2]. Given the artist names and song titles, we fetched the corresponding 30-second audio previews from the media delivery company 7digital. There are 1,644 clips in AMG1644. The VA values are unavailable.

3.2 Frame-based Acoustic Features

In this work, we adopt the bag-of-frames feature representation and extract features from each short-time frame of a music clip, as a frame captures detailed acoustic features of music. As different emotion perceptions may be associated with different patterns of music features [4], we use MIRToolbox 1.3 [22] to extract the following four types of frame-based acoustic features: *dynamic*, *spectral*, *timbre*, and *tonal* features. All of the frame-based features are extracted with the same frame size of 50ms and 50% hop size. A frame is represented by a 70-D vector that is constructed by concatenating all the four types of features.

3.3 Model Fitting and Evaluation Results

We need to learn three models from the data, including acoustic GMM, ATB and AEG. We first use all the frames in MER60 to form the global frame vector set $\mathcal{F}$ (in total about 72K 70-D frame vectors) to train the acoustic GMMs. Then, we convert all clips into $\boldsymbol{\theta}$s using the acoustic GMM (cf. Eqs. 2 and 3). For ATB, we adopt the same implementation reported by Hoffman *et al.* [17] for model learning. As for AEG, we adopt the same learning specification in [18].

The following two settings are adopted in our experiments. First, in the MER60-alone case, the acoustic GMM, ATB, and AEG are learned from MER60 only, as this dataset has both tag-based and VA-based annotations. Second, in the AMG1644+MER60 case, the acoustic GMM and AEG are learned from MER60, but ATB is fitted with AMG1644. Note that the purpose of our experiments is to evaluate the quality of the mapping between the categorical mood tag space and the dimensional VA space, so this is not a train/test set splitting scenario. In both settings, the categorical-dimensional mapping is learned in an unsuper-

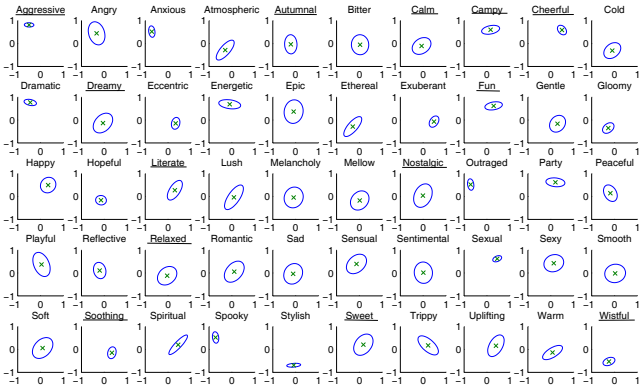


Figure 4: The VA Gaussians of 50 Last.fm mood tags for the MER60-alone setting. Both the ATB and AEG models are learned with $K = 256$ on MER60.

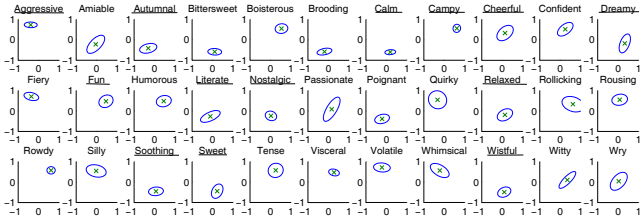


Figure 5: The VA Gaussians of 33 AMG mood tags for the MER60+AMG1644 setting. The ATB and AEG models are learned with $K = 32$ on AMG1644 and MER60, respectively.

vised fashion, in the sense that *we do not know a prior which VA region a tag should be mapped to*.

Figure 3 illustrates two learned VA GMMs (K=32 and K=256) for the MER60-alone setting, from which we observe that different latent VA Gaussians correspond to different areas and the VA GMM cover almost all the VA space. It seems that the learned VA GMM is able to generate all kinds of representative VA Gaussians by using different $\boldsymbol{\theta}$s.

Figures 4 and 5 show the representative VA Gaussians for a number of tags for the MER60-alone and the AMG1644+ MER60 settings, respectively. We use a green cross and a blue ellipse to denote the center and the covariance of the VA Gaussian for each tag. It can be observed that the categorical-dimensional mapping is fairly reasonable. For example, the tags Aggressive, Campy, Wistful, and Calm are mapped to the four different quadrants, as we may have expected. For affective terms with clear semantic meaning, the radius of the associated covariance is also small, such as Outraged and Sexual. Generally, there seems to be an inverse correlation between the *emotion intensity* (distance from the origin) and the covariance (or uncertainty) of a tag. Moreover, by comparing the location of the overlapping tags[6] between MER60 and AMG1644 in Figures 4 and 5, it can be found that the results for the two settings are reasonably consistent, except Literate and Sweet.

To offer a quantitative evaluation, we compare the results in Figures 4 and 5 with those proposed by psychologists. Specifically, we refer to the VA values of 30 affective terms, which are also used in this work, proposed by Whissell and Plutchik (WP) [20] and by the Affective Norms for English Words (ANEW) [21].[7] For each tag, the deviation of VA values between the one estimated by our method and the one defined by the psychologists can be quantified by computing the Euclidean distance. In addition, as a baseline method, we set the estimated VA values of each tag to the origin (representing a non-effective tag mapping likely given by a method) and also measure the VA deviation. One can interpret the baseline as the average emotion intensity of all the tags given by psychologists. From Table 1 we can see that the average VA deviations of tags derived by our method are significantly smaller than that of the baseline, indicating that the proposed framework is indeed effective.[8]

Table 1: The comparison of VA deviation in terms of Euclidean distance (the smaller the better).

Method	WP	ANEW
VA intensity (baseline)	0.615	0.549
Proposed framework	**0.327**	**0.303**

3.4 Evaluation of Generated Mood Tag Clouds

In Figure 6 (a) and (b), we show two mood tag clouds derived respectively from the results in Figures 4 and 5 using an open source graph visualization software, Graphviz [23]. Specifically, we initialize the position of each tag at its Gaussian center (cf. Eq. 17), and then use the Voronoi diagram-based heuristic to avoid overlapping.[9] The font size is set in proportion to the probability density at the Gaussian center. In other words, the font size of a mood tag would be larger if the covariance is smaller (i.e., more specific). Similarly, we also use Graphviz to visualize the tags with the corresponding VA values proposed by WP and ANEW in Figure 6 (c). Since we do not have the variance information of these psychologist tags, each tag is shown with equal size. By comparing the subfigures in Figure 6, it can be seen that the automatically generated tag clouds present reasonably matched results with that of psychologists.

4. DISCUSSION

Due to the space limitation, we only show the results for certain K values. From our empirical observation, the estimated mean vector of a tag VA Gaussian is generally not sensitive to the value of K, though the covariance seems to increase along with K for the MER60+AMG1644 setting.

The observed inverse correlation between the VA intensity and the covariance of a tag in fact implies a way to detect whether an arbitrary tag (e.g., Jazz, Rock, Male, or Guitar) has a strong emotional meaning or not. Our ongoing work has confirmed that a non-emotion-related tag would be mapped to the central area of the VA space, with a large covariance (or a small font when represented in a tag cloud). Such a function for detecting non-emotion-related tags is not available for conventional dimension reduction techniques.

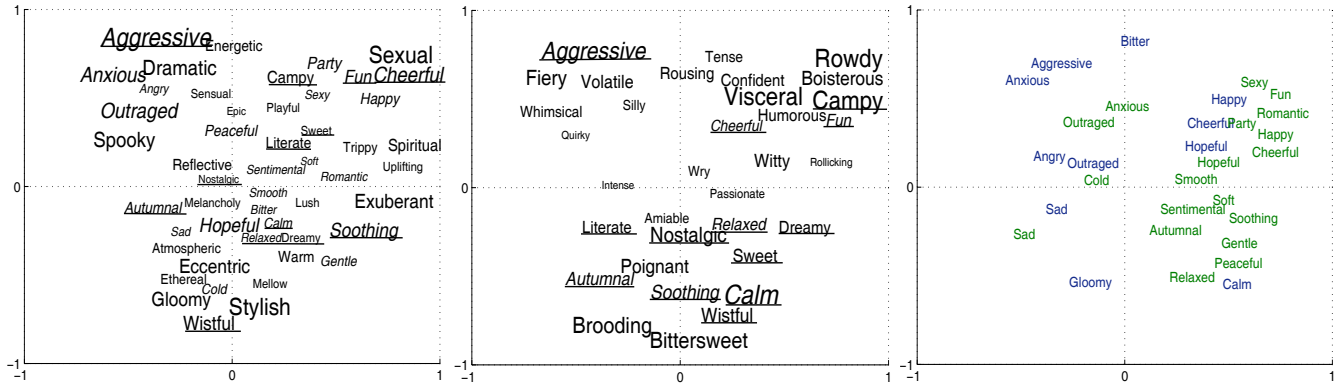
This paper does not focus on the effectiveness of the ATB and AEG models, which in fact have been well demonstrated in the previous work. For example, ATB has won the first

[6]The 13 overlapping tags are underlined in Figures 4, 5, 6(a) and 6(b). They are Aggressive, Autumnal, Calm, Campy, Cheerful, Dreamy, Fun, Literate, Nostalgic, Relaxed, Soothing, Sweet, and Wistful.

[7]11 and 19 tags are respectively from WP and ANEW.

[8]Admittedly, the baseline method considered in this evaluation is overly simple. Our future work will include other methods that are more reasonable.

[9]The tag position originated from its VA Gaussian may be shifted slightly by the Graphviz software to avoid overlapping with other tags in the tag cloud. Therefore, the valence and activation axes in Figure 6 could be slightly rotated or re-scaled.



(a) Last.fm tag cloud generated from the result of Figure 4.

(b) AMG tag cloud generated from the result of Figure 5.

(c) Reference tag cloud with the affective terms created by psychologists.

Figure 6: Demonstration of three tag cloud layouts generated by GraphViz.[10]

title (AUC Clip) in Mood Tag Classification of MIREX 2009 [17], and AEG has outperformed its rivals such as SVR and MLR for automatic emotion regression [14, 18, 24]. It seems that the two models are adequate for modeling the relationship between acoustic features and emotion labels/values, although more investigations for better feature representations that capture the nuance of emotion perception are still needed to further boost the prediction accuracy.

5. CONCLUSION

In this paper, we have proposed a novel probabilistic framework that unifies the categorical and dimensional emotion semantics of music. By making use of the acoustic GMM posterior, two emotion modalities are aligned to each other in a content-based fashion. The proposed learning-based framework is *scalable* and can do better if more annotated data is available. Based on this framework, we have demonstrated how to map an arbitrary tag to the VA space as a 2-D Gaussian distribution. In addition, we have generated two mood tag clouds respectively for two datasets and shown that the resulting visualization of mood tags bears high resemblance to the ones proposed by psychologists. We have also reported qualitative evaluations that further validate the effectiveness of the proposed method. For future work, we will conduct a user study of the tag-based music retrieval interface powered by the proposed framework.

6. REFERENCES

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

[2] X. Hu, *et al.* 2008. The 2007 MIREX audio mood classification task: Lessons learned. In *Proc. ISMIR*.

[3] C. Laurier, M. Sordo, J. Serrà, P. Herrera. 2009. Music mood representations from social tags. In *Proc. ISMIR*.

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

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

[10] In subfigures (a) and (b), the location and font size of each tag are set based on the parameters of the VA Gaussian of the tag. Subfigure (c) shows the tags of psychologists, in which terms in blue and green are proposed by WP [20] and ANEW [21], respectively. The overlapping terms from psychologists are italicized in subfigures (a) and (b).

[6] Y. E. Kim, *et al.* 2010. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, pp. 255–266.

[7] J. A. Russell. 1980. A circumplex model of affect. *J. Personality and Social Science*, 39, 6, pp. 1161–1178.

[8] J. A. Sloboda and P. N. Juslin. 2011. Psychological Perspectives on Music and Emotion. *Music and Emotion: Theory and Research*, New York: Oxford University Press.

[9] E. Pampalk, *et al.* 2002. Content-based organization and visualization of music archives. In *Proc. ACM Multimedia*.

[10] C. M. Whissel. 1989. The dictionary of affect in language. *Emotion: Theory, Research and Experience: Vol. 4, The Measurement of Emotions*, New York: Academic.

[11] R. Plutchik. 1980. *Emotion: A Psychoevolutionary Synthesis*, New York: Harper & Row.

[12] M. Zentner, D. Grandjean and K. R. Scherer. 2008. Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion*, 8, 4, pp. 494–521.

[13] Y.-H. Yang and X. Hu. 2012. Cross-cultural music mood classification: A comparison of English and Chinese songs. In *Proc. ISMIR*.

[14] Y.-H. Yang and H. H. Chen. 2011. Predicting the distribution of perceived emotions of a music signal for content retrieval. *IEEE TASLP*, 19, 7, pp. 2184–2196.

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

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

[17] M. Hoffman, D. Blei and P. Cook. 2009. Easy as CBA: A simple probabilistic model for tagging music. In *Proc. ISMIR*, pp. 369–374.

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

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

[20] R. Cowie, *et al.* 2001. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 18, 1, pp. 32–80.

[21] M. Bradley and P. J. Lang. 1999. Affective norms for English words *ANEW*: Instruction manual and affective ratings. *Technical Report C-1*. The Center for Research in Psychophysiology, University of Florida.

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

[23] J. Ellson, *et al.* 2001. Graphviz – Open source graph drawing tools. In *Proc. Graph Drawing*, pp. 483–484.

[24] E. M. Schmidt and Y. E. Kim. 2010. Prediction of time-varying musical mood distributions from audio. In *Proc. ISMIR*, pp. 465–470.