Personalized Music Emotion Recognition via Model Adaptation

Ju-Chiang Wang*†, Yi-Hsuan Yang†‡, Hsin-Min Wang†‡ and Shyh-Kang Jeng*
*Department of Electrical Engineering, National Taiwan University, Taiwan
†Institute of Information Science, Academia Sinica, Taiwan
‡Research Center for Information Technology and Innovation, Academia Sinica, Taiwan
E-mail: {asriver, yang, whm}@iis.sinica.edu.tw; skjeng@cc.ee.ntu.edu.tw;

***Abstract*—In the music information retrieval (MIR) research, developing a computational model that comprehends the affective content of music signal and utilizes such a model to organize music collections have been an essential topic. Emotion perception in music is in nature subjective. Consequently, building a general emotion recognition system that performs equally well for every user could be insufficient. In contrast, it would be more desirable for one's personal computer/device being able to understand his/her perception of music emotion. In our previous work, we have developed the acoustic emotion Gaussians (AEG) model, which can learn the broad emotion perception of music from general users. Such a general music emotion model, called the background AEG model in this paper, can recognize the perceived emotion of unseen music from a general point of view. In this paper, we go one step further to realize the personalized music emotion modeling by adapting the background AEG model with a limited number of emotion annotations provided by a target user in an online and dynamic fashion. A novel maximum a posteriori (MAP)-based algorithm is proposed to achieve this in a probabilistic framework. We carry out quantitative evaluations on a well-known emotion annotated corpus, MER60, to validate the effectiveness of the proposed method for personalized music emotion recognition.**

I. Introduction

State-of-the-art systems for speaker recognition are usually built upon two models: a large Gaussian mixture model (GMM), also called the Universal Background Model (UBM) [1], that is trained to represent the speaker-independent distribution of acoustic features, and a speaker-dependent GMM that is obtained by updating the parameters of the UBM via the model adaptation techniques with the speech data of a specific speaker, who is interacting with the system [2]. This system design has been proved successful as it captures both the commonality among general speakers and the individuality of the target speaker.

In this paper, we propose to apply the idea of GMM-UBM and model adaption to the challenging task of automatic music emotion recognition (MER), which has received increasing attention in recent years [3], [4], [5].[1] MER is considered important as it holds the promise of managing the ever increasing volume of digital music in a content-based and intuitive way. However, due to the complicated mental processes involved in the perception of music, MER is different from conventional pattern recognition tasks in that oftentimes emotion perception is fairly *user-dependent* [6], [7]. For example, heavy metal music can be pleasant to some people, yet annoying to others. The subjective nature of emotion perception indicates the requirement for personalizing the MER system [8], [9]. As argued in [10], although developing a general MER system that performs equally well for every user is great, it is rather more sufficient if one's personal computer or mobile device is able to understand his or her perception of emotion and adapt to each individual in a dynamic and real-time fashion [11].

Despite that the subjective nature of emotion perception is well recognized, little effort has been invested to take the subjectivity into account. Most existing work avoids dealing with this issue by assuming a common consensus can be achieved (particularly for classic music) [12], discarding songs that a common consensus cannot be achieved [13], or simply leaving it as future work [14]. Although some preliminary attempts have been made to personalized MER, most of them are built upon discriminative models that lack a solid and theoretical computational framework [15], [16], [11]. The performance of existing MER systems are still limited from both theoretical and practical points of view.

For the model adaptation techniques to be applicable, a prerequisite is that the target pattern recognition task can be represented in a *parametric* form, such that model adaptation can be performed efficiently on-line by adapting the model parameters. A novel probabilistic model, which is outlined below, is developed to learn a UBM-like background model for music emotion, which is viewed as a parametric and probabilistic distribution over the so-called emotion space instead of static mood labels. Personalizing the background model can then be realized by adapting the parameters of this model. To the best of our knowledge, few attempts if any have been made to develop a principled probabilistic framework that has a sound statistical foundation in concern with the subjectivity issue and personalization scenario in emotion-based music information systems.

A. The Acoustic Emotion Gaussians Model

In our recent work, we have proposed a novel *Acoustic Emotion Gaussians* (AEG) model that realizes the generative

[1] We define music emotion as the emotion human *perceives* as being expressed in a piece of music, rather than the emotion *felt* in response to the piece. This distinction is made because we may not feel sorrow when listening to a sad tune [6].

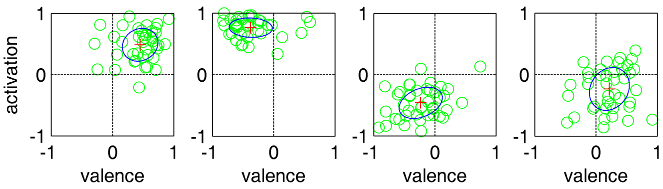

Fig. 1. Subjects' annotations in the VA space [18] for four 30-second clips, which from left to right are *Dancing Queen* by ABBA, *Civil War* by Guns N' Roses, *Suzanne* by Leonard Cohen, and *All I Have To Do Is Dream* by the Everly Brothers, respectively. Each circle corresponds to a subject's annotation for a clip, and the overall annotations (in total 40) for a clip can be modeled by a 2-D Gaussian distribution (i.e., the blue ellipse) [19].

process of emotion perception in music from acoustic features [17]. The AEG model learns from data two Gaussian mixture models (GMMs), namely an *acoustic GMM* and a *VA GMM*, to describe the low-level acoustic feature space and high-level emotion space, respectively. A set of *latent feature classes* is introduced to play the end-to-end linkage between the two spaces and align the two GMMs. As a principled probabilistic model, AEG is applicable to both emotion-based music annotation (i.e., MER) and retrieval.

Specifically, to better account for the subjective and stochastic natures of emotion perception, the proposed AEG model represents the perceived emotion of music as a mixture of *bivariate Gaussian distributions* in a two-dimensional emotion space spanned by *valence* (or pleasantness; positive/negative affective states) and *activation* (or arousal; energy and stimulation level) – the two most fundamental dimensions found by psychologists [18].[2] The valence-activation space is referred to as the VA space in this paper hereafter. Figure 1 shows the ground truth VA annotations of four music clips, each labeled by multiple subjects. We can see that the annotations for each clip appear to be approximately expressed by a 2D Gaussian. Therefore, we can learn an AEG model from this type of emotion annotations and utilize it to predict the emotion distribution for a music clip as a 2-D Gaussian. In this way, developers of an emotion-based music retrieval system can better understand how likely a specific emotional expression (expressed as a VA-based probabilistic distribution) would be elicited when listening to a clip.

B. Personalizing the AEG Model via Model Adaptation

As the AEG model is parametric, it can be easily extended to incorporate additional user information, such as individual emotion perception survey, personal profile, purchasing records, and listening history, for personalization. Therefore, in this paper, we go one step further to realize the personalization scenario for the AEG model. Due to the use of VA GMM in modeling the VA annotations underlying the AEG framework, we can apply the GMM-based adaptation methods to personalizing the AEG model with a person's annotations. Specifically, we first treat the VA GMM learned from broad subjects as a background emotion model, and then derive the *maximum a posteriori* (MAP) [20] based method to adapt the background (general) VA GMM using a small number of user-provided annotations in an on-line fashion. In practice, the quantity of personal annotations is usually sparse, and sometimes only one annotation is available at a time instance. Therefore, it is preferable to incrementally adapt the model parameters.[3]

The remainder of the paper is organized as follows. Section II introduces the generative process of AEG as well as the model learning and emotion predicting procedures underlying the AEG framework. Sections III describes the technical details of the model adaptation method. The corpus, evaluation setup, and metric used in this work and the evaluation results are presented in Section IV. Finally, we conclude the paper in Section V.

II. The Acoustic Emotion Gaussians Model

This section introduces the AEG model and presents how to apply it to the VA-based emotion prediction of a music clip, as illustrated in Figure 2.

A. Acoustic GMM Posterior Representation

To start the generative process of the AEG model, we utilize a universal acoustic GMM to span the probabilistic space for the acoustic GMM posterior representation. The acoustic GMM, which is pre-learned using the EM algorithm [21] on a universal set of frame-based acoustic feature vectors, $\mathcal{F}$, is expressed 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.

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

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

In our implementation, the mixture prior (i.e., π_k and π_h) in Eq. 2 is replaced by $\frac{1}{K}$, because it was not useful in the previous work [22].

The clip-level acoustic GMM posterior $\{\theta_{ik}\}_{k=1}^{K}$ (cf. Figure 2) can be summarized by

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

[3] It can be noted that because the MAP-based method is content-based, the target user can choose whatever clips to annotate, for example, songs that he/she is familiar with, not limited to the ones utilized in training the background model.

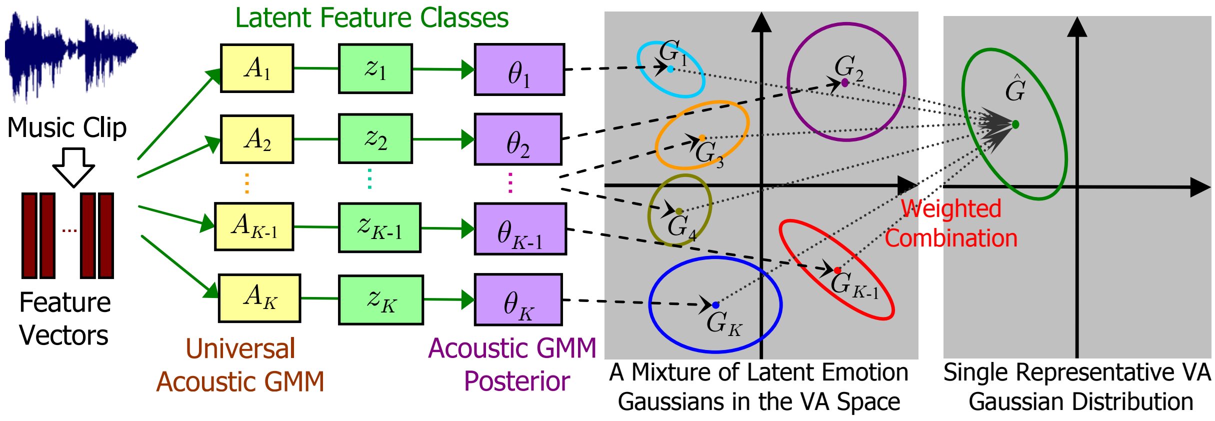

Fig. 2. Illustration of the generative process of the AEG model. Music emotion distribution can be generated from the acoustic features.

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

One can see from Eq. 3 that a component acoustic posterior probability θ_k 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. Finally, the acoustic GMM posterior of s_i is represented by vector $\boldsymbol{\theta}_i$, whose k-th component is θ_{ik}.

B. User Prior Model for VA Annotation

To cover the emotion perception of different subjects, typically each clip s_i in $\mathcal{X}$ is annotated by multiple subjects. Given the emotion annotations $\mathbf{e}_{ij}, j = 1, \ldots, U_i$, of s_i, where $\mathbf{e}_{ij}$ denotes the annotation given by the j-th subject u_{ij} and U_i denotes the number of subjects who have annotated s_i, we build a user prior model γ with the following Gaussian distribution,

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

where $\mathbf{a}_i = \frac{1}{U_i}\sum_{j=1}^{U_i}\mathbf{e}_{ij}$ and $\mathbf{B}_i = \frac{1}{U_i}\sum_{j=1}^{U_i}(\mathbf{e}_{ij} - \mathbf{a}_i)(\mathbf{e}_{ij} - \mathbf{a}_i)^T$. The annotation prior of $\mathbf{e}_{ij}$ can be estimated based on the likelihood computed by Eq. 4. Therefore, if an annotation is far away from other annotations for the same clip, it would be considered less reliable.

To be incorporated into the learning process of the VA GMM, the annotation prior probability of an annotation of a clip (denoted by γ_{ij}) is derived from its corresponding user prior model $\gamma(\mathbf{e}_{ij}|u_{ij}, s_i)$ as follows

$$\gamma_{ij} \leftarrow p(u_{ij}, s_i|\mathcal{X}) = \frac{\gamma(\mathbf{e}_{ij}|u_{ij}, s_i)}{\sum_{q=1}^{N}\sum_{r=1}^{U_q}\gamma(\mathbf{e}_{qr}|u_{qr}, s_q)}. \tag{5}$$

C. Learning the VA GMM

In the right hand side of Figure 2, each z_k maps an audio pattern into an area G_k in the VA space, where G_k can be modeled by a bivariate Gaussian distribution, denoted as a latent VA Gaussian. The mixture of latent VA Gaussians is called the VA GMM hereafter. To infer the VA GMM, we assume that $\mathbf{e}_{ij}$ of s_i by u_{ij} in $\mathcal{X}$ can be generated from a weighted VA GMM governed by the acoustic GMM posterior $\boldsymbol{\theta}_i$ of s_i,

$$p(\mathbf{e}_{ij}|u_{ij}, s_i, \boldsymbol{\theta}_i) = \sum_{k=1}^{K}\theta_{ik}\mathcal{N}(\mathbf{e}_{ij}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{6}$$

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

For each clip s_i, we have computed its acoustic GMM posterior $\boldsymbol{\theta}_i$ and annotation prior γ_{ij}. The clip-level likelihood is generated by a weighted sum over all subjects who have annotated the clip, and the total likelihood is generated by the weighted sum of the clip-level likelihoods over all clips in $\mathcal{X}$ as follows,

$$p(\mathbf{E}|\mathcal{X}) = \sum_i\sum_j\gamma_{ij}\sum_k\theta_{ik}\mathcal{N}(\mathbf{e}_{ij}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{7}$$

where $\sum_i\sum_j\gamma_{ij} = 1$, and $\mathbf{E}$ denotes $\{\mathbf{e}_{ij}\} \in \mathcal{X}, \forall i, j$. According to the Jensen's inequality, the logarithm of Eq. 7 has the following property,

$$\log p(\mathbf{E}|\mathcal{X}) \geq \sum_{i,j}\gamma_{ij}\log\sum_k\theta_{ik}\mathcal{N}(\mathbf{e}_{ij}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{8}$$

We apply the EM algorithm to maximize Eq. 8 with respect to $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ [21]. In the E-step, the posterior probability of z_k given a subject's annotation for s_i is

$$p(z_k|\mathbf{e}_{ij}, \boldsymbol{\theta}_i) = \frac{\theta_{ik}\mathcal{N}(\mathbf{e}_{ij}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{h=1}^{K}\theta_{ih}\mathcal{N}(\mathbf{e}_{ij}|\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}. \tag{9}$$

In the M-step, we derive the following update rules [17]:

$$\boldsymbol{\mu}'_k \leftarrow \frac{\sum_{i,j}\gamma_{ij}p(z_k|\mathbf{e}_{ij}, \boldsymbol{\theta}_i)\mathbf{e}_{ij}}{\sum_{i,j}\gamma_{ij}p(z_k|\mathbf{e}_{ij}, \boldsymbol{\theta}_i)}, \tag{10}$$

$$\boldsymbol{\Sigma}'_k \leftarrow \frac{\sum_{i,j}\gamma_{ij}p(z_k|\mathbf{e}_{ij}, \boldsymbol{\theta}_i)(\mathbf{e}_{ij} - \boldsymbol{\mu}'_k)(\mathbf{e}_{ij} - \boldsymbol{\mu}'_k)^T}{\sum_{i,j}\gamma_{ij}p(z_k|\mathbf{e}_{ij}, \boldsymbol{\theta}_i)}. \tag{11}$$

D. Emotion Prediction

Given a test music clip with the acoustic GMM posterior $\hat{\boldsymbol{\theta}}$, the AEG model can generate the predicted emotion distribution as a GMM $\sum_k \hat{\theta}_k \mathcal{N}(\mathbf{e}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ is the learned VA GMM. However, the resulting predicted VA GMM may be unnecessarily complicated and difficult for a user to interpret the result of emotion prediction. Instead, a single and representative VA Gaussian $\mathcal{N}(\mathbf{e}|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ is practically more useful, since there is only one set of Gaussian parameters that makes it straightforward to comprehend the predicted emotion. The representative VA Gaussian can be derived by the weighted combination of all latent VA Gaussians as shown in the rightmost part in Figure 2. This can be resorted to the information theory to calculate the mean vector and covariance matrix of the representative VA Gaussian by solving the following optimization problem,

$$\begin{aligned}&\mathcal{N}(\mathbf{e}|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \\ &\underset{\{\boldsymbol{\mu},\boldsymbol{\Sigma}\}}{\operatorname{argmin}} \sum_{k=1}^{K} \hat{\theta}_k D_{\text{KL}}(\mathcal{N}(\mathbf{e}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \| \mathcal{N}(\mathbf{e}|\boldsymbol{\mu}, \boldsymbol{\Sigma})),\end{aligned} \tag{12}$$

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

$$\begin{aligned}D_{\text{KL}}(\mathcal{N}_A \| \mathcal{N}_B) =& \frac{1}{2}\left(\text{tr}(\boldsymbol{\Sigma}_A \boldsymbol{\Sigma}_B^{-1}) - \log |\boldsymbol{\Sigma}_A \boldsymbol{\Sigma}_B^{-1}|\right) \\ &+ \frac{1}{2}(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)^T \boldsymbol{\Sigma}_B^{-1} (\boldsymbol{\mu}_A - \boldsymbol{\mu}_B) - 1,\end{aligned} \tag{13}$$

The optimal mean vector and covariance matrix for Eq. 12 are obtained by [23]:

$$\hat{\boldsymbol{\mu}} = \sum\nolimits_{k=1}^{K} \hat{\theta}_k \boldsymbol{\mu}_k, \tag{14}$$

$$\hat{\boldsymbol{\Sigma}} = \sum\nolimits_{k=1}^{K} \hat{\theta}_k \left(\boldsymbol{\Sigma}_k + (\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}})^T\right). \tag{15}$$

III. Personalizing the AEG Model

So far the VA GMM learned from quite a few of subjects and their corresponding annotated music clips can be a sufficient representation for general user opinions as a background model. Similar to the idea of the GMM-UBM speaker recognition system, we treat the background VA GMM as a user-independent background model, which is considered having the well-trained parameters for generalizing the emotion modeling. When the annotated clips of a target (or new) user are available, the personalized VA GMM can be adapted from the background VA GMM. Motivated by speaker adaption in speech recognition, we adopt the *maximum a posteriori* (MAP) [20], [2] criterion derived from the Bayesian learning theory as the GMM-based adaptation method. The MAP-based approach is regarded as fairly efficient in speaker adaptation and as having a tight coupling between the personalized model and UBM, without the loss of model generalizability. It is therefore an ideal candidate for online personalization of emotion-based MIR applications.

To personalize the AEG model, the system may ask a target user u_* to annotate a few number of music clips in advance and then uses the personal annotations to adapt the background VA GMM. We are given a pre-trained background VA GMM denoted as $\{\mathcal{N}(\mathbf{e}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K$, and a set of music clips and their corresponding VA values $\{\mathbf{e}_m, \boldsymbol{\theta}_m\}_{m=1}^M \in \mathcal{X}_*$ that the target user u_* has rated. Since the VA GMM models the emotion annotations based on the acoustic GMM posterior that is generated from the fixed acoustic GMM, the music clips in $\mathcal{X}_*$ can be exclusive to the music corpus $\mathcal{X}$ used to learn the background VA GMM. This means that u_* is allowed to annotated his/her familiar music.

The first step of VA GMM adaptation is equivalent to the E-step of the EM algorithm that computes the posterior probabilities over z_k using $\mathcal{X}_*$,

$$p(z_k|\mathbf{e}_m, \boldsymbol{\theta}_m) = \frac{\theta_{mk} \mathcal{N}(\mathbf{e}_m|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{h=1}^{K} \theta_{mh} \mathcal{N}(\mathbf{e}_m|\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}. \tag{16}$$

Then, we derive the expected sufficient statistics of $\mathcal{X}_*$ over the posterior probability $p(z_k|\mathbf{e}_m, \boldsymbol{\theta}_m)$ for the effective number (weight), mean, and covariance parameters:

$$M_k = \sum\nolimits_{m=1}^{M} p(z_k|\mathbf{e}_m, \boldsymbol{\theta}_m), \tag{17}$$

$$E(\boldsymbol{\mu}_k) = \frac{\sum_{m=1}^{M} p(z_k|\mathbf{e}_m, \boldsymbol{\theta}_m)\mathbf{e}_m}{M_k}, \tag{18}$$

$$E(\boldsymbol{\Sigma}_k) = \frac{\sum_{m=1}^{M} p(z_k|\mathbf{e}_m, \boldsymbol{\theta}_m)\mathbf{e}_m\mathbf{e}_m^T}{M_k}. \tag{19}$$

Finally, the new parameters of the personalized VA GMM can be obtained according to the MAP criterion with a set of conjugate prior distributions [20]. The resulting update rules are the forms of interpolations between the expected sufficient statistics (i.e., $E(\boldsymbol{\mu}_k)$ and $E(\boldsymbol{\Sigma}_k)$) and the parameters of the background VA GMM (i.e., $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$) as follows:

$$\boldsymbol{\mu}_k^* \leftarrow \alpha_k^{\boldsymbol{\mu}} E(\boldsymbol{\mu}_k) + (1 - \alpha_k^{\boldsymbol{\mu}})\,\boldsymbol{\mu}_k, \tag{20}$$

$$\boldsymbol{\Sigma}_k^* \leftarrow \alpha_k^{\boldsymbol{\Sigma}} E(\boldsymbol{\Sigma}_k) + \left(1 - \alpha_k^{\boldsymbol{\Sigma}}\right)\left(\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T\right) - \boldsymbol{\mu}_k^*(\boldsymbol{\mu}_k^*)^T. \tag{21}$$

Note that there is no need to update the weight parameters since the mixture weights are replaced by the acoustic GMM posterior probabilities in emotion prediction. Personalizing the background VA GMM is very efficient because we only need to perform the adaptation procedure once. The complexity mainly depends on K times of computing expected sufficient statistics and updating the parameters.

The interpolation coefficients for updating the mean vector and covariance matrix in Eqs. 20 and 21 are data-dependent and defined as

$$\alpha_k^{\boldsymbol{\mu}} = \frac{M_k}{M_k + \delta^{\boldsymbol{\mu}}}, \quad \text{and} \quad \alpha_k^{\boldsymbol{\Sigma}} = \frac{M_k}{M_k + \delta^{\boldsymbol{\Sigma}}}, \tag{22}$$

where $\delta^{\boldsymbol{\mu}}$ and $\delta^{\boldsymbol{\Sigma}}$ are the fixed relevance factors for mean and covariance, respectively. The personalized VA GMM adapted from $\mathcal{X}_*$ can be used to personalize music emotion recognition mentioned in Section II-D. The personalized VA GMM can be also incorporated into the emotion-based music retrieval introduced in our recent work [17].

IV. Evaluation

This section presents the emotion annotated corpus, evaluation setup, and metrics used in this work. As for the performance study, we first evaluate general MER using the background AEG, and then investigate the performance of the personalized AEG for personalized MER in an incremental learning scenario.

A. Music Corpora

We use the MER60 corpus consisting of 60 clips that comes with VA annotations [19].[4] These clips, in which each is 30-second long, were selected from the chorus section of English pop songs, and each of them was annotated by 40 subjects for VA values. Each subject was asked to annotate the VA values by using a graphic interface that displays the VA space in a silent computer lab. The VA values, which are numerical values ranging from -1 to 1, are entered by clicking a point in the emotion space. Among the 40 users, 6 users have annotated all the clips. Therefore, we can evaluate the performance of personalization on these 6 users.

B. Frame-based Acoustic Features

In this work, we adopt the bag-of-frames modeling and extract frame-based musical features for acoustic modeling [22], [24], [25]. A frame that captures detailed temporal features can facilitate the ability of clip-level acoustic modeling of the acoustic GMM posterior representation. Instead of analyzing the emotion of a specific frame, we aggregate all the frames in a clip into the acoustic GMM posterior vector $\boldsymbol{\theta}$ (cf. Eq. 3) and perform our analysis of emotion at the clip level. Although it may be interesting to extract long-term mid-level features such as melody, rhythm, structure, or harmonic progression that directly characterizes the musical information of a clip, such features are not used because the extraction of them is still not perfect and they may introduce noises (in feature extraction) to the system.

We utilize MIRToolbox 1.3 [26] to extract the following four types of frame-based acoustic features: *dynamic* (root-mean-squared energy), *spectral* (centroid, spread, skewness, kurtosis, entropy, flatness, rolloff 85%, rolloff 95%, brightness, roughness, and irregularity), *timbre* (zero crossing rate, spectral flux, 13 MFCCs, 13 delta MFCCs, and 13 delta-delta MFCCs), and *tonal* (key clarity, key mode possibility, HCDF, 12-bin chroma, chroma peak, and chroma centroid). All of the frame-based features are extracted with the same frame size of 50ms and 50% hop size to ensure easy alignment. Each dimension in all extracted frame vectors is normalized to have zero mean and one standard deviation. Two frame vector representations are considered in the performance evaluation: a 39-D vector that consists of MFCC-related features only and a 70-D vector that concatenates all the features.

C. Evaluation of Background AEG for General MER

To learn the acoustic GMM, we use an external music collection to form the global frame vector set $\mathcal{F}$ containing 235K frames. Then, the acoustic GMMs with several K values are learned using the EM algorithm [21]. We restrict the covariance matrix of the acoustic GMM to be diagonal. To learn the respective VA GMM, we initialize all the VA Gaussian components with the sample mean vector and covariance matrix of the VA annotations of the training set, and use a full covariance matrix for each latent VA Gaussian component.

We perform six-fold cross-validation for the MER evaluation. That is, 50 clips are used for training and the remaining 10 clips are used for testing. Each set of ground truth annotations of a clip is summarized by a ground truth Gaussian to represent the general emotion perception. The error (prediction deviation) can be evaluated by the one-way KL divergence (cf. Eq. 13) between the predicted and ground truth Gaussians for a clip. The overall performance is evaluated in terms of the average KL divergence (AKL) over the test set. Smaller AKL corresponds to better performance.

The following factors in the background AEG are considered: the frame-based acoustic features (either 39-D MFCCs or 70-D concatenated features), the number of latent feature classes K, and whether to use the annotation prior described in Section II-B or not. For example, "AEG-APrior-70DConcat" means using the annotation prior with the 70-D concatenated features. We test the AEG model with K ranging from 16 to 1,024. When the annotation prior is not used, we simply replace all γ_{ij} by 1 in the learning process. We compare the AEG method with support vector regression (SVR) [27], which is regarded as one of the state-of-the-art methods widely used in the MER task [28], [19], with different acoustic features. The SVR-Melody method, which uses the melody features, was the best performed setting reported in [19].[5] We also investigate the performance of SVR using the acoustic features used in our method.

Figures 3 shows the AKL performance. It is clear that, as a general music emotion recognizer, the background AEG consistently outperforms the SVR method in almost all cases. Particularly, AEG-APrior-70DConcat (K=32) significantly outperforms SVR-Melody with p-value$<1\%$ under the two-tailed t-test. In general, the annotation prior model improves the performance, and the 70-D concatenated features outperform the 39-D MFCCs when K is small.

D. Evaluation for Personalized MER

So far we have demonstrated the effectiveness of the background AEG for general MER. Next, we conduct the personalized MER task in an incremental setting as the scenario described in Section I-B that the personal annotations of a target user are not available during training the background VA GMM, and more and more data are gradually available in the future. In the MER60 corpus, there are 6 users who

[4] http://mac.iis.sinica.edu.tw/ yang/MER/NTUMIR-60/

[5] Currently we do not consider the melody features for AEG, since they belong to the long-term mid-level features. We leave this for our future work.

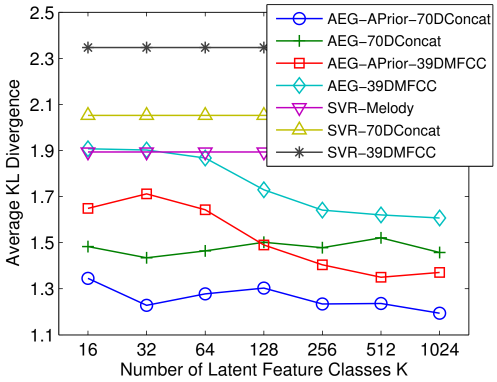

Fig. 3. Performance of general MER in terms of AKL (the smaller the better) evaluated on the MER60 corpus.

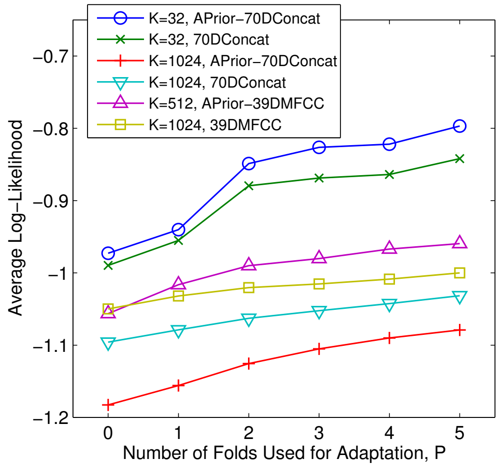

Fig. 4. Performance comparison (ALLi, the larger the better) of the average performance over the 6 users with 6 experimental settings (different K, acoustic features, and number of folds of personal annotations used for adaptation) for personalized MER on the MER60 corpus.

have annotated all the clips. Therefore, we train and test the personal MER task with the annotations of the 6 users.

For each test user, we perform six-fold cross-validation that holds out 1-fold of music clips for testing in each validation run. The remaining five folds are utilized for training the personalized model in an incremental way: the training data are available for model adaptation fold-by-fold (instead of one-by-one). In a nutshell, the experiment for a target user is organized as the following procedures:

1) Randomly split all the clips and the corresponding annotations of the user into $P+1$ folds (here $P=5$).
2) Perform $P+1$-fold cross-validation ($P+1$ validation trials):
    a) Hold out one fold unseen for testing.
    b) Train a background VA GMM with music clips in the rest P folds using the emotion annotations of all the subjects we have except for the target user.
    c) Add one fold into the adaptation pool until all the P folds are used (P incremental adaptation trials).
        i) Use the current adaptation pool to adapt the background VA GMM.
        ii) Evaluate the prediction accuracy for the clips in the test fold using the adapted VA GMM.
3) Summarize the performance for the target user.

Note that we have used P to denote the number of folds used for adaptation. The main purpose of the experiment is to investigate the effectiveness of model adaptation against the quantity of personal data available for the system. The performance of the personalized MER model can be evaluated by feeding the ground truth annotation of the target user into the predicted VA Gaussian, i.e., $\log \mathcal{N}(\mathbf{e}_* | \hat{\boldsymbol{\mu}}_*, \hat{\boldsymbol{\Sigma}}_*)$, where $\mathbf{e}_*$ is the ground truth annotation of user u_*, and $\{\hat{\boldsymbol{\mu}}_*, \hat{\boldsymbol{\Sigma}}_*\}$ is the predicted single VA Gaussian with the test music clip. The average log-likelihood (ALLi) over the test set is regarded as the final performance; larger ALLi corresponds to more accurate performance.

In our preliminary study, we have found that updating the covariance matrices of VA GMM may sometimes lead to the degradation of performance of model adaptation, possibly because the limited number of personal data we have may not be sufficient for estimating a proper covariance. This observation is also in line with the findings in speaker adaptation [2]. Therefore, in the experiment we only adapt the mean vector of the VA GMM according to the personal annotations. As a result of model adaptation, the positions of the component latent VA Gaussians may shift (c.f., Figure 2), but the shape and size will remain unchanged. The relevance factor $\delta^{\boldsymbol{\mu}}$ is empirically set to 0.1.

Figure 4 shows the result of personalized MER in terms of ALLi. We select 6 settings, which have performed well in the evaluation of background AEG, to investigate in the evaluation of personalized MER. In general, the performance gets improved in all cases as more personal annotations are available, i.e., the performance is positively proportional to the value of P. In particular, the performance difference of APrior-70DConcat with $K=32$ is significant (p-value$<5\%$ under the two-tailed t-test) between $P=0$ and $P=2$, indicating that the adaptation method can achieve a significant improvement with only 20 personal annotations. In summary, these observations demonstrate the effectiveness and efficiency of the proposed personalization method.

Two more observations can be made from the result. First, among the 6 settings, APrior-70DConcat with $K=32$ performs the best as suggested in the previous evaluation. In contrast, APrior-70DConcat with $K=1,024$ leads to the worst result, suggesting that a complicated model considering a larger K and more diverse acoustic features may not be useful. We attribute this to the limited number of adaptation data, which may be insufficient for updating the parameters of such a complicated model. Second, the annotation prior does not benefit the 70DConcat feature as K is large, possibly also because it increases the model complexity.

V. Conclusion

In this paper, we have presented a novel MAP-based adaptation technique for personalizing the AEG model. The performance study has also demonstrated the effectiveness of the proposed method for the personalized MER task in an incremental learning scenario. In the near future, we will investigate the *maximum likelihood linear regression* (MLLR) [29] based technique that learns a linear transformation over the parameters of the AEG model for personalization.

VI. Acknowledgement

This work was supported in part by the National Science Council of Taiwan under Grants: NSC 101-2221-E-001-019-MY3, NSC 101-2221-E-001-017, and NSC 101-2631-H-001-007 (TELDAP).

References

[1] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. European Conf. Speech Communication and Technology*, 1997.

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

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

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

[5] 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. Society for Music Information Retrieval Conference*, 2010, pp. 255–266.

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

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

[8] C.-C. Yeh, S.-S. Tseng, P.-C. Tsai, and J.-F. Weng, "Building a personalized music emotion prediction system." in *PCM*, ser. Lecture Notes in Computer Science. Springer, 2006, pp. 730–739.

[9] B. Zhu and T. Liu, "Research on emotional vocabulary-driven personalized music retrieval," in *Edutainment*, 2008, pp. 252–261.

[10] R. W. Picard, *Affective Computing*. The MIT Press, 1997.

[11] Y.-H. Yang, Y.-C. Lin, and H. H. Chen, "Personalized music emotion recognition," in *Proc. ACM Int. Conf. SIGIR*, 2009, pp. 748–749.

[12] M. Wang, N. Zhang, and H. Zhu, "User-adaptive music emotion recognition," in *Proc. IEEE Int. Conf. Signal Processing*, vol. 2, 2004, pp. 1352–1355.

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

[14] T. Li and M. Ogihara, "Detecting emotion in music," in *Proc. Int. Conf. Music Information Retrieval*, 2003, pp. 239–240.

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

[16] 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 on Human-Centered Multimedia*, 2007, pp. 13–21.

[17] 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. Int. ACM Conf. Multimedia*, 2012.

[18] J. A. Russell, "A circumplex model of affect," *J. Personality and Social Science*, vol. 39, no. 6, pp. 1161–1178, 1980.

[19] Y.-H. Yang and H. H. Chen, "Predicting the distribution of perceived emotions of a music signal for content retrieval," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 19, no. 7, pp. 2184–2196, 2011.

[20] J. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

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

[22] 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. Society for Music Information Retrieval Conference*, 2011, pp. 85–90.

[23] J. V. Davis and I. S. Dhillon, "Differential entropic clustering of multivariate gaussians," in *NIPS*, 2006.

[24] J.-C. Wang, Y.-C. Shih, M.-S. Wu, H.-M. Wang, and S.-K. Jeng, "Colorizing tags in tag cloud: A novel query-by-tag music search system," in *Proc. Int. ACM Conf. Multimedia*, 2011, pp. 293–302.

[25] 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 and Expo*, 2011.

[26] O. Lartillot and P. Toiviainen, "A matlab toolbox for musi-cal feature extraction from audio," in *Proc. Int. Conf. Digital Audio Effects*, 2007.

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

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

[29] M. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.