# LINEAR REGRESSION-BASED ADAPTATION OF MUSIC EMOTION RECOGNITION MODELS FOR PERSONALIZATION

*Yu-An Chen,[1] Ju-Chiang Wang,[2] Yi-Hsuan Yang,[2] and Homer Chen[1]*

[1] National Taiwan University, Taiwan
[2] Academia Sinica, Taiwan
Emails: b96901042@ntu.edu.tw, {asriver, yang}@iis.sinica.edu.tw, homer@ntu.edu.tw

## ABSTRACT

Personalization techniques can be applied to address the subjectivity issue of music emotion recognition, which is important for music information retrieval. However, achieving satisfactory accuracy in personalized music emotion recognition for a user is difficult because it requires an impractically huge amount of annotations from the user. In this paper, we adopt a probabilistic framework for *valence-arousal* music emotion modeling and propose an adaptation method based on linear regression to personalize a background model in an online learning fashion. We also incorporate a component-tying strategy to enhance the model flexibility. Comprehensive experiments are conducted to test the performance of the proposed method on three datasets, including a new one created specifically in this work for personalized music emotion recognition. Our results demonstrate the effectiveness of the proposed method.

***Index Terms***— Personalization, music, emotion recognition, MLLR, MAPLR

## 1. INTRODUCTION

Music emotion recognition (MER) is important as it facilitates music organization, indexing, and retrieval based on emotion semantic [9], [19], [25]. Most MER systems adopt the valence-arousal (VA) emotion plane [14] to describe music emotion. Valence (or pleasantness) corresponds to the positive/negative affective state, and arousal (or activation) indicates the energy and stimulation level of the emotion. Since the perceived music emotion may differ from person to person, personalized MER (PMER) is needed [4], [7], [26], [27]. However, training a PMER model from scratch for a user needs a huge amount of annotations from the user [26], [27]. A practical personalization method with a reasonable annotation load is needed.

Motivated by the classic paradigm of speaker adaptation in speech recognition [5], [13], Wang et al. [20] proposed a probabilistic technique that performs personalization by adapting a background MER model to a user to reduce the annotations required. The background model, which is learned from a group of subjects, can be considered a model that averages the emotion perceptions of all subjects. This technique suggests that exploiting the available background model and tailoring it to a specific user is an effective approach to PMER.

Instead of using a maximum a posteriori (MAP) estimation method for adaptation, we propose a linear regression (LR)-based adaptation method for personalizing the MER model. The LR approach is adopted for the following two reasons. First, it has been shown in speech recognition that the LR approach works effectively when limited adaptation data are available [3], [12]. Second, the LR approach allows the incorporation of domain knowledge for component-tying [3] to enhance model flexibility. We exploit the fact that songs corresponding to neighboring points in the VA plane have similar affective characteristics [25] to tie Gaussian components in the same VA quadrant.

The contributions of this paper are twofold. First, we develop a novel maximum a posteriori linear regression (MAPLR) adaptation method for PMER. Experimental results show that the proposed method is superior to the previous MAP method [20] for PMER. Second, we create a new dataset, AMG240, for the design and evaluation of PMER algorithms.

This paper is organized as follows. Section 2 reviews prior work on personalized music emotion recognition. Section 3 briefly introduces the MER model used in this work. Section 4 describes the details of the proposed method. The evaluation setup and experimental results are presented in Section 5. Finally, Section 6 concludes this paper.

## 2. PRIOR WORK

The model retraining approach proposed by Yang et al. [22] represents one of the first attempts that explore PMER under the conventional VA regression framework using support vector regression as the predictor. In contrast to the model retraining approach, where information of the background model is discarded, the two-stage approach described in [23] personalize music emotion recognition without retraining. The first stage creates the background model for predicting the perception of all users, and the second stage predicts the

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

difference (perceptual residual) between the general perception and a target user. Similar to the probabilistic technique [20] discussed earlier, this approach retains the background model and enables dynamic adaptation. However, it does not model the correlation between valence and arousal and is unable to provide the confidence value of each prediction.

Although its primary focus is on categorical MER, the work done by Su et al. [17] falls into the category of PMER because it incorporates active learning to minimize user participation in the personalization process. However, it is a model retraining method, where PMER is formulated as a 4-class classification problem and the support vector machine is adopted to train classifiers for each subject. Because model adaption is not considered, this method still requires sufficient personal annotations to train personalized SVMs.

## 3. THE MER MODEL

Since diverse annotations are collected from different users of the same song, the emotion distribution of a song over the subjects is often modeled by a Gaussian distribution [15], [16], [19], [24]. In this work, we employ the acoustic emotion Gaussians (AEG) approach [19] to train our background MER model. AEG is a generative approach that uses a mixture of bivariate Gaussian distributions to model the association between the music signal of a song and the VA emotion values of the song. It is a state-of-the-art method for MER.

As depicted in Fig 1, a standard AEG model contains an acoustic Gaussian mixture model (GMM) $\{A_k\}_{k=1}^{K}$ and a VA GMM $\{G_k\}_{k=1}^{K}$, where $K$ denotes the number of Gaussian components. For the acoustic GMM, each component $A_k$ represents a certain type of acoustic features and is assumed to be associated with a VA (bivariate) Gaussian $G_k$ in the VA GMM. Then, a song's frame-based acoustic features are mapped into a $K$-dimensional probabilistic vector $\boldsymbol{\pi} = [\pi_1\ \pi_2\ \ldots \pi_K]$, which contains the posterior probability of each component of the acoustic GMM. The vector $\boldsymbol{\pi}$ serves as the prior weight for VA GMM to relate the acoustic features of a song to the emotion of the song. Therefore, we can use $\sum_{k=1}^{K} \pi_k G_k$ to generate the VA distribution of a song. Finally, given a set of training songs with their annotations and acoustic posterior vectors $\{X_n, \boldsymbol{\pi}_n\}_{n=1}^{N}$, where $X_n$ denotes the annotations of song $n$ by the subjects, we can learn the mean and covariance matrix of each VA Gaussian component using the expectation-maximization (EM) algorithm described in [1], [19].

In the prediction phase, AEG provides a mixture of Gaussian distributions in the VA plane. If we want to use a single Gaussian to represent the predicted emotion of a song [19], we can approximate the mixture of Gaussian distributions with a single Gaussian based on their KL-divergence [6], [19]. Moreover, the mean of the single Gaussian can be taken as a single point prediction (i.e., a pair of VA values instead of a probabilistic distribution), as illustrated in the rightmost part of Fig 1.

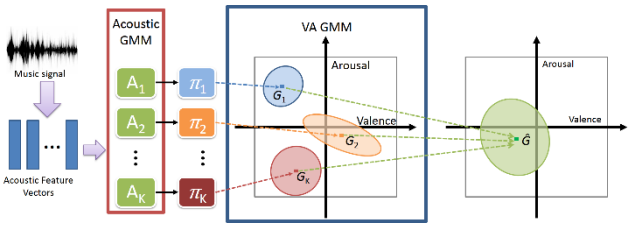


**Fig. 1.** Illustration of the basic idea of the acoustic emotion Gaussian (AEG) approach [19].

## 4. LINEAR REGRESSION FOR MODEL ADAPTATION

### 4.1. Linear regression and component tying

LR-based adaptation methods seek an optimal linear transformation to update each model parameter when given a set of personal data. For VA GMM, ideally we can learn a linear transformation for each Gaussian component. However, the personal annotations are usually insufficient to learn a transformation for each component individually. Therefore, we apply the component-tying strategy to group a number of components into a *tied-group,* and each tied-group shares a linear transformation. This strategy alleviates the complexity of the LR problem and at the same time increases the generalization ability [12].

In this paper, we propose a *quad-tied strategy* that ties the Gaussian components in the same quadrant of the VA plane, because many studies have indicated that emotions within a quadrant are highly correlated [9], [14], [25] and because the four quadrants in fact correspond to the four most representative emotion semantics [8], [25]. Given a well-trained background VA GMM with model parameters $\Lambda = \{\mu_k, \Sigma_k\}_{k=1}^{K}$, the parameter update criterion for the mean of the $k$-th Gaussian is defined as

$$\hat{\mu}_k = W_g \cdot m_k \in \mathbb{R}^2\ , \qquad (1)$$

where $W_g = [b_g\ R_g] \in \mathbb{R}^{2\times 3}$ is the transformation matrix of the $g$-th tied-group formed by concatenating a translation vector $b_g$ and a rotation matrix $R_g$, and $m_k = [1\ \mu_k^T]^T \in \mathbb{R}^3$ makes the translation possible. We denote each set of tied components by $\mathcal{K}_g$, $g \in \{1, \ldots, G\}$.

To solve the aforementioned linear regression problem, we vectorize the transformation matrix $W_g$ so that $\hat{\mu}_k = \widetilde{M}_k \cdot w_g$, where $\widetilde{M}_k = I_2 \otimes m_k^T \in \mathbb{R}^{2\times 6}$, $\otimes$ is the Kronecker product, and $w_g = [W_g^{(1)}, W_g^{(2)}]^T$ concatenates the two rows of $W_g$. When $G = K$, and $k = \mathcal{K}_k$, $\forall k$, the problem is generalized to the non-tying case.

### 4.2. MAPLR

MLLR [12] and MAPLR [2] are two main LR-based methods for speaker adaptation. MLLR aims at maximizing the likelihood on the adaptation data with respect to the LR parameters.

In contrast, MAPLR performs a Bayesian treatment for MLLR that involves the prior distribution of the LR parameter so that the posterior distribution is maximized. Our discussion below would focus on MAPLR since it is relatively more general.

Given the personal annotations of a subject and the associated posterior probability vectors, together denoted by $S = \{x_n, \boldsymbol{\pi}_n\}_{n=1}^{H}$, the posterior distribution function $p(w_g|S, \Lambda) \sim \mathcal{N}$ is defined by

$$p(w_g|S,\Lambda) \propto \prod_{n=1}^{N}\prod_{g=1}^{G} p(w_g) \cdot \sum_{k\in\mathcal{K}_g} \pi_{nk}|\Sigma_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_n-\hat{\mu}_k)^T\Sigma_k^{-1}(x_n-\hat{\mu}_k)\right\}, \quad (2)$$

where $p(w_g) \sim \mathcal{N}(w_g|\overline{w},\ \lambda_g \cdot I_6)$ is the prior distribution of $w_g$ with $\overline{w} = [0,1,0,0,0,1]^T$ and precision $\lambda_g$ (instead of variance). Note that $\overline{w}$ represents a non-effective transformation for $w_g$ to keep the adapted model close to the background VA GMM, and $\lambda_g$ plays the role of balancing between the background model and the personal annotations.

We employ the EM algorithm to maximize (2) iteratively. In the expectation step, we compute the posterior probability of the $k$-th Gaussian component for each personal annotation,

$$\gamma_{nk} = \frac{\pi_{nk}\cdot\mathcal{N}(x_n|\widetilde{M}_k w,\Sigma_k)}{\sum_h \pi_{nh}\cdot\mathcal{N}(x_n|\widetilde{M}_k w',\Sigma_h)}, \quad (3)$$

where $w = w_g$ for $k \in \mathcal{K}_g$, and $w' = w_g$ for $h \in \mathcal{K}_g$. In the maximization step, the updated $w_g$ is obtained by

$$\widehat{w}_g = \left(\sum_{k\in\mathcal{K}_g}\sum_{n=1}^{N}\gamma_{nk}\widetilde{M}_k^T\Sigma_k^{-1}\widetilde{M}_k + \lambda_g I_6\right)^{-1} \cdot \left(\sum_{k\in\mathcal{K}_g}\sum_{n=1}^{N}\gamma_{nk}\widetilde{M}_k^T\Sigma_k^{-1}x_n + \lambda_g\overline{w}\right). \quad (4)$$

The precision $\lambda_g$ is designed to be data-dependent and is defined by,

$$\lambda_g = C \cdot \exp\left\{-\max\left(\sum_n\sum_{k\in\mathcal{K}_g}\gamma_{nk} - \tau, 0\right)\right\}, \quad (5)$$

where $C \geq 0$ is a parameter to be determined, $\max(v, 0)$ is a hinge function, and $\tau$ is set to 5 empirically. As (5) shows, if sufficient personal annotations are available and have effects on the $g$-th tied-group, a large value for $\sum_n\sum_{k\in\mathcal{K}_g}\gamma_{nk}$ will be obtained, leading to a small $\lambda_g$. Thus, the adaptation learning will depend more on the personal annotations. MAPLR reduces to MLLR when $C = 0$.

## 5. EXPERIMENTAL RESULTS

This section describes the dataset, the experimental setup, and the experimental results.

### 5.1. Datasets and acoustic features

A new dataset was created specifically for evaluating the performance of PMER methods. The dataset is referred to as AMG240, as it contains 240 diverse music clips that have been labeled with discrete mood tags in All Music Guide (AMG), a professional music service website. Specifically, we retrieved the list of songs associated with 34 AMG mood tags [18] and the corresponding 30-second audio previews through the 7digital API. This initial set has diverse music content. In order to balance the distribution of the songs in the VA plane, we used the method proposed in [21] to project the 34 tags into the VA plane and randomly picked 60 clips from each quadrant according to the VA values of the tags. Finally, ten casual subjects were recruited to annotate the VA values of all the 240 clips in a silent computer lab. The subjects were well-instructed, and a graphic interface with retrospect functions was used to annotate the VA values [22] for better annotation consistency.

We also used two other datasets, MER60 [24] and DEAP [10], in the experiments. MER60 is comprised of 60 pieces of 30-second clips with VA annotations contributed by 99 subjects, where each clip is annotated by 40 subjects on average. However, unlike AMG240, only six subjects completely annotated all the 60 clips for MER60. Therefore, it makes more sense to use MER60 to train a background model. On the other hand, DEAP is composed of DEAP120 and DEAP40: The former includes 120 pieces of 60-second clips with quantized VA values annotated by 14–16 subjects using online self-assessment tools, whereas the latter encompasses 40 pieces of 60-second clips and 32 subjects who annotated all of the 40 clips using continuous VA scales in a lab environment.

In addition, we use the MIRToolbox [11] to extract 70 features encompassing dynamic, spectral, timbre and tonal descriptors according to the procedure described in [19].

### 5.2. Experimental settings and evaluation criteria

Since AMG240 and MER60 were collected by similar means (the ranges of VA values are [-1, 1]) while DEAP40 and DEAP120 have more in common (the VA values are both in the range of [1, 9]), the following two experimental settings were adopted. First, we used MER60 for training the background model and AMG240 for evaluating PMER. Because AMG240 has abundant personal annotations, we performed 4-fold cross-validation for PMER that uses 3 folds for testing and one fold for adaptation. Second, DEAP120 is used to train the background model, and 5-fold cross-validation is performed on DEAP40, one fold for testing and 4 folds for adaptation. For both experiments, we repeated the PMER evaluation for five times, each with a different random permutation to divide the folds. There was no overlap between the training, adaptation, and testing sets.

We adapted the background VA GMM in an incremental manner to evaluate the performance of PMER. At each iteration, we added four more randomly selected personal annotations from the training set to the adaptation pool, which was then used to adapt the background model until all the training data were used. Finally, the accuracy was computed

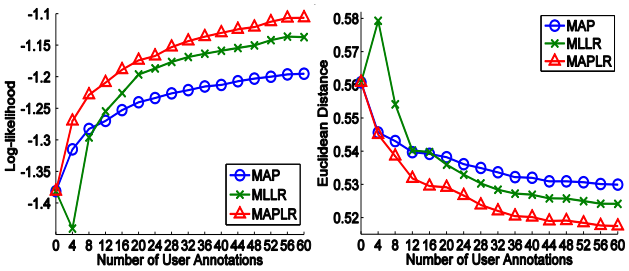


**Fig. 2.** The performance of personalized music emotion recognition evaluated on AMG240.

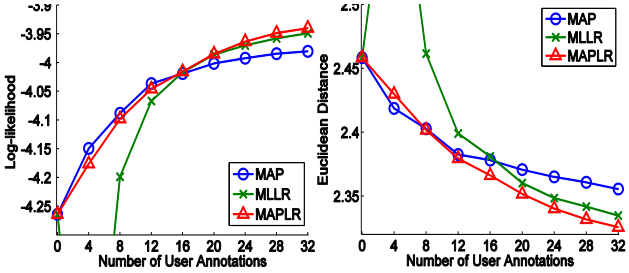


**Fig. 3.** The performance of personalized music emotion recognition evaluated on DEAP40.

in each iteration and averaged over all subjects and repetitions. Note that all the methods were tested under the same experimental settings described above.

The prediction results were evaluated in two ways. First, we calculated the log-likelihood (LL) [20] of the ground truth annotation on the predicted single VA Gaussian and evaluated the accuracy of both the predicted mean and covariance. Larger LL represents better performance. Second, we measured the Euclidean distance (ED) [15], [19] between the ground truth annotation and the mean vector of the predicted VA Gaussian. Smaller ED corresponds to better accuracy.

### 5.3. Experimental results

We compared the performance of the three model adaptation methods, MAP, MLLR and MAPLR, using MAP as the baseline [20]. Figures 2 and 3 show the results on AMG240 and DEAP40, respectively. The following observations can be made. First, the results clearly show that MAPLR outperforms MAP except for some cases with scarce adaptation data on DEAP40. Second, MLLR suffers from over-fitting when very few adaptation data are used. See, for example, the case with 4 annotations. On the contrary, thanks to the Bayesian treatment, MAPLR reduces the risk of over-fitting and exhibits better generalization ability. Accordingly, MAPLR outperforms MLLR when adaptation data are scarce. Third, MAPLR and MLLR perform similarly on DEAP40 when the adaptation data are abundant. This is because that the factor $\lambda_g$ is designed to decay rapidly once sufficient adaptation data are available. In contrast, the baseline method MAP is less flexible; its performance on both datasets is excessively influenced by the background model even when plentiful adaptation data are used.

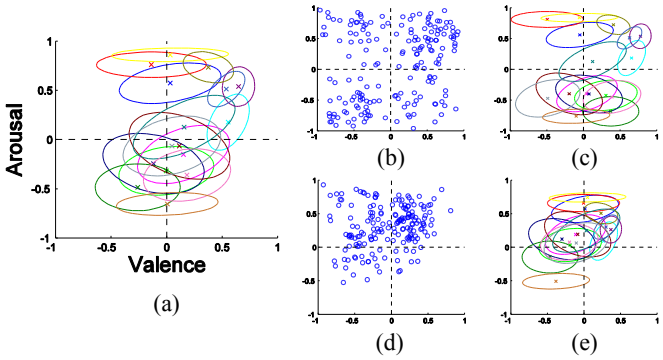


**Fig. 4.** Qualitative illustration of the performance of PMER. (a) The background VA GMM trained with MER60, where each ellipse stands for a Gaussian component. (b) Distribution of the test annotations labeled by Subject #1 in AMG240, where each circle corresponds to the VA annotation of a song. (c) VA GMM personalized by MAPLR with 16 annotations labeled by Subject #1. (d) Distribution of the test annotations labeled by Subject #2 in AMG240. (e) VA GMM personalized by MAPLR with 16 annotations labeled by Subject #2.

To gain more insights into the capability of MAPLR, we showcase two train/test data pairs selected from the first experiment, which uses AMG240 as the test data. Fig. 4 shows the distributions of the background model on MER60, the test annotations of two subjects on the test set of AMG240, and their personalized models learned from the training annotations. We can see that the Gaussian components of the background model (cf. Fig 4 (a)) generally spread over the VA plane but somewhat closer to the origin. For Subject #1, the background model appears conservative since the annotation distribution (cf. Fig. 4 (b)) is more diverse, covering even the boundaries of the VA plane. By learning from the adaptation data, our method drives the Gaussian components of the background model towards the boundaries and forms a new personalized VA GMM (cf. Fig. 4 (c)) covering a larger area of the plane. For subject #2, we can see that the personalized model (cf. Fig. 4 (e)) adapts well and performs well in predicting the ground truth distribution of the test set (cf. Fig. 4 (d)).

## 6. CONCLUSIONS

In this paper, we have proposed an LR-based model adaptation method to personalize a background MER model in an online fashion. We have also provided empirical evidences to show the effectiveness of the proposed method. The proposed method works effectively across a wide range of available data for adaptation and is particularly useful when only a limited amount of adaptation data are available from the user.

## 7. ACKNOWLEDGMENT

This work was supported by the NSC of Taiwan under Grant NSC 100-2221-E-002-198-MY3, and by an Academia Sinica Fellowship to Ju-Chiang Wang, sponsored by Academia Sinica, Taiwan.

## REFERENCES


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

[2] C. Chesta, O. Siohan, and C.-H. Lee. "Maximum a posteriori linear regression for hidden Markov model adaptation," in *Proc. Eurospeech*, 1999.

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

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

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

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

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

[8] 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. Society Music Information Retrieval Conf.*, 2008.

[9] 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 Music Information Retrieval Conf.*, 2010, pp. 255–266.

[10] S. Koelstra, C. M¨uhl, 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 Computing*, vol. 3, no. 1, pp. 18–31, 2012.

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

[12] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer speech and language*, vol. 9, no. 2, p. 171, 1995.

[13] 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.

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

[15] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions from audio," in *Proc. Int. Society Music Information Retrieval Conf.*, 2010.

[16] E. M. Schmidt, and Y. E. Kim, "Prediction of time-varying musical mood distributions using Kalman filtering," in *Proc. Int. Conf. Machine Learning and Applications*, 2010, pp. 655-660.

[17] D. Su and P. Fung, "Personalized music emotion classification via active learning," in *Proc. ACM Workshop Music Information Retrieval with User-centered and Multimodal Strategies*, 2012.

[18] [online] http://www.allmusic.com/moods

[19] 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*, pp. 89–98, 2012.

[20] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng, "Personalized music emotion recognition via model adaptation," in *Proc. Asia-Pacific Signal and Information Processing Association Annu. Summit and Conference*, 2012.

[21] J.-C. Wang, Y.-H. Yang, K. Chang, H.-M. Wang, and S.-K. Jeng, "Exploring the relationship between categorical and dimensional emotion semantics of music," in *Proc. ACM Workshop Music Information Retrieval with User-centered and Multimodal Strategies*, 2012, pp. 63–68.

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

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

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

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

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

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