Design Studies and Intelligence Engineering
L.C. Jain et al. (Eds.)

doi:10.3233/FAIA220004

PSIC3839: Predicting the Overall Emotion and Depth of Entire Songs

Liang XU[a], Zongyang YUN[a], Zaoyi SUN[b], Xin WEN[a], Xianan QIN[c], and Xiuying QIAN[a,1]
[a] *Department of Psychology and Behavioral Sciences, Zhejiang University, China.*
[b] *College of education, Zhejiang University of Technology, China.*
[c] *College of Textile Science and Engineering, Zhejiang Sci-Tech University, China.*

Abstract. Music emotion recognition (MER) studies have made great progress in detecting the emotions of music segments and analyzing the emotional dynamics of songs. The overall emotion and depth information of entire songs may be more suitable for real-life applications in certain scenarios. This study focuses on recognizing the overall emotion and depth of entire songs. First, we constructed a public dataset containing 3839 popular songs in China (PSIC3839) by conducting an online experiment to collect the arousal, valence, and depth annotation of each song. Second, we used handcrafted feature-based method to predict the overall emotion and depth values. Support vector regressions using Mel frequency cepstrum coefficients features as inputs achieve good model performance (arousal: $R^2 = 0.609$; valence: $R^2 = 0.354$; and depth: $R^2 = 0.465$). Finally, the groupwise and personalized results were also investigated by training a unique regressor for each group or individual, which provides a reference for future research.

Keywords. Music emotion recognition, music depth recognition, audio signal processing, handcrafted spectral feature.

1. Introduction

Music emotion recognition (MER), a research area constructing computation models to automatically identify the perceived music emotion [1], has made great progress in the past decade [2, 3]. The detected emotion information is widely used in music information retrieval (MIR) [4] and music recommendation (MR) [5] systems. Since the emotional content of a song fluctuates, most MER studies have focused on detecting the emotions of music segments [4] or analyzing the dynamics of a song at as granular a level as possible [2, 6]. Regarding real-life use, the emotional labels of the entire song may be more practical in MIR and MR systems, that is, we sometimes need the overall emotion information of a song and not the emotional label of the segment or the dynamic change of a song's emotion. However, predicting the overall emotion of an entire song is more difficult than predicting the perceived emotion of music segments. This task may not only be related to the human auditory system [2], but it may also be affected by working memory [7], short-term memory [24], psychological states [8], and so forth [3]. For example, emotions (annotators' emotional state) are more likely to be aroused after

[1] Corresponding Author: Xiuying Qian. Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou 310028, China. E-mail: xqian@zju.edu.cn.

listening to an entire song, which may affect the annotation of the perceived emotion. Therefore, recognizing the overall emotion of an entire song is an urgent issue, which is of practical value but has new challenges.

To the best of our knowledge, existing public music emotion datasets are usually collected for music emotion variation detection (e.g., the DEAM [9] and MTurk [10] datasets), or they only contain emotional labels of music segments (e.g., MER60 [11]). There is still no dataset for the overall emotional evaluation of songs. In addition, the valence-arousal (VA) model [12] is widely used in MER studies to describe emotions because it can be applied to form the corresponding relations between music features and music emotions simply and intuitively through the coordinates [2]. However, the depth, a crucial dimension recently found by psychologists [13], has rarely been considered in previous work. "Depth" is a basic music perception attribute, positively correlated with intelligent, sophisticated, inspiring, complex, poetic, deep, emotional, and thoughtful attributes while negatively correlated with party music and danceable attributes [13]. Thus, the present study first constructed a public dataset containing 3839 songs popular in China (PSIC3839) to predict the overall music emotion and depth. The arousal, valence, and depth values were collected for each song in this dataset. Additionally, several participants annotated more than 650 songs, which make this dataset also available for constructing personalized models [3, 11].

Handcrafted feature-based methods are commonly used in traditional MER studies to analyze low- and mid-level audio features, such as Mel frequency cepstrum coefficients (MFCCs), the timbre, and the intensity [2, 18]. In addition, many studies have used machine learning methods, such as the Bayesian model [16], support vector regression (SVR) [17], and Gaussian mixture model [4], to map the relations between handcrafted features and perceived emotions. Therefore, this study tests whether the above methods are also suitable for music overall emotion and depth recognition.

The rest of this study is organized as follows. Section 2 introduces the construction method and data distribution of the PSIC3839 dataset. Section 3 describes the signal processing methods in detail. Section 4 presents the general and personalized results. Finally, Section 5 concludes the work.

2. The PSIC3839 Dataset

To construct the PSIC3839 dataset, we first selected songs from the websites of QQ Music (https://y.qq.com/) and NetEase CloudMusic (https://music.163.com/), which are two popular music sites in China. After browsing the popular songs posted from January 2018 to June 2018, a total of 6916 songs were first considered. However, limited by copyrights, only 3839 songs were retained for annotation. Most of the songs are pop music (68.95%), followed by soundtrack, nursery rhymes, world music, dance music, rock, folk and others (see Figure 1). The annotation method is introduced in Section 2.1, and the annotation results are presented in Section 2.2.

2.1. Annotation Collection

Before the formal annotation experiment, a pretest was conducted to select the participants. After explaining the differences between the "felt" and "perceived" emotions, we introduced the meanings of the words "valence", "arousal" and "depth" to

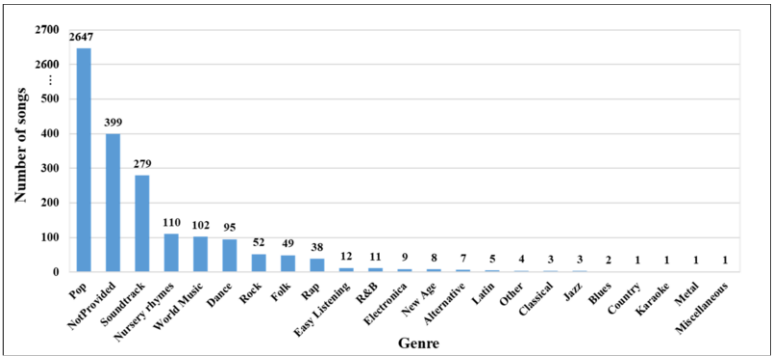

Figure 1. Genre distribution of the PSIC3839 dataset. The genre information was collected from the websites of QQ Music (https://y.qq.com/). As shown, most of the songs are pop music (68.95%), followed by soundtrack (7.27%), nursery rhymes (2.86%), and world music (2.66%). 10.39% of songs' genre information was not found.

Table 1. Correspondence between different dimensions and adjectives.

Target dimensions		Adjectives
Arousal	High	Intense, tense, aggressive, angry, abrasive, strong, thrilling, manic.
	Low	Mellow, gentle, calming, warm, reflective, relaxing, romantic, sensual.
Valence	Positive	Happy, fun, merry, joyful, enthusiastic, lively, animated, amusing.
	Negative	Sad, depressing.
Depth	Deep	Intelligent, sophisticated, inspiring, complex, poetic, dreamy, thoughtful, emotional.
	Not deep	Party music, danceable.

the participants in both the pretest and formal experiments by describing their corresponding adjectives. For example, high arousal music can be described as intense, tense, aggressive, angry, abrasive, and strong; and low arousal music can be described as mellow, gentle, calming, and sensual. According to the findings in [13], the correspondence between the different dimensions and adjectives used in this study is shown in Table 1. Notably, this study used relatively few adjectives for negative valence, which may bias the process. A balanced way, providing adequate and similar number of adjectives, should be advocated in future work.

Next, sixteen songs were selected as test songs, and their arousal, valence, and depth values were annotated by various authors and volunteers. These songs were distributed in each quadrant of the arousal-valence-depth plane (e.g., two songs were positive, highly aroused, and deep). The test songs were used for the pretest by asking participants to annotate them with emotion and depth labels. Participants can participate in the formal experiment only when the accuracy rate is above 85%.

The formal experiment was designed such that each song was heard and annotated by at least five unique participants. After a brief description of the experiment, participants were asked to listen to music in the listening order given. Each song was followed by a self-reported questionnaire asking participants to evaluate the overall estimation of the perceived emotion and depth. Arousal and depth were evaluated on a 5-point Likert scale from -2 (not at all) to 2 (very much), and valence was evaluated from

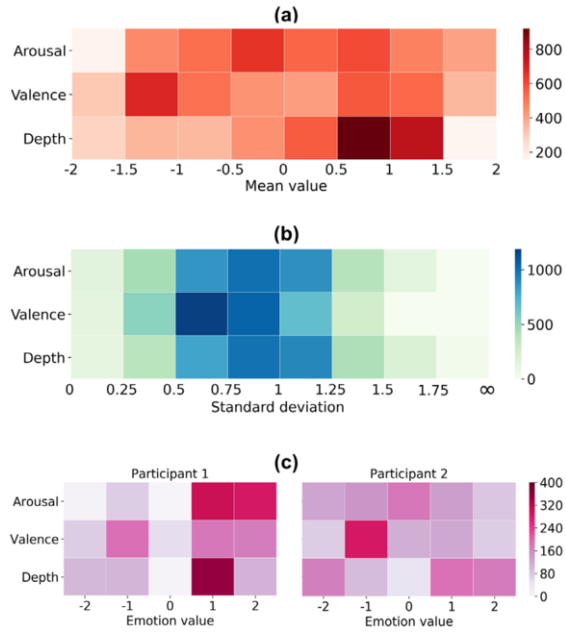

Figure 2. The overall distribution of emotion annotation results. Each square represents the number of songs in the target interval.

-2 (*negative*) to 2 (*positive*). The annotation task was conducted online. The following rules were implemented to ensure the quality of the annotation:

1) The listening order was independently randomized to minimize the influence of the presentation order [14].
2) The listening task was limited to 50 songs per day to avoid fatigue.
3) All participants were asked to use the same online music player to play the songs in the same format.
4) To concentrate on the music, listening tasks were conducted with participants' eyes closed.
5) Other activities that may distract the annotators were not allowed during the listening task.

2.2. Annotation Results

A total of 87 individuals, recruited from campus, participated in the formal experiment. Unfortunately, two annotators' information was lost, and the rest of them (52 females, 33 males) were aged 21.91 years (SD = 2.03). Most of them often listen to music (96.47%), while only eight participants had received professional music training.

Each song was finally annotated five to seven times. The overall distribution of the annotation results is shown in Figure 2a. Similar to previous work [10], the valence value shows a fuzzy bipolar distribution; the arousal value is difficult to distinguish, dispersing from the middle to both sides; and the depth value presents a right-skewed distribution. Deep songs may be more popular in our life, so most of songs were annotated as deep music. In addition, we observed that the arousal ratings are positively correlated with the

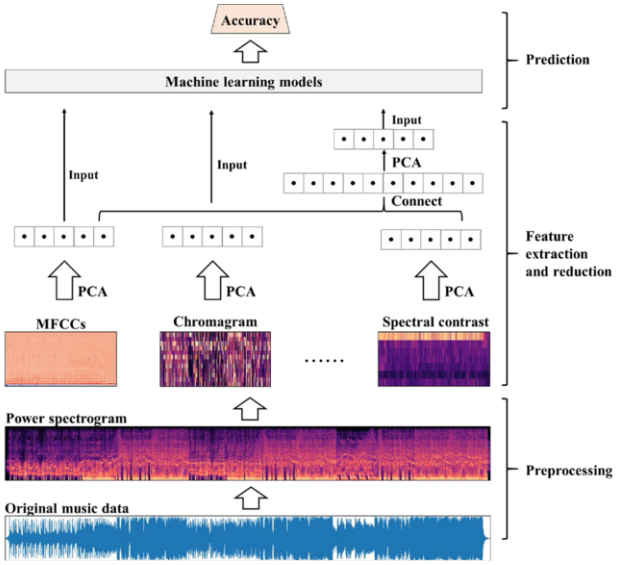

Figure 3. The framework of the proposed method.

valence ratings (r(3838) = 0.556, $p < 0.001$) and negatively correlated with the depth ratings (r(3838) = -0.630, $p < 0.001$), and the valence ratings are also negatively correlated with the depth ratings (r(3838) = -0.628, $p < 0.001$). The correlation results are similar to other datasets [10], [26] and the findings in psychological research [13]. This finding may provide a reference for future music psychology research.

Different people listening to the same music may experience different perceived emotions [15]; thus, the standard deviation (SD) is an effective indicator for the selection of songs for subsequent modeling. The distribution of the SD for all songs is depicted in Figure 2b. The Friedman test shows that the SDs of valence are significantly lower than arousal and depth ($\chi^2 = 247.829$, $p < 0.001$), denoting that individuals are more consistent in annotating valence. We assume that, since full songs were considered, most pop songs do not vary greatly on valence (from happy to sad) but have several segments with lower and higher intensity (from verses to chorus). Thus, this variation in valence might generate more agreement between annotators.

In addition, five participants annotated more than 650 songs. The annotation results of the unique individuals are retained for personalized MER studies. Two samples are shown in Figure 2c. We can see that participant 1 annotated almost all songs as high arousal (87.18%) and deep music (68.85%). The annotation results of participant 2 are more even.

3. Methods

As shown in Figure 3, handcrafted feature-based method contains three main components: preprocessing, feature extraction and reduction, and prediction. The preprocessing and feature extraction and reduction approaches of the handcrafted feature-based method is introduced here. Section 4 will describe the prediction approach and corresponding results.

3.1. Preprocessing

We uniformly extract 180 seconds of audio at the beginning of each song for subsequent feature reduction, even though the song durations are not equal. A small number of songs less than 180 seconds are repeated so that they are set length. The sampling rate is 22050 Hz for each song. The short-time Fourier transform is then applied to obtain the power spectrogram by computing discrete Fourier transforms over Hann windows [25], and the window length is 1.025 seconds with a 256 milliseconds step size. We finally obtained a 1025 × 704 spectrogram, which is used for subsequent feature extraction.

3.2. Handcrafted Feature Extraction and Reduction

Various handcrafted features are considered here. First, the MFCCs, reflecting the nonlinear frequency sensitivity of the human auditory system, are extract-ed. Each spectrogram is filtered by a Mel filter bank to generate the Mel-scaled spectrogram, and the MFCCs are computed by selecting lower cepstral coefficients [20]. Second, a chroma- gram, describing the pitch components of music, is computed from the power spectrogram using the approach of [21]. Third, we compute the spectral contrast by dividing each spectrogram into sub-bands. For each sub-band, the energy contrast is estimated by comparing the mean energy in the top quantile to that of the bottom quantile [22]. Regarding other timbre-related features, we extract the spectral centroid, bandwidth, roll-off, and flatness, related to tone structure [9], from the power spectrogram. In addition, tonal centroid features (tonnetz), useful for harmonic change detection [19], are calculated by projecting chromagram features onto a 6-dimensional basis in this study. The tempo (beats per minute) of each song is also considered.

The above process is achieved through the librosa toolkit [23]. Principal component analysis (PCA) is then used to reduce the dimensionality of each type of handcrafted feature (with 99% of the variance). The reduced features are used as the inputs of the machine learning models, respectively. In addition, the handcrafted features are also combined as model inputs.

4. Results

4.1. General Results

Music overall emotion and depth recognition is formulated as a regression problem in this study. The handcrafted spectral features are used as model inputs to investigate their effects on music arousal, valence, and depth recognition. The effects of multiple linear regression (MLR), Bayesian ridge regression (BRR), and SVR algorithms in predicting the overall emotion and depth of entire songs are compared here. In addition, the ground truth value is scaled to a value of 0 to 1 via min-max scaling. The model's performance is evaluated using the ten-fold cross-validation technique, and the prediction accuracy is measured using the R^2 statistics as follows: $R^2 = 1 - \frac{\sum_{i=1}^{N}(X_i - Y_i)^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$, where X_i is the ground truth and Y_i is the predicted result. The general performance of the models constructed with different inputs is presented in Table 2. The SVR model with the radial basis function kernel performs the best. Regarding the inputs, MFCCs features perform the best in predicting arousal (R^2 = 0.609), valence (R^2 = 0.354), and depth (R^2 = 0.465).

Table 1. The performances of the general models.

Inputs	Arousal	Valence	Depth
Chromagram	0.425	0.217	0.264
Combined features	0.450	0.252	0.285
MFCCs	**0.609**	**0.354**	**0.465**
Spectral bandwidth	0.277	0.197	0.182
Spectral centroid	0.377	0.240	0.235
Spectral contrast	0.576	0.337	0.421
Spectral flatness	0.369	0.183	0.245
Spectral rolloff	0.365	0.222	0.232
Tonnetz	0.484	0.301	0.343

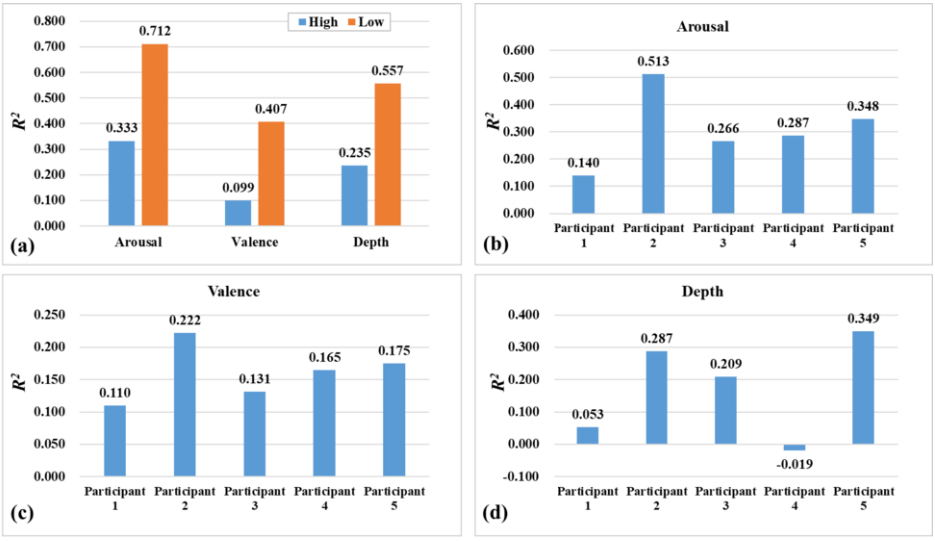

Figure 4. The performances of the groupwise (a) and personalized (b-d) models.

We also tested whether the emotional and deep annotation consistency will affect the model. We partition one-third of the songs with the highest standard deviations (SDs) and one-third of the songs with the lowest SDs into two groups and train a regressor using the SVR for each group. For each dimension (arousal, valence, or depth), the prediction accuracy for the high or low SD group is evaluated separately. The results are presented in Figure 3d. A ten-fold cross-validated paired t-test shows that the high SD groups perform significantly worse than the low SD group (arousal: $t = -19.719$, $p < 0.001$; valence: $t = -9.500$, $p < 0.001$; and depth: $t = -4.605$, $p < 0.005$). This means that the perceived emotion of songs, which are commonly agreed across individuals, are more likely to be predicted by audio features. Thus, the scoring consistency (SD in this study) may help us determine which songs are easy for emotion detection.

4.2. Personalized Results

To evaluate the datasets annotated by unique individuals, we train a personalized regressor for each individual who annotated more than 650 songs. The SVR algorithm is also used to train the model, and the model performance is evaluated via the ten-fold cross-validation technique. As shown in Figures 4b-4d, different individuals' recognition models are significantly different in predicting arousal ($\chi^2 = 22.000$, $p < 0.001$), valence ($\chi^2 = 13.420$, $p < 0.001$), and depth ($\chi^2 = 27.920$, $p < 0.001$). In addition, the personalized models only reach a mean R^2 of 0.311 in predicting arousal, a mean R^2 of 0.161 in predicting valence, and a mean R^2 of 0.176 in predicting depth, which perform significantly worse than the general models. This may be caused by an insufficient number of training samples, and combining the general and personalized models can be attempted in the future to solve the above problem.

5. Conclusion

In this study, we present a new and challenging issue with the aim of recognizing the overall emotion and depth of entire songs. First, we created a novel public dataset for the above research topic, and the dataset contains a large number of songs. Next, the handcrafted feature-based method was applied to predict the overall emotion and depth of songs. The SVR using handcrafted spectral features as inputs achieves good performance in this study (arousal: $R^2 = 0.609$; valence: $R^2 = 0.354$; and depth: $R^2 = 0.465$). Finally, the results of general, groupwise, and personalized models show the need for a large dataset, emphasize the importance of scoring consistency, and provide a reference for future work.

References

[1] S.H. Chen, Y.S. Lee, W.C. Hsieh, and J.C. Wang, Music emotion recognition using deep Gaussian process, in *Asia-Pac. Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Hong Kong, 2015, 495–498.

[2] X. Yang, Y. Dong, and J. Li, Review of data features-based music emotion recognition methods, *Multimedia Syst* **24** (2018), 365–389.

[3] L. Xu et al., Effects of individual factors on perceived emotion and felt emotion of music: Based on machine learning methods, *Psychol. Music* 2020. DOI: 10.1177/0305735620928422.

[4] 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. 20th ACM Int. Conf. Multimedia*, Nara, Japan, 2012, 89-98.

[5] S.H. Park, S.Y. Ihm, W.I. Jang, A. Nasridinov, and Y.H. Park, A Music Recommendation Method with Emotion Recognition Using Ranked Attributes, in *Computer Science and its Applications*, Berlin, Heidelberg, 2015, 1065–1070.

[6] E.M. Schmidt and Y. E. Kim, Prediction of Time-varying Musical Mood Distributions from Audio, in *Proc. 11th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Utrecht, Netherlands, 2010, 465–470.

[7] A. D. Baddeley and G. Hitch, Working Memory, in *Psychology of Learning and Motivation*, G.H. Bower, Ed. New York, NY, USA: Academic Press, 1974, 47–89.

[8] K. Kallinen and N. Ravaja, Emotion perceived and emotion felt: Same and different, *Music Sci.* **10** (2016), 191–213.

[9] M. Soleymani, M. N. Caro, E. M. Schmidt, C.Y. Sha, and Y.H. Yang, 1000 songs for emotional analysis of music, in *Proc. 2nd ACM Int. Workshop Crowdsourcing Multimedia*, Barcelona, Spain, 2013, 1–6.

[10] J.A. Speck, E.M. Schmidt, O.G. Morton, and Y.E. Kim, A Comparative Study of Collaborative Vs. Traditional Musical Mood Annotation, in *Proc. 12th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Miami, Florida, USA, 2011, 549–554.

[11] Y.H. Yang, Y.F. Su, Y.C. Lin, and H.H. Chen, Music emotion recognition: the role of individuality, in *Proc. 2nd ACM Int. Workshop on Human-centered Multimedia*, Augsburg, Bavaria, Germany, 2007, 1–9.

[12] J.A. Russell, A circumplex model of affect," *J. Pers. Soc. Psychol.* **39** (1980), 1161–1178.

[13] D.M. Greenberg, M. Kosinski, D.J. Stillwell, B.L. Monteiro, D.J. Levitin, and P.J. Rentfrow, The Song Is You: Preferences for Musical Attribute Dimensions Reflect Personality, *Soc. Psychol. Personal Sci.*, **7** (2016), 597–605.

[14] N.N. Vempala and F.A. Russo, Modeling Music Emotion Judgments Using Machine Learning Methods, *Front. Psychol.*, **8** (2018), 2239.

[15] A. Gabrielsson, Emotion perceived and emotion felt: Same or different?, *Music Sci*. **5** (2001), 123–147.

[16] B. Wu, E. Zhong, A. Horner, and Q. Yang, Music Emotion Recognition by Multi-label Multi-layer Multi-instance Multi-view Learning, in *Proc. ACM Conf. Multimedia*, Orlando, Florida, USA, 2014, 117–126.

[17] H. Xianyu et al., SVR based double-scale regression for dynamic emotion prediction in music, in *Proc. 41st IEEE Int. Conf. Acoustics, Speech and Signal Processing* (ICASSP), Shanghai, China, 2016.

[18] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks, *IEEE Trans. Multimedia* **16** (2014), 2203–2213.

[19] C. Harte, M. Sandler, and M. Gasser, Detecting harmonic change in musical audio. in *Proc. 1st ACM Workshop on Audio and Music Computing Multimedia*, Santa Barbara, CA, USA, 2006, 21-26.

[20] B.T. Meyer and B. Kollmeier, Complementarity of MFCC, PLP and Gabor features in the presence of speech-intrinsic variabilities, in *Proc. 10th Annu. Conf. Int. Speech. Commun. Assoc. (INTERSPEECH)*, Brighton, UK, 2009, 2755–2758.

[21] D. Ellis, Chroma Feature Analysis and Synthesis, 2017. From: https://labrosa.ee.columbia.edu/matlab/chroma-ansyn

[22] D.N. Jiang, L. Lu, H.J. Zhang, J.H. Tao, and L.H. Cai, Music type classification by spectral contrast feature, in *Proc. IEEE Int. Conf. Multimedia. Expo* (ICME), 2002, 113–116.

[23] B. McFee et al., librosa: Audio and Music Signal Analysis in Python, in *Proc. 14th Python in Science Conf.* (SCIPY), Austin, Texas, 2015, 18–24.

[24] S. Hochreiter and J. Schmidhuber, Long Short-Term Memory, *Neural Comput.*, **9** (1997), 1735–1780.

[25] E. R. Kanasewich, *Time sequence analysis in geophysics*, 2nd ed. Edmonton, Alta: University of Alberta Press, 1975, 106–108.

[26] Y.A. Chen, Y.H. Yang, J.C. Wang, and H. Chen, The AMG1608 dataset for music emotion recognition, in *Proc. 40st IEEE Int. Conf. Acoustics, Speech and Signal Processing* (ICASSP), Brisbane, QLD, 2015, 693–697.