applied sciences

Article

Comparison and Analysis of Acoustic Features of Western and Chinese Classical Music Emotion Recognition Based on V-A Model

Xin Wang [1], Li Wang [1] and Lingyun Xie [2,*]

[1] School of Musical and Recording Art, Communication University of China, Beijing 100024, China; metero_wx@cuc.edu.cn (X.W.); wwli@cuc.edu.cn (L.W.)
[2] Key Laboratory of Media Video & Audio, Communication University of China, Beijing 100024, China
* Correspondence: xiely@cuc.edu.cn

Abstract: Music emotion recognition is increasingly becoming important in scientific research and practical applications. Due to the differences in musical characteristics between Western and Chinese classical music, it is necessary to investigate the distinctions in music emotional feature sets to improve the accuracy of cross-cultural emotion recognition models. Therefore, a comparative study on emotion recognition in Chinese and Western classical music was conducted. Using the V-A model as an emotional perception model, approximately 1000 pieces of Western and Chinese classical excerpts in total were selected, and approximately 20-dimension feature sets for different emotional dimensions of different datasets were finally extracted. We considered different kinds of algorithms at each step of the training process, from pre-processing to feature selection and regression model selection. The results reveal that the combination of MaxAbsScaler pre-processing and the wrapper method using the recursive feature elimination algorithm based on extremely randomized trees is the optimal algorithm. The harmonic change detection function is a culturally universal feature, whereas spectral flux is a culturally specific feature for Chinese classical music. It is also found that pitch features are more significant for Western classical music, whereas loudness and rhythm features are more significant for Chinese classical music.

Keywords: music emotion recognition; classical music; V-A model; feature selection; extreme random tree

check for updates

Citation: Wang, X.; Wang, L.; Xie, L. Comparison and Analysis of Acoustic Features of Western and Chinese Classical Music Emotion Recognition Based on V-A Model. *Appl. Sci.* **2022**, *12*, 5787. https://doi.org/10.3390/app12125787

Academic Editor: Philippe Esling

Received: 12 May 2022
Accepted: 5 June 2022
Published: 7 June 2022

1. Introduction

With the development and popularization of the Internet, online music applications have gradually become the main channel for people to listen to music. Now, people can listen to all kinds of music from various countries on the Internet. One of the challenges for online music services is to automatically classify large amount of music that meets people's listening needs. Emotion-based music organization and retrieval technology are feasible, and the core is music emotion recognition (MER). The main work is to construct an emotion calculation model based on audio content and text information to realize the process of automatic recognition of music emotion [1,2]. The three research directions in MER are music emotion classification (MEC), emotion regression prediction, and music emotion variation detection, with acoustic feature extraction being one of the most important steps [3]. This addresses the analysis and extraction of meaningful information from audio signals to obtain a compact and expressive description that is machine processable [4]; this significantly contributes to the performance of the classification or regression system, which is always the focus of MER research.

In this study, we adopted the valence-arousal emotional model (hereinafter termed V-A model) instead of a categorical model, treating MER as a continuous regression problem [5]. The V-A model, proposed by Russell, believes that the emotional state is a point distributed in a two-dimensional space containing valence and arousal [6]. The valence reflects the degree of positive and negative emotions, and the degree of arousal reflects the intensity of emotions [7]. Two reasons for employing the V-A model are (1) to avoid ambiguities with the terms used to label emotions and (2) to generate playlists with smooth transitions from one emotion to another.

There have been many studies on the acoustic feature set of MER focusing on Western music [8–11]. Liu et al., proposed an algorithm called multi-emotion similarity preserving embedding and found that the mean and standard deviation of spectral flux, the first component of mel-frequency cepstral coefficients (MFCC), and the first and second beat histogram peaks in BPM are the most important features for MEC [12]. Yang et al., summarized the methods of MER based on data features. They mentioned that the acoustic features usually used in MER include MFCC, octave-based spectral contrast, statistical spectrum descriptors (spectral centroid, flux, roll-off, and flatness), and Chromagram [2]. Zhang et al., applied a shrinkage method to feature selection in the arousal emotional dimension and proposed that low specific loudness sensation coefficients, root mean square, and loudness-flux are the most useful features [13]. Grekow examined the influence of different feature sets on valence and arousal prediction and concluded that a combination of different types of features can improve the results [14]. Rhythm features are important for arousal prediction, and tonal features are useful for detecting valence.

A large quantity of research and applications of MER also exist in the related fields of neuroscience. Rolls described a theory of the neurobiological foundations of aesthetics and art, which have roots in emotion [15]. The latest research explored the high association between emotional arousal and neuro-functional brain connectivity measures [16]. Furthermore, as an important application of neuroscience, deep learning has recently become a popular topic in MER studies. Many paradigms have been proposed such as Convolutional Neural Networks (CNNs) applied to spectrograms or MFCC trajectories, Long Short-Term Memory layers (LSTM) for modeling of longer temporal context in music, and other deep learning strategies [17–19]. Although these end-to-end models have gradually become the mainstream algorithm, they do not rely on handcrafted features and are incapable of computing feature importance.

As a result of the progress in music globalization, an increasing number of researchers have begun to focus on the influence of different cultures on the emotional perception of music. Studies have shown that in addition to the characteristics of music, the cultural environment affects people's perception of music emotions [20], and people from different cultural backgrounds may have significant differences in the way they perceive emotions in the same music [21]. In cross-cultural dataset music emotion recognition (MER), most studies employ the MEC model [22–24]. Studies based on emotion regression models, especially for Western and Chinese classical music, are still limited. A study by Yang et al. represents one of the first attempts to adopt a regression approach for MER based on Western, Chinese, and Japanese pop music datasets [25]. However, their best results for the emotional regression model were 58.3% for arousal and 28.1% for valence. Similarly, Hu and Yang explored the generalizability of emotion regression models for Western and Chinese pop music [26]. It was found that loudness and timbre features work well for both valence and arousal prediction. The size of the training datasets and the annotation reliability level of training and testing datasets can affect the regression performances on both valence and arousal, especially for cross-cultural and cross-dataset music emotional prediction. Although Hu and Yang evaluated three distinct cross-cultural datasets, all of them were composed of pop music. Because Western pop music is now ubiquitous in China, the standard instrument arrangement of pop music is relatively unified. There were no significant differences between Western and Chinese pop music in terms of melody, harmony, and orchestration, which may make it difficult to accurately extract culturally representative feature sets. Therefore, for our study we focused on more culturally representative Western and Chinese classical music as our research object in order to explore the influence of culture on MER. Furthermore, we also considered different kinds of algorithms at each step of the training process to maximize the performance of a MER system based on the regression approach, from pre-processing to feature selection and classification model selection. Our study aimed to extend the previous research and focused on classical music to investigate the following three research questions:

1. For Western and Chinese classical MER datasets, what kind of combination algorithm of pre-processing and feature selection methods can achieve the optimal effect of emotion regression prediction?
2. For Western and Chinese classical music, which acoustic features should be selected as being the most culturally representative and effective for participants' MER, respectively?
3. Based on extracted feature sets for different music datasets, what are the differences in the influences of different music elements on emotion regression prediction for different datasets?

Figure 1 shows the workflow of emotional feature set analysis using regression prediction designed to investigate the research questions. The findings from our study try to establish a connection between musical elements and emotion perception in a cross-cultural context, which is useful for musicians to understand and apply when composing different emotional music.

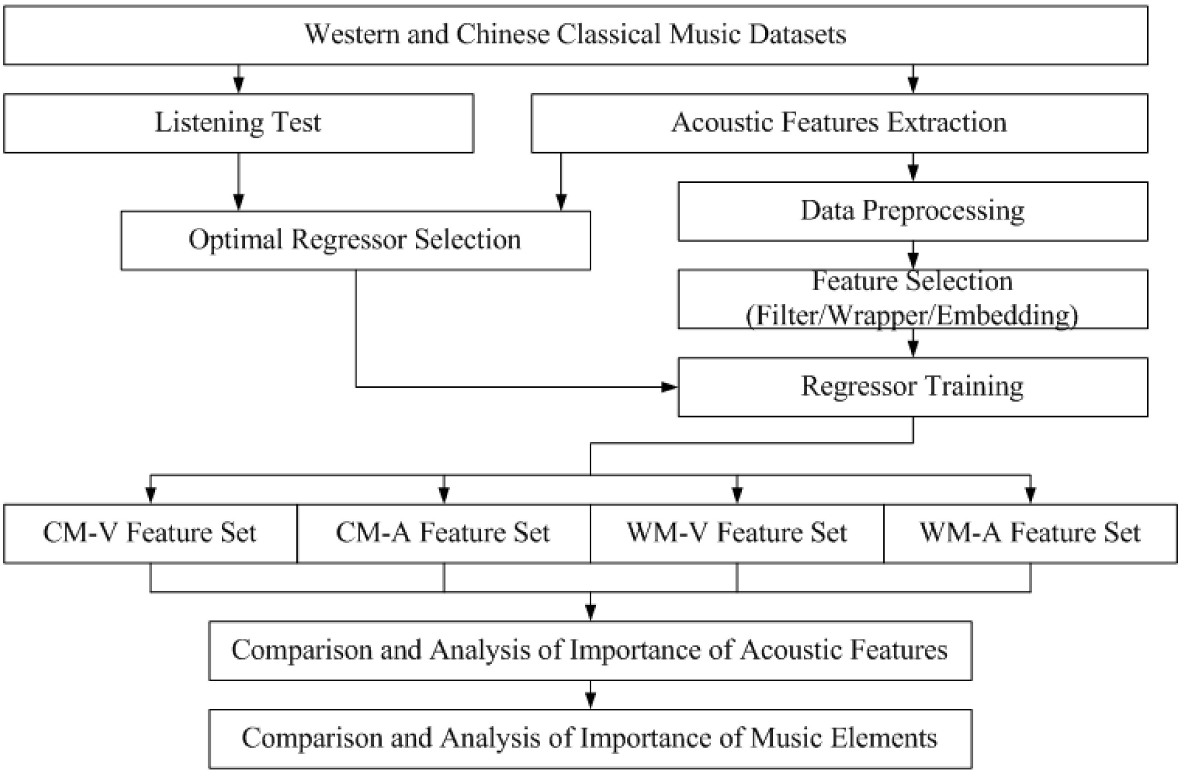

Western and Chinese Classical Music Datasets
Listening Test
Acoustic Features Extraction
Data Preprocessing
Optimal Regressor Selection
Feature Selection (Filter/Wrapper/Embedding)
Regressor Training
CM-V Feature Set
CM-A Feature Set
WM-V Feature Set
WM-A Feature Set
Comparison and Analysis of Importance of Acoustic Features
Comparison and Analysis of Importance of Music Elements

Figure 1. Workflow of emotional feature sets analysis using regression prediction. Acronyms: CM = Chinese classical music, WM = Western classical music, V = valence, A = arousal.

The remainder of this paper is organized as follows. Section 2 introduces the datasets used in this study and the method for extracting acoustic features. Section 3 describes the process of MER predictions for Western and Chinese classical music, along with feature sets for different datasets, which are obtained based on the optimal combination of pre-processing and feature selection methods. The experimental results and detailed discussions are presented in Section 4. Then, we draw conclusions by summarizing the findings of this study in Section 5.

2. The Datasets and Acoustic Features

2.1. The Datasets

Dataset quality depends on three factors: the number of excerpts in the dataset, the number of ratings per excerpt, and the method by which the data are annotated. We focus on these three factors to guarantee the validity of our datasets.

- Western Classical Music Dataset

The emoMusic [27] and Soundtracks [28] datasets that focus on music emotional analysis were selected as Western classical music datasets. The emoMusic dataset consists of 1000 representative songs selected from the Free Music Archive (FMA), with a 45 s clip excerpted from each song. The V-A model was adopted as the emotional model. More than 350 participants (72% from the USA) on Amazon Mechanical Turk were recruited to participate in this annotation after passing a qualification test. Participants were asked to score the perceived emotion rather than evoking one. Each participant annotated 3 songs on arousal and valence dimensions on a continuous scale ranging from 1 to 9, and each excerpt was annotated by a minimum of 10 participants. The average scores of each excerpt across participants were taken as the ground truth data. Krippendorff's alpha was calculated to measure the inter-annotation agreement, which indicated "fair" agreement (0.32 for valence and 0.35 for arousal). From this dataset, 78 excerpts of Western classical music were chosen for our study, including solo and ensemble music.

The Soundtracks dataset contains 360 musical excerpts from a large selection of film soundtracks, with individual excerpts being between 10 and 30 s long. Twelve Western musicians who had studied Western musical instruments for 10 years or more were recruited to choose excerpts of different target emotions, such as happy, angry, sad, and calm. No excerpt was to contain lyrics, dialogue, or sound effects, and each was to be a Western music ensemble or instrumental solo. A three-dimensional emotion model including valence, tension arousal, and energy arousal was adopted [29], and the same group of musicians took part in this annotation to rate the three dimensions on a continuous scale of 1–7. Cronbach's alpha was calculated to measure the inter-annotation agreement, which indicated that all scales had good internal consistency (0.92 for valence, 0.90 for energy arousal, and 0.93 for tension arousal). The average scores across all participants were taken as the ground truth. Because Pearson's correlation between valence and tension arousal was -0.91, the data of valence and energy arousal were selected to represent the ground truth of the V-A model. From this dataset, 360 musical excerpts were selected for our study.

Overall, the total number of excerpts in our Western classical music dataset was 438.

- Chinese Classical Music Dataset

We created the Chinese classical music dataset for this study, consisting of 500 musical excerpts from famous Chinese classical music albums, national instrumental music compilation discs, and online collected folk music albums, including instrumental solo and ensemble music of bowed chordophones, plucked chordophones, aerophones, and membranophones. The length of each excerpt was approximately 30 s with stable emotion and complete phrases. The stimuli were stored on a laptop computer with Windows 10 and played back through Audio-technica CKM77 headphones. All stimuli were calibrated for loudness by volunteers. The monitor level was adjusted to a comfortable level by the participants. A total of 20 participants with Chinese backgrounds majoring in audio technology annotated the valence and arousal emotions on a continuous scale of 1–9. Cronbach's alpha indicated good inter-annotation agreement (0.95 for valence and 0.93 for arousal). The average values across all participants were used as the ground truth.

In summary, the total number of musical excerpts in our cross-cultural dataset was 938.

2.2. Acoustic Features

The current music acoustic features that can be extracted mainly include the underlying physical features and high-level semantic features [30]. The underlying physical features are based on the time-frequency attributes of an audio signal, including spectral, temporal, and spectro-temporal features. High-level semantic features are abstract semantic features extracted by signal processing, image processing, and other methods, which can depict the inherent elements of music. In this study, we selected the acoustic features that had clear physical meanings to clarify the mechanism of emotional perception and classified them into four categories based on music elements: timbre, pitch, loudness, and rhythm. Although the physical meaning of MFCC is not very clear, many researchers have confirmed that it is an effective feature for MER. Therefore, it was also included in this research [31,32]. The list of acoustic features is presented in Table 1.

Table 1. List of acoustic features selected for our study.

Category	Feature	Dimensionality
Timbre	Spectral characteristics—Centroid, Complexity, Decrease, Entropy, Skewness, Kurtosis, RMS, Rolloff, Strongpeak, Spread, Contrast Coeffs (1–6), Irregularity, Spectral Flux, High Frequency Content (HFC)	78
	Spectral characteristics of ERB Bands—Crest, Flatness, Kurtosis, Skewness, Spread	20
	Temporal characteristics—Lowenergy, ZCR	5
	Mel-frequency cepstral coefficients (MFCC)	52
	Perception characteristic—Dissonance	4
Pitch	Harmonic Pitch Class Profile (HPCP), HPCP Entropy	148
	Harmonic Change Detection Function (HCDF)	6
	Tuning-related characteristics—Tuning Diatonic Strength, Tuning Equal Tempered Deviation, Tuning Frequency, Tuning Nontempered Energy Ratio	4
Loudness	Dynamic Complexity	1
	Silence Rate (30 dB/60 dB)	8
	ERB Bands Energy	160
	Spectral Bands Energy—High, Middle High, Middle Low, Low	16
Rhythm	Onset Rate, BPM, Tempo	6
	BPM Histogram—First Peak BPM, First Peak Spread, First Weight, Second Peak BPM, Second Peak Spread, Second Weight	24
	Beats Loudness Band	24
	Danceability	1

Essentia [33] and MIRtoolbox [34] were employed to extract acoustic features. Low-level acoustic features are based mostly on the global static features that describe the entire piece of music, yet the time-varying characteristics of low-level acoustic features are the main factors affecting the emotional perception of music. The features extracted through Essentia were computed over frames of approximately 25 ms with a frame step of 10 ms. In the MIRtoolbox, the frame size was 50 ms, and an overlap of 50% between successive frames was used. To reflect these time-varying characteristics, statistical values of features were introduced as an extension of static features. Essentia has algorithms for computing the mean, variance (var), mean of first-order difference (dmean), and variance of first-order difference (dvar), whereas MIRtoolbox calculates the mean, standard deviation (std), envelope, periodic frequency (PeriodFre), and period entropy (PeriodAmp) as time-varying statistics. There were 557 dimensions of acoustic features in total.

- Timbre

Timbre is a multi-dimensional property [35,36]. For our study, five types of timbre features were extracted, as shown in Table 1. General spectral characteristics can describe the shape and structure of the signal spectrum. Spectral characteristics of ERB bands were chosen to show the characteristics of human auditory perception. Roughness is calculated to represent dissonance [34].

- Pitch

The pitch features extracted in this study refer to all features related to the tone of the sound. The fundamental frequency is defined as the lowest frequency of a harmonic stationary audio signal, which in turn can be qualified as a tonal sound. In music, tonality is a system that organizes the notes of a musical scale according to musical criteria. Moreover, tonality is related to the notion of harmonicity [37]. The harmonic change detection function (HCDF), which is used to represent the change in harmonic content in music, describes the change in harmony between consecutive frames. The harmonic pitch class profile (HPCP) is a 36-dimension vector that represents the intensities of 36 subdivisions of the 12 semitone pitch classes (corresponding to notes from A to G#).

- Loudness

Loudness represents the strength of the music, which is related to the objective amplitude of sound and the subjective psychological perception of hearing [30]. Therefore, the features of loudness were extracted from both objective and subjective perspectives. Dynamic complexity is defined as the average absolute deviation from the global loudness level estimated on the dB scale. It is related to the dynamic range and the amount of fluctuation in loudness. Silence rate describes the proportion of silent segments in the music. Spectral band energy is the energy of each of five subbands spanning the entire spectrum. ERB band energy is the frequency band energy through ERB filter banks.

- Rhythm

Rhythm is a significant feature of music that describes the tempo and beat. The tempo-related features and beat-related features were extracted. The BPM histogram characteristics provide a general tempo perspective and summarize the beat tempos present in music. The beat-loudness band computes the spectrum energy of beats in an audio signal given their beat positions. Danceability estimates the danceability of an audio signal, and the algorithm is derived from the detrended fluctuation analysis (DFA) described in [38].

3. Music Emotion Regression Experiment

MER models were devised for valence and arousal of Western and Chinese classical music datasets to obtain the best representative music emotion feature sets. Due to the redundancy and irrelevance within the 557-dimension acoustic features, the optimal combination algorithm of pre-processing and feature selection is essential. First, the optimal regression model was determined based on a cross-cultural dataset, as mentioned previously. Then, five data pre-processing and four feature selection methods were combined in pairs to form 20 combination algorithms for the next step. The optimal combination algorithm was determined through analysis and comparison. Eventually, an optimal regression framework was established. Figure 2 outlines the above procedure.

Later, four feature sets for different datasets were obtained through this optimal framework. Based on the feature sets, the importance of acoustic features and music elements for different emotional dimensions were analyzed and compared in a cross-cultural context.

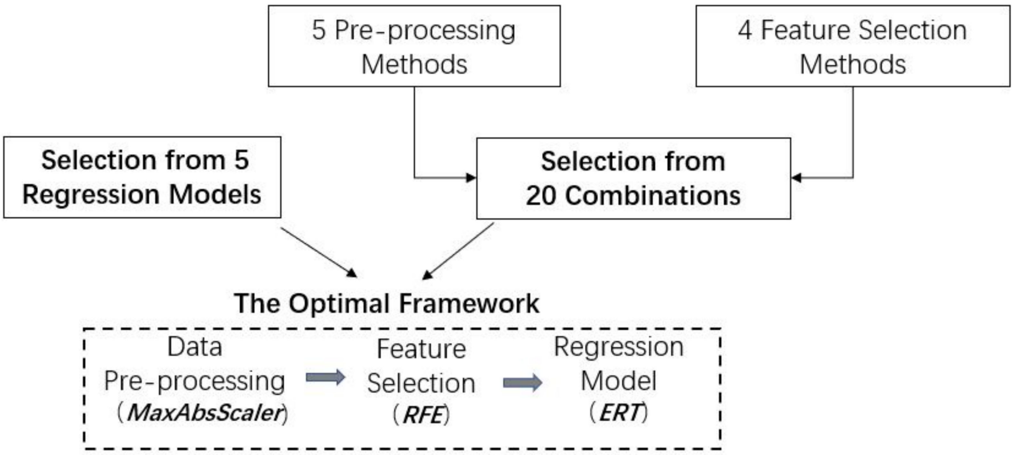

5 Pre-processing Methods
4 Feature Selection Methods
Selection from 5 Regression Models
Selection from 20 Combinations
The Optimal Framework
Data Pre-processing (*MaxAbsScaler*)
Feature Selection (*RFE*)
Regression Model (*ERT*)

Figure 2. The procedure to establish the optimal framework for music emotion regression. Italicized words in the bottom brackets indicate the selected methods of each step, as detailed in the following.

3.1. Optimal Regression Model Selection

An appropriate regression model is a prerequisite for music emotion regression prediction. Five typical machine learning regression models were selected, and the 557-dimension acoustic features in Table 1 were extracted from a cross-cultural dataset of 938 excerpts to compare and analyze the regression effects for valence and arousal. The five regression models were linear regression (LR), support vector regression (SVR), extremely randomized tree regression (ERT), XGBoost, and K-nearest neighbor (KNN). They were all implemented in the Python library of Scikit-Learn, and adopted in this paper with default parameters [39]. The R^2 scores as the evaluation standard are presented in Table 2, and were obtained through the five-fold cross-validation method. It was found that compared with other algorithms, ERT had the best emotion regression prediction for valence and arousal. Thus, it was chosen as the regression model in subsequent experiments.

Table 2. R^2 scores of five different regression models.

Dataset/Classifier	LR	SVR	XGBoost	ERT	KNN
Valence	−1.37	−0.01	0.62	0.66	−0.15
Arousal	−1.37	−0.01	0.39	0.52	−0.11

3.2. Optimal Combination Algorithm for Pre-Processing and Feature Selection

Because the 557-dimension acoustic features were extracted by different toolboxes and each feature had its own physical significance, there were large dynamic range differences or dimensional differences between feature values, which would seriously reduce the performance of regression prediction. Data pre-processing can expand the discrimination of features and improve the accuracy of the prediction. Five mainstream pre-processing methods in Scikit-Learn were adopted with default parameters: StandardScaler, MinMaxScaler, MaxAbsScaler, QuantileTransformer using a uniform distribution (QTU), and QuantileTransformer using Gaussian distribution (QTG).

Generally, feature selection methods can be divided into three categories: wrapper, filter, and embedded. We chose four feature selection methods in Scikit-Learn: a filter method using correlation measurement (Filter-C), a filter method using information entropy measurement (Filter-I), a wrapper method using recursive feature elimination (RFE) based on extremely randomized trees (Wrapper), and an embedded method based on extremely randomized trees (Embedded). The RFE-based wrapper method uses backward selection, which starts with all features and successively removes features if performance improves. Note that when a wrapper approach is used, the feature selection block in Figure 1 actually contains within it a module for regressor training and performance evaluation.

Using the extremely randomized tree method as the basic regression model, 20 combinations of pre-processing and feature selection were evaluated to obtain the optimal method. The R^2 scores are shown in Table 3, with the best combinations for the different datasets underlined. It was found that MaxAbsScaler performed best for two datasets (Western Music Valence and Chinese Music Arousal) and QTU and MinMaxScaler were the optimal methods for the remaining datasets. When MaxAbsScaler was taken as a pre-processing method for all datasets, regression performance was not significantly reduced compared with the QTU or MinMaxScaler as best method. Therefore, the MaxAbsScaler was selected as the optimal pre-treatment method for all datasets.

Table 3. R^2 scores of different combinations of pre-processing and feature selection methods, the best combinations for the different datasets underlined.

Dataset	Method	StandardScaler	MinMaxScaler	MaxAbsScaler	QTU	QTG
Western Music Valence	Filter-C	0.5743	0.5623	0.6289	0.6492	0.6725
	Filter-I	0.6520	0.6401	0.6698	0.6244	0.6693
	Wrapper	0.6442	0.6697	0.6804	0.6614	0.6232
	Embedded	0.6718	0.6798	0.6772	0.6497	0.6520
Western Music Arousal	Filter-C	0.6164	0.5268	0.5277	0.5178	0.5917
	Filter-I	0.6140	0.6037	0.5326	0.5861	0.5822
	Wrapper	0.5377	0.6193	0.5967	0.6036	0.5797
	Embedded	0.5917	0.5659	0.5927	0.5931	0.5821
Chinese Music Valence	Filter-C	0.5851	0.6055	0.6248	0.5964	0.5658
	Filter-I	0.5561	0.5895	0.6084	0.5161	0.6247
	Wrapper	0.6101	0.6424	0.6707	0.6747	0.6189
	Embedded	0.6628	0.5429	0.6306	0.6909	0.6737
Chinese Music Arousal	Filter-C	0.6661	0.6280	0.6712	0.6155	0.5823
	Filter-I	0.5746	0.4715	0.6027	0.6325	0.5890
	Wrapper	0.6786	0.6654	0.6821	0.6589	0.6344
	Embedded	0.6113	0.6222	0.6146	0.6787	0.6069

The wrapper method was optimal for three datasets (Western Music Valence, Western Music Arousal, and Chinese Music Arousal). Therefore, the combination of the optimal pre-processing and feature selection methods is MaxAbsScaler and the wrapper method using RFE based on extremely randomized trees.

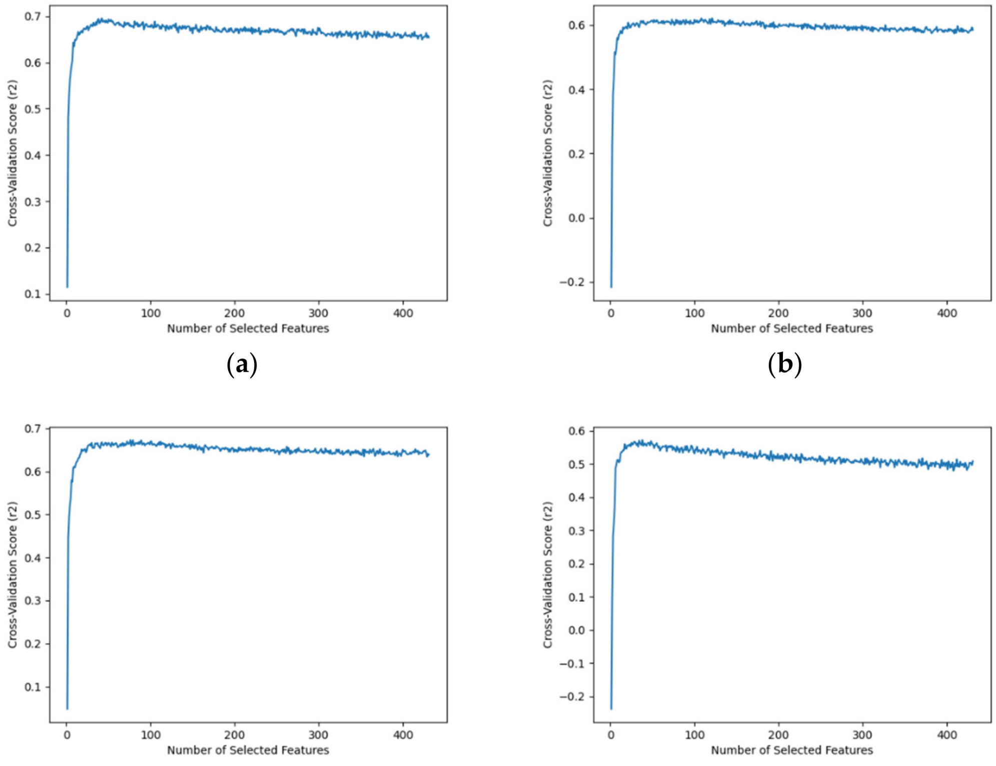
3.3. Determination of Music Emotion Feature Sets

In Section 3.1, the preliminarily optimal combination of pre-processing and feature selection was detailed. This section introduces the specific method to determine the music emotional feature set according to the combination. First, a data cleaning process was applied and 18 acoustic features with many missing values in the datasets were removed. Then, pre-processing was carried out using the MaxAbsScaler method.

The feature selection stage is divided into two steps. In the first step, univariate analysis was carried out for each dataset. The filtering feature selection based on the correlation attribute evaluation method was used to remove 20% of the features that had the least influence on the regression results, reducing the number of features for the next step.

In the second step, RFE was carried out using extremely randomized trees. In this algorithm, the prediction model was trained on the original features, and then the importance weight of each feature was obtained and sorted. The feature with the least weight was removed from the feature set, and this procedure was called recursively until only one feature was left. Then, a five-fold cross-validation method was adopted. The number of features that best ensured a high recognition rate was determined by calculating the regression R^2 score after each iteration.

To select the number of features, two methods were adopted. One was the global optimal feature number, which is the number of features corresponding to the point where the maximum value of R^2 is found on the RFE curve, as shown in Figure 3. However, this could lead to a lot of redundant information, resulting in an excessive number of features. Consequently, the local optimal feature number was selected, which is the number of features corresponding to the first peak point on the curve envelope. In this algorithm, the minimum horizontal distance in feature numbers between neighboring peaks was 20. Smaller peaks were removed until the condition was fulfilled for all remaining peaks. In this way, a more appropriate number of features was obtained.



(**a**) (**b**)

(**c**) (**d**)

Figure 3. The RFE curve of five-fold cross validation of feature selection. (**a**) Western Classical Music—Valence. (**b**) Western Classical Music—Arousal. (**c**) Chinese Classical Music—Valence. (**d**) Chinese Classical Music—Arousal.

Figure 3 shows the RFE curve of five-fold cross-validation, and the evaluation score is R^2. Table 3 displays the global optimal feature number, local optimal feature number, and their R^2 scores. As can be seen in Figure 3, the RFE curves of the four datasets were very flat, and all of them had very small variances. Table 4 shows that the number of locally optimal features of the four datasets was substantially reduced compared with that of the global optimal features, whereas the R^2 scores of five-fold cross-validation barely changed. Therefore, we finally selected the local optimal feature set for the following discussion. The experimental results are presented in Section 4.

Table 4. Global and local optimal R^2 scores and number of features.

Data Subset	Global Optimal Score (R^2)	Global Optimal Feature Number	Local Optimal Score (R^2)	Local Optimal Feature Number
Western Classical Music—V	0.695	42	0.674	21
Western Classical Music—A	0.572	38	0.560	17
Chinese Classical Music—V	0.621	121	0.607	21
Chinese Classical Music—A	0.673	76	0.652	22

4. Results and Discussion

4.1. Feature Sets of Western and Chinese Classical Music Emotion Regression

The feature sets for the valence and arousal of Western and Chinese classical music datasets are displayed in Table 5. The results indicate that some acoustic features, such as MFCC, spectral complexity, and spectral contrast, are important for all scenarios, whereas some other acoustic features such as spectral flux and dissonance are unique for specific emotion regression prediction. We further analyze and classify acoustic features as culturally universal or culturally specific in the next step.

Table 5. Music emotional feature sets of Western and Chinese classical music.

Data Subset	Feature Sets
Western Classical Music—Valence	HCDF_mean, HPCP_dmean_28, Onset_Rate, MFCC_dmean_1, MFCC_dmean_2, Spectral_complexity_dmean, Spectral_flux _mean Spectral_complexity_dvar, Spectral_complexity_mean, Spectral_complexity_var, Spectral_entropy_dmean, Spectral_entropy_mean, Spectral_contrast_coeffs_dmean_1, Spectral_contrast_coeffs_dmean_2, Spectral_contrast_coeffs_dmean_6, Spectral_contrast_coeffs_dvar_1, Spectral_contrast_coeffs_dvar_6, Spectral_contrast_coeffs_mean_6, Dissonance_mean, Beats_loudness_band_ratio_mean_5, Spectral_centroid_dmean
Western Classical Music—Arousal	Tuning_diatonic_strength, Tuning_equal_tempered_deviation, Spectral_contrast_coeffs_mean_2, Spectral_contrast_coeffs_mean_3, Spectral_contrast_coeffs_dmean_2, Spectral_contrast_coeffs_dmean_6, Spectral_contrast_coeffs_dvar_6, Spectral_skewness_dvar, Spectral_complexity_var, HPCP_entropy_dmean, HPCP_dmean_12, HPCP_dvar_12, HPCP_entropy_mean, HCDF_Mean, HCDF_PeriodAmp, MFCC_var_5, MFCC_var_6
Chinese Classical Music—Valence	Onset_rate, HCDF_mean, Spectral_strongpeak_dmean, Spectral_complexity_dmean, HFC_dmean, Spectral_flux_mean, Spectral_flux_dmean, Spectral_contrast_coeffs_mean_2, Spectral_contrast_coeffs_dmean_4, Spectral_strongpeak_dvar, Spectral_strongpeak_var, Beats_loudness_band_ratio_mean_3, BPM_histogram_first_peak_weight_mean, Silence_rate_60 dB_var, Dynamic_complexity, MFCC_dmean_2, MFCC_mean_6, MFCC_dmean_9, MFCC_dvar_13, Dissonance_mean
Chinese Classical Music—Arousal	Spectral_flux_mean, Spectral_complexity_dmean, Dissonance _mean, Dissonance_var, HCDF_PeriodAmp, Spectral_entropy_mean, AHFC_dmean, Spectral_skewness_mean, Spectral_strongpeak_dmean, Spectral_contrast_coeffs_mean_2, Onset_rate, HCDF_mean, ERBbands_mean_30, ERBbands_spread_dmean, Silence_rate_60 dB_var, HPCP_entropy_mean, HPCP_dvar_1, HPCP_dmean_1, MFCC_dmean_2, MFCC_mean_1, MFCC_dmean_7, MFCC_dmean_8

4.2. Importance of Acoustic Features for Different Feature Sets

The importance of acoustic features in the four music emotion feature sets was further analyzed based on the importance coefficients obtained through RFE feature selection using extremely randomized trees. The percentage of the importance of acoustic features for each feature set is displayed in Figure 4, in which the different colors of acoustic features represent different musical elements. For multi-dimensional acoustic features, the importance coefficients of multiple dimensions are summed to obtain the percentage of importance. In each feature set, the top five important features are shown in Table 6.

For the Western classical music dataset, the number of important acoustic features is relatively small, and there are some features with a much higher percentage of importance, such as spectral contrast. For the Chinese classical music dataset, the number of important features is relatively large, and the percentage of feature importance is more evenly distributed. Some important acoustic features were analyzed and compared further to discover the emotional perception mechanism for Western and Chinese classical music datasets.

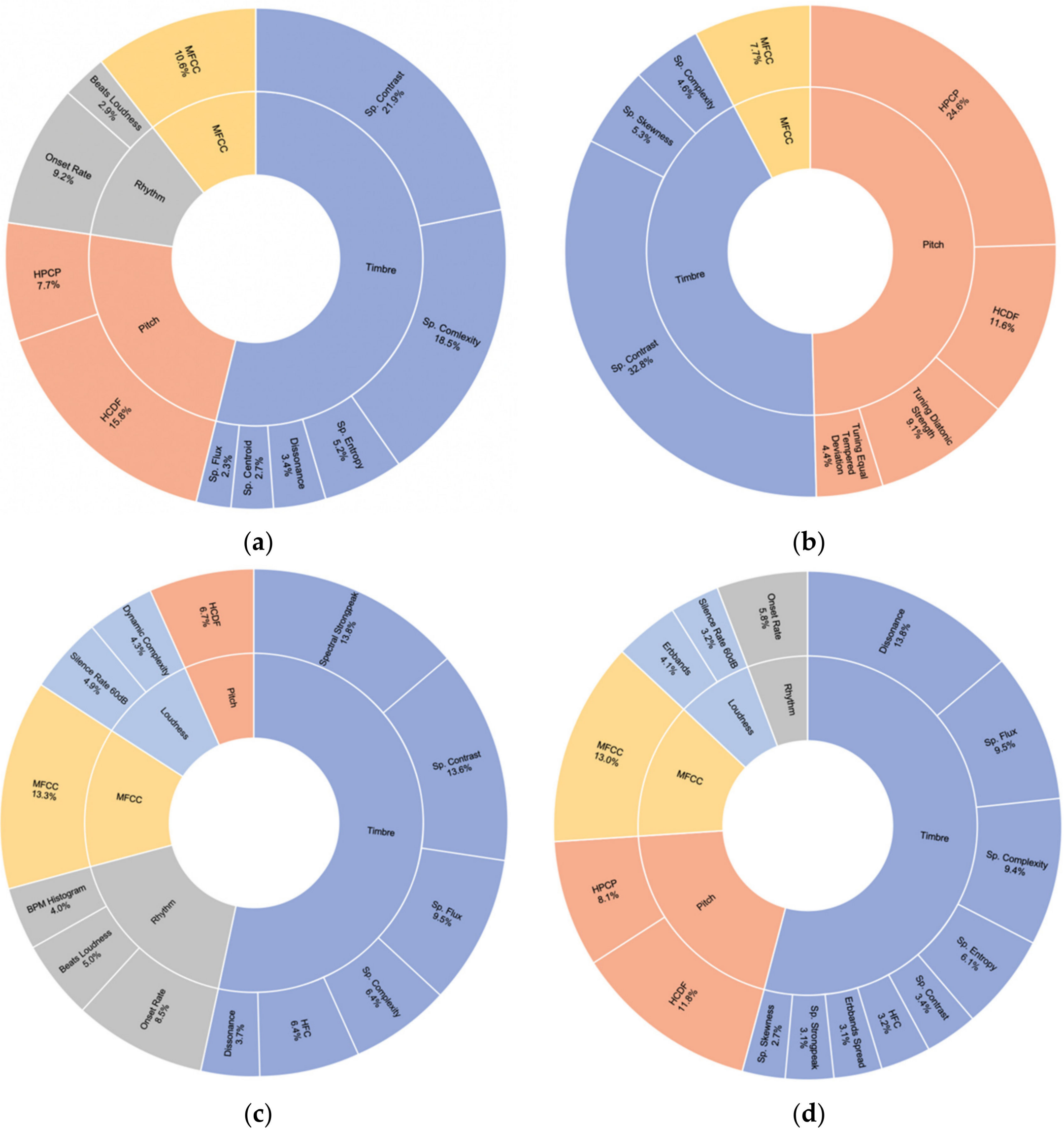


Figure 4. Percentage of importance of acoustic features for each feature set. (**a**) Western Classical Music—Valence. (**b**) Western Classical Music—Arousal. (**c**) Chinese Classical Music—Valence. (**d**) Chinese Classical Music—Arousal.

Table 6. Top five important acoustic features and their percentage of importance for each feature set.

	1	2	3	4	5
W-V	Spectral contrast 21.9%	Spectral complexity 18.5%	HCDF 15.8%	MFCC 10.6%	Onset rate 9.2%
W-A	Spectral contrast 32.8%	HPCP 24.6%	HCDF 11.6%	Tuning diatonic strength 9.1%	MFCC 7.7%
C-V	Spectral strongpeak 13.8%	Spectral contrast 13.6%	MFCC 13.3%	Spectral flux 9.5%	Onset rate 8.5%
C-A	Dissonance 13.6%	MFCC 13.0%	HCDF 11.8%	Spectral flux 9.5%	Spectral complexity 9.4%

Acronyms: W = Western classical music, C = Chinese classical music; V = valence, A = arousal.

- Spectral Contrast

As shown in Figure 5, the spectral contrast is the difference in strength between the spectral peaks and valleys in each spectral subband divided by octave-scale filters. For most music, the peaks of the spectrum roughly correspond to the harmonic components, whereas the valleys represent most of the non-harmonic components or noise. Consequently, spectral contrast can reflect the relative distribution of harmonic and non-harmonic components in the spectrum [40].

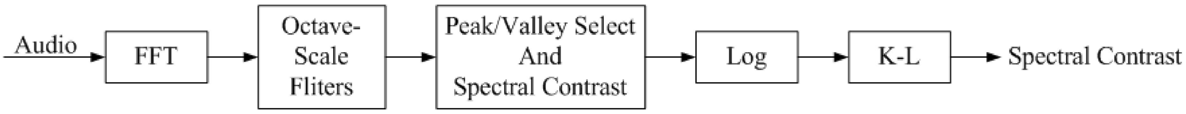


Figure 5. Flow diagram of spectral contrast.

Spectral contrast is very important for the valence and arousal perception of Western classical music. It can be inferred that the contrast of peak and valley values of the music spectrum affects people's emotional judgement. The ratio of the harmonic component to the non-harmonic component plays an important role in emotion recognition of Western classical music, perhaps because Western musical instruments have more regular harmonics than Chinese instruments, making spectral contrast more significant for emotional perception of Western classical music.

- HCDF

HCDF is the harmony change detection function. The feature extraction method is shown in Figure 6. This feature represents the change in harmony between consecutive frames, and its essence is the flux of the tonal centroid [41]. HCDF is significant for both valence and arousal perception of Western and Chinese classical music, which indicates that HCDF is culturally universal. It can be concluded that the more frequent the harmony changes in music, the more obvious the changes in pleasure and activation perceived by people.

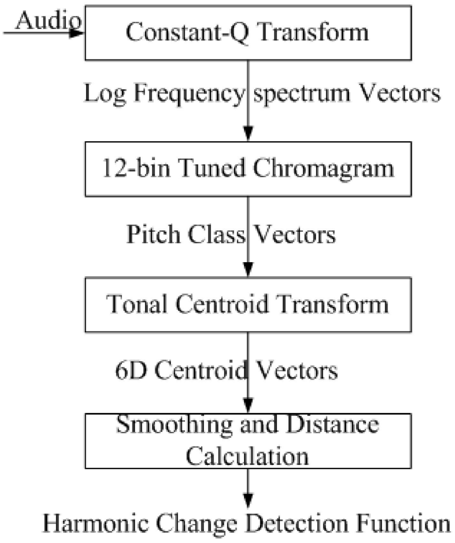


Figure 6. Flow diagram of the HCDF system.

- Onset Rate

The onset rate represents the number of notes per second, which indirectly describes the tempo of music. In the algorithm, peak-picking is automatically performed on the onset detection curve to estimate the positions of the notes [42]. In both Western and Chinese classical music, the influence of onset rate on valence perception is stronger than that of arousal, which indicates that valence is highly related to the tempo of music.

- Spectral Complexity

Spectral complexity is a measure of the number of spectral peaks: the higher the number of spectral peaks, the higher the spectral complexity. Laurier et al. showed that relaxed music usually had lower spectral complexity, whereas non-relaxed music had higher spectral complexity [43]. Spectral complexity plays an important role in all scenarios, especially for the valence perception of Western classical music and the arousal perception of Chinese classical music, which indicates that the influence of spectral complexity on the emotional perception of Western and Chinese classical music is a somewhat different.

- Spectral Flux

Spectral flux is the difference in the spectrum amplitude between consecutive frames. This feature measures the rate of change in the spectrum structure and reflects the dynamic characteristics of the audio signal. Spectral flux is very important for the valence and arousal perception of Chinese classical music. Therefore, it is inferred that the greater the spectral flux, the more intense the spectral-domain energy of the music changes over time, and the higher the valence and arousal perception for Chinese classical music. Spectral flux is a culturally specific feature for Chinese classical music.

- Dissonance

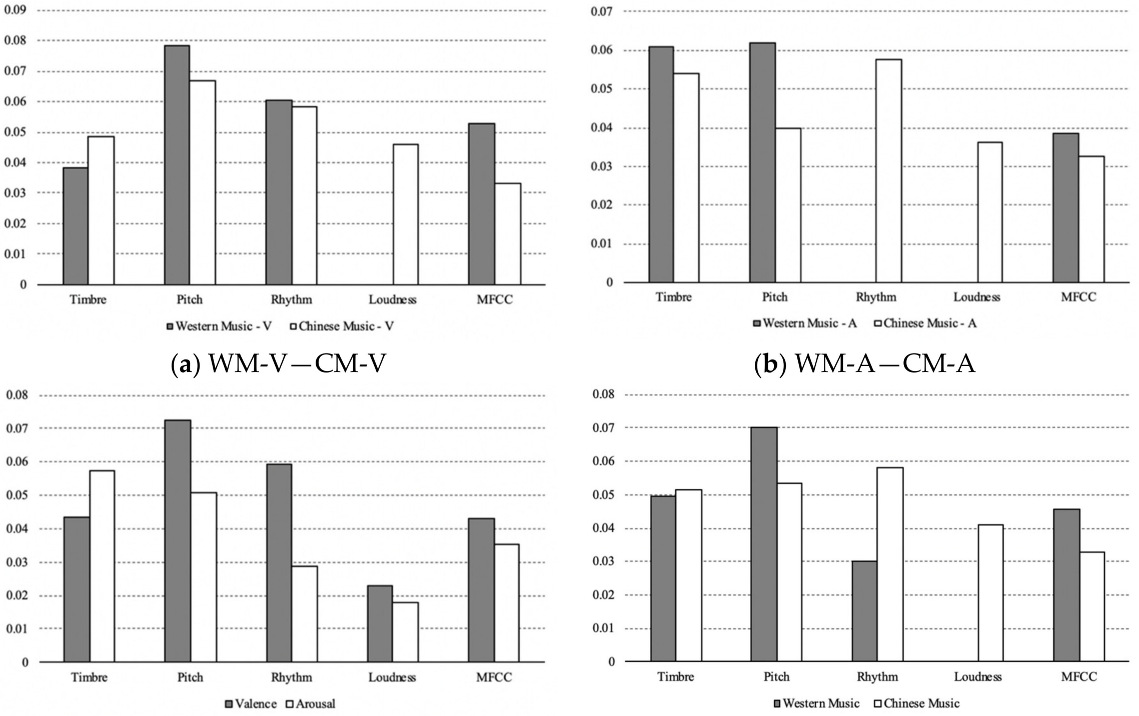
Dissonance is a timbre perceptual feature, and is also defined as roughness, based on a harmonic series of complex tones to study the influence of spectral features on sensory dissonance [44]. Dissonance is the most important feature for arousal perception of Chinese classical music. Because Chinese classical music emphasizes the personalization of musical instruments rather than the fusion [45], dissonance is an important aesthetic feature of Chinese classical music [46]. It can be concluded that dissonance is a culturally specific feature of Chinese classical music.

4.3. Importance of Musical Elements for Different Feature Sets

The importance of musical elements for different feature sets was compared and analyzed according to the corresponding relationship between acoustic features and musical elements, combined with human auditory perception. The importance coefficients of different musical elements, including timbre, loudness, pitch, and rhythm for each feature set, were calculated as the average values of the importance coefficients of acoustic features belonging to each musical element. The reason for introducing the average value of importance is to avoid distortion by the number of statistical features or multi-dimensional features. Although the MFCC reflected the timbre characteristic of the audio signal to a certain extent and should be classified as a timbre feature, the MFCC of a single dimension had no corresponding relationship in timbre perception. Therefore, the timbre features were divided into MFCC and other timbre features.

The importance of different musical elements on the valence of Western and Chinese classical music is compared in Figure 7a. For Western classical music, the musical elements, ranked in order of the importance from high to low, are pitch, rhythm, MFCC, and timbre, whereas the ranking for Chinese classical music elements is pitch, rhythm, timbre, loudness, and MFCC. Although pitch and rhythm are very important for both forms of classical music, loudness is specifically significant for Chinese classical music. As shown in Figure 7b, the musical elements that affect the arousal perception of Western classical music are very different from those for Chinese classical music. Pitch and timbre are the most important musical elements for Western classical music, whereas all musical elements have a significant impact on Chinese classical music, especially rhythm.

Based on our cross-cultural dataset, Figure 7c indicates that pitch and rhythm have a greater influence on valence perception, whereas timbre has a greater influence on arousal perception. This result is consistent with our usual experience of listening to music. By comparison, from the results shown in Figure 7d of emotion perception between Western and Chinese classical music, it can be seen that pitch is the most important musical element for Western classical music, whereas almost all musical elements are equally important for Chinese classical music. Furthermore, it suggests that the features related to rhythm and loudness of Chinese classical music are more important than those of Western classical music. A possible reason is that Western classical music has a strict and perfect rhythm theory, with most repertoires following a regular rhythm and speed, whereas Chinese classical music places more emphasis on personalization and personification. Some Chinese classical music does not even have a fixed rhythm, and the dynamic fluctuation is very large, which leads to the fact that rhythm and loudness affect people's emotional perception of Chinese classical music to a large extent.



(**a**) WM-V—CM-V

(**b**) WM-A—CM-A

(**c**) Valence—Arousal

(**d**) WM—CM

Figure 7. The distribution of the average importance of musical elements in four feature sets: (**a**) comparing the results of valence between Western and Chinese classical music, (**b**) comparing the results of arousal between Western and Chinese classical music, (**c**) comparing the results of valence and arousal based on the cross-cultural dataset, and (**d**) comparing the results of Western and Chinese classical music.

5. Conclusions

A comparative study of emotional feature sets between Western and Chinese classical music was conducted using emotion regression prediction. To compare the combination algorithms of pre-processing and feature selection methods, the optimal combination algorithm was determined. Twenty-dimension feature sets for valence and arousal of Western and Chinese classical music datasets were extracted from over 500-dimension acoustic features based on emotion regression prediction. The importance of representative acoustic features and musical elements of different feature sets was analyzed and compared, and the following conclusions were obtained:

- Based on our cross-cultural dataset, the optimal combination algorithm of pre-processing and feature selection is MaxAbsScaler pre-processing and the wrapper method using RFE based on extremely randomized trees.
- The number of important acoustic features for Western classical music dataset is larger than that for Chinese classical music dataset. For the Western classical music dataset, the distribution of the importance of acoustic features is mainly concentrated

on several features such as spectral contrast, whereas for the Chinese classical music dataset, the difference in the importance of each feature is relatively small.

- Spectral contrast is the most significant feature for both valence and arousal perception of the Western classical music dataset. HCDF is significant for both valence and arousal perception of the Western and Chinese classical music datasets, which indicates that HCDF is culturally universal. Regardless of whether it is a Western or Chinese classical music dataset, the onset rate's influence on valence perception is stronger than its influence on arousal. Compared with Western classical music dataset, spectral flux is more important for valence and arousal perception of the Chinese classical music dataset. Dissonance is a culturally specific feature of the Chinese classical music dataset.
- For valence, although pitch and rhythm are very important for both cultures' classical music dataset, loudness is specifically significant for Chinese classical music dataset. For arousal, pitch and timbre are the most important musical elements for the Western classical music dataset, whereas all musical elements have a significant impact on the Chinese classical music dataset, especially rhythm features.

These research conclusions may provide some inspiration in the field of MIR, MER, music creation, and artificial intelligence composition. In particular, for cross-cultural music research, this paper can provide the data basis and acoustical perspective for further discussion. Future expansions of this study may include evaluating more datasets and applying a more detailed emotional model for evaluation to accumulate more experience in exploring cross-cultural music emotion perception.

Author Contributions: Conceptualization, X.W.; Data curation, L.W. and L.X.; Project administration, X.W. and L.X.; Supervision, L.X.; Writing—original draft, L.W.; Writing—review & editing, X.W. and L.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Humanity and Social Science Youth Foundation of the Ministry of Education of China (15YJCZH178) and the Science Foundation of the Communication University of China (HG1608-1) to Xin Wang and Major Project of Beijing Social Science Foundation (20ZD20) to Xiuqing Jia.

Institutional Review Board Statement: The study was approved by the Research Ethics Committee of Communication University of China.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kim, Y.E.; Schmidt, E.M.; Migneco, R.; Morton, B.G.; Richardson, P.; Scott, J.; Speck, J.A.; Turnbull, D. State of the Art Report: Music Emotion Recognition: A State of the Art Review. In Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), Utrecht, The Netherlands, 9–13 August 2010; pp. 255–266.
2. Yang, X.; Dong, Y.; Li, J. Review of Data Features-Based Music Emotion Recognition Methods. *Multimed. Syst.* **2018**, *24*, 365–389. [CrossRef]
3. Yang, Y.H.; Chen, H.H. *Music Emotion Recognition*; CRC Press: Boca Raton, FL, USA, 2011.
4. Mitrovic, D.; Zeppelzauer, M.; Breiteneder, C. Features for Content-Based Audio Retrieval. *Adv. Comput.* **2010**, *78*, 71–150.
5. Juslin, P.N.; Sloboda, J.A. *Handbook of Music and Emotion: Theory, Research, Applications*; Oxford University Press: New York, NY, USA, 2010.
6. Russell, J.A. A Circumplex Model of Affect. *J. Pers. Soc. Psychol.* **1980**, *39*, 1161–1178. [CrossRef]
7. Posner, J.; Russell, J.A.; Peterson, B.S. The Circumplex Model of Affect: An Integrative Approach to Affective Neuroscience, Cognitive Development and Psychopathology. *Dev. Psychopathol.* **2005**, *17*, 715–734. [CrossRef] [PubMed]
8. Eerola, T.; Vuoskoski, J.K. A Review of Music and Emotion Studies: Approaches, Emotion Models, and Stimuli. *Music Percept.* **2013**, *30*, 307–340. [CrossRef]
9. Panda, R.; Rocha, B.; Paiva, R.P. Music Emotion Recognition with Standard and Melodic Audio Features. *Appl. Artif. Intell.* **2015**, *30*, 313–334. [CrossRef]
10. Downie, J.S.; Ehmann, A.F.; Bay, M.; Jones, M.C. The Music Information Retrieval Evaluation Exchange: Some Observations and Insights. In *Advances in Music Information Retrieval*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 93–115.

11. Schmidt, E.M.; Turnbull, D.; Kim, Y.E. Feature Selection for Content-Based, Time-Varying Musical Emotion Regression. In Proceedings of the International Conference on Multimedia Information Retrieval, Philadelphia, PA, USA, 29–31 March 2010; Association for Computing Machinery: New York, NY, USA; pp. 267–274.
12. Liu, Y.; Liu, Y.; Zhao, Y.; Hua, K.A. What Strikes the Strings of Your Heart?—Feature Mining for Music Emotion Analysis. *IEEE Trans. Affect. Comput.* **2015**, *6*, 247–260. [CrossRef]
13. Zhang, J.L.; Huang, X.L.; Yang, L.F.; Xu, Y.; Sun, S.T. Feature Selection and Feature Learning in Arousal Dimension of Music Emotion by Using Shrinkage Methods. *Multimed. Syst.* **2017**, *23*, 251–264. [CrossRef]
14. Grekow, J. Audio Features Dedicated to the Detection and Tracking of Arousal and Valence in Musical Compositions. *J. Inf. Telecommun.* **2018**, *2*, 322–333. [CrossRef]
15. Rolls, E.T. Neurobiological Foundations of Aesthetics and Art. *New Ideas Psychol.* **2017**, *47*, 121–135. [CrossRef]
16. Kılıç, B.; Aydın, S. Classification of Contrasting Discrete Emotional States Indicated by EEG Based Graph Theoretical Network Measures. *Neuroinformatics* **2022**, 1–15. [CrossRef] [PubMed]
17. Jeon, B.; Kim, C.; Kim, A.; Kim, D.; Park, J.; Ha, J.-W. Music Emotion Recognition via End-to-End Multimodal Neural Networks. In Proceedings of the RecSys'17 Posters Proceedings, Como, Italy, 27–31 July 2017.
18. Hizlisoy, S.; Yildirim, S.; Tufekci, Z. Music emotion recognition using convolutional long short term memory deep neural networks. *Eng. Sci. Technol. Int. J.* **2020**, *24*, 760–767. [CrossRef]
19. Orjesek, R.; Jarina, R.; Chmulik, M. End-to-End Music Emotion Variation Detection using Iteratively Reconstructed Deep Features. *Multimed. Tools Appl.* **2022**, *81*, 5017–5031. [CrossRef]
20. Lee, J.H.; Hill, T.; Work, L. What Does Music Mood Mean for Real Users? In Proceedings of the iConference, Toronto, ON, Canada, 7–10 February 2012; pp. 112–119.
21. Lee, J.H.; Hu, X. Cross-Cultural Similarities and Differences in Music Mood Perception. In Proceedings of the iConference, Berlin, Germany, 4–7 March 2014; pp. 259–269.
22. Wu, W.; Xie, L. Discriminating Mood Taxonomy of Chinese Traditional Music and Western Classical Music with Content Feature Sets. In Proceedings of the IEEE Congress on Image and Signal Processing, Sanya, China, 27–30 May 2008; pp. 148–152.
23. Hu, X.; Lee, J.H.; Choi, K.; Downie, J.S. A Cross-Cultural Study on the Mood of K-POP Songs. In Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR), Taipei, Taiwan, 27–31 October 2014; pp. 385–390.
24. Yang, Y.H.; Hu, X. Cross-Cultural Music Mood Classification: A Comparison on English and Chinese Songs. In Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR), Porto, Portugal, 8–12 October 2012; pp. 19–24.
25. Yang, Y.H.; Lin, Y.C.; Su, Y.F.; Chen, H.H. A regression approach to music emotion recognition. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 448–457. [CrossRef]
26. Hu, X.; Yang, Y.H. Cross-Dataset and Cross-Cultural Music Mood Prediction: A Case on Western and Chinese Pop Songs. *IEEE Trans. Affect. Comput.* **2017**, *8*, 228–240.
27. Soleymani, M.; Caro, M.N.; Schmidt, E.M.; Sha, C.Y.; Yang, Y.H. 1000 Songs for Emotional Analysis of Music. In Proceedings of the ACM International Workshop on Crowdsourcing for Multimedia, Barcelona, Spain, 21–25 October 2013; pp. 1–6.
28. Eerola, T.; Vuoskoski, J.K. A Comparison of the Discrete and Dimensional Models of Emotion in Music. *Psychol. Music* **2010**, *39*, 18–49.
29. Schimmack, U.; Grob, A. Dimensional Models of Core Affect: A Quantitative Comparison by Means of Structural Equation Modeling. *Eur. J. Pers.* **2000**, *14*, 325–345.
30. Lerch, A. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*; Wiley-IEEE Press: Piscataway, NJ, USA, 2012.
31. Vijayavani, E.; Suganya, P.; Lavanya, S.; Elakiya, E. Emotion Recognition Based on MFCC Features using SVM. *Int. J. Adv. Res. Comput. Sci. Manag. Stud.* **2014**, *2*, 31–36.
32. Panda, R.; Malheiro, R.; Paiva, R. Novel Audio Features for Music Emotion Recognition. *IEEE Trans. Affect. Comput.* **2018**, *11*, 614–626. [CrossRef]
33. Homepage—Essentia 2.1-Beta6-Dev Documentation. Available online: https://essentia.upf.edu (accessed on 5 April 2020).
34. Lartillot, O.; Toiviainen, P. A MATLAB Toolbox for Musical Feature Extraction from Audio. In Proceedings of the 10th International Conference on Digital Audio Effects (DAFx), Bordeaux, France, 10–15 September 2007; pp. 237–244.
35. Zwicker, E.; Feldtkeller, R. *Das Ohr als Nachrichtenempfänger*, 2nd ed.; S. Hirzel Verlag: Stuttgart, Germany, 1967.
36. Moore, B. *An Introduction to the Psychology of Hearing*, 4th ed.; Academic Press: London, UK, 1997.
37. Eyben, F. *Real-Time Speech and Music Classification by Large Audio Feature Space Extraction*; Springer: Cham, Switzerland, 2016.
38. Streich, S.; Herrera, P. Detrended Fluctuation Analysis of Music Signals: Danceability Estimation and further Semantic Characterization. In Proceedings of the AES 118th Convention, Barcelona, Spain, 28–31 May; 2005; pp. 765–773.
39. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
40. Jiang, D.; Lu, L.; Zhang, H.; Tao, J.; Cai, L. Music Type Classification by Spectral Contrast Feature. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'02), Lausanne, Switzerland, 26–29 August 2002; pp. 113–116.
41. Harte, C.; Sandler, M.; Gasser, M. Detecting Harmonic Change in Musical Audio. In Proceedings of the ACM International Multimedia Conference and Exhibition, Santa Barbara, CA, USA, 23–27 October 2006; pp. 21–26.

42. Brossier, P.M.; Bello, J.P.; Plumbley, M.D. Fast Labelling of Notes in Music Signals. In Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR), Barcelona, Spain, 10–14 October 2004; pp. 331–336.
43. Laurier, C.; Meyers, O.; Serrà, J.; Blech, M.; Herrera, P.; Serra, X. Indexing Music by Mood: Design and Integration of an Automatic Content-Based Annotator. *Multimed. Tools Appl.* **2009**, *48*, 161–184. [CrossRef]
44. Sethares, W.A. *Tuning, Timber, Spectrum, Scale*; Springer: Berlin/Heidelberg, Germany, 2004.
45. Zhou, W. A study on change of the aesthetics of timbre of Chinese pop music. In Proceedings of the 1st Asia International Symposium on Arts, Literature, Language and Culture, Fuzhou, China, 25–26 May 2019; pp. 105–110.
46. Xin, W.; Zihou, M. The consonance evaluation method of Chinese plucking instruments. *Acta Acust.* **2013**, *38*, 486–492.