Continuous Response to Perceived Affective Intentions in Music Listening: Differences in Music Training

Lena Heng, Claire Wei, & Stephen McAdams
McGill University, Montreal, Canada

Perception of affective intentions in music is a complex, yet commonplace, phenomenon that is influenced by listeners' experience with different musical traditions and the style of the music. This study explores the continuous responses of listeners to a piece of Chinese orchestral music and how their musical backgrounds influence affective responses, as well as the acoustic and musical features used in the perception of affective intentions. Three groups of listeners (trained in Chinese and Western art music traditions and nonmusicians) were presented with a 15.5-minute piece of music and responded continuously on a unidimensional emotional intensity scale and a two-dimensional valence and arousal interface in separate blocks. Functional data analysis compared differences between listener groups' response profiles. Time series analysis explored how each listener group utilized different acoustic and musical features over the course of the music. Results show significant differences between listener groups over different sections of the music. Valence responses diverge more than arousal or emotional intensity responses. The perception of affective intentions in music is influenced by the degree of familiarity listeners have with a musical tradition, the content implicated in the music, and the complex sonic environment created by the composer's creation and the musicians' interpretation.

Received: April 10, 2024, accepted March 3, 2025.

Key words: timbre features, musical features, musical affect, continuous response, cross-cultural differences

Many different ideas, including affective intentions, can be expressed through music. Musical meaning emerges from the interpretation of organized musical information. The communication of emotions or affect takes place when listeners perceive emotional meaning that is expressed by performers in music (Juslin, 2013a, 2013b). The interpretation of musical information, however, does not rely predominantly on single and direct sound-to-meaning relationships, but arises more through relationships when sounds come together. Different elements in music carry important information.

The affective content in music has often been framed in terms of discrete emotions using, for example, Ekman's (1992) basic emotions or dimensional scales such as Russell's (1980) two-dimensional circumplex model of arousal and valence (activation-pleasantness). A large amount of previous work has explored the relationship between musical and acoustic features with perceived affect in listeners. Musical structures such as pitch relations are perceptually salient and provide important information for listeners (e.g., Gabrielsson & Lindström, 2010; Krumhansl, 1998; Krumhansl & Kessler, 1982). The perception of consonance also plays an important role in the music listening process—combinations of tones that are consonant are perceived as more positively valenced than dissonant ones (Harrison & Pearce, 2020). Timbre has also been shown to carry perceptually useful information for affect encoding. Eerola and colleagues (2009) attempted to predict affective ratings in music from audio using multivariate regression models and found that several factors can be used to predict the ratings: timbral elements of spectral centroid, spectral spread, and spectral entropy; harmonic characteristics such as key clarity, harmonic change, and a measure of "majorness"; registral cues; rhythmic characteristics including clarity of pulsation and rhythmic periodicity; articulation; and musical structure estimated through the computation of novelty curves. Acoustic features have also been found to combine in different ways to provide cues for decoding affective intentions. In a previous study, Heng and McAdams (2024) presented musical stimuli to participants having different musical backgrounds (Chinese musicians, Western musicians, and nonmusicians). The stimuli consisted of a single melody played with different affective intentions on Western and Chinese instruments and presented as isolated notes, measures, and phrases. Participants were asked to make a global judgment of their perceived affective intention on the two-dimensional scale of valence and arousal. With more information from increasing musical

Music Perception, volume 43, issue 3, pp. 235–264, issn 0730-7829, electronic issn 1533-8312. DOI: https://doi.org/10.1525/mp.2025.2488483

context, participants were more accurate in their affective judgments. Furthermore, features do not inform affective intent independently. A particular acoustic feature might contribute positively to an affective dimension when combined with a particular set of features, but negatively to the same affective dimension when it is combined with a different set of acoustic features.

In addition to the affective dimensions of valence and arousal, emotional intensity appears anecdotally to be an important aspect of perceived responses to music and could also be an important aspect in a listener's perception of music. Emotional intensity is defined in this study as how strong an emotion is, regardless of the type of emotion experienced. Following Schubert (2001) and McAdams et al. (2004), it is likely that the unidimensional emotional intensity scale taps into the arousal or activity level component of emotion rather than the valence component. Even though there are several studies exploring the intensity of emotions in music (e.g., Krumhansl, 1998; McAdams et al., 2004; Sloboda & Lehmann, 2001), surprisingly few have directly compared emotional intensity with other dimensions of affect. Reisenzein (1994) explored the relationship between emotional intensity and the two-dimensional pleasure-arousal scale with various affect terms (words only). The results suggest that participants' understanding of emotional intensity is related to the strength of each of these emotion concepts on the pleasure-arousal scale. Exploring listeners' perceived emotional intensity when listening to a piece of music and how it relates to the strength of both valence and arousal or only arousal alone will contribute to increasing our understanding of the quality and structure of perceived affective intentions in music.

Studies examining these various musical dimensions frequently use abstract experimental stimuli such as chords, melodies, or instruments on their own so as to isolate effects due to experimental manipulation more clearly. Less has been done with longer and more complex musical stimuli, or actual musical pieces where musical context plays a larger role and interactions and reactions to these stimuli could be much more complex. Schubert observes that an important, fundamental property of music is also the fact that it "requires time to exist [and] any definition of music must include this property, even if by implication" (2010, p. 223).

Functional data and time series analyses are two useful techniques for modeling dynamic changes and continuous responses. They allow for the testing of hypotheses about the relationship between musical cues and affective responses even in the context of long and complex pieces of music. Vines and colleagues (2005), for example, emphasized the importance of contextual cues in playing an important role in a listener's perception. The perception of a stimulus depends both on what has been played before and on expectations of what is to come after. In considering the dynamic relationship between acoustic and musical elements in music listening, Dean and Bailes (2010) examined acoustic influences on listeners' perception of change, arousal, and valence. Listeners' perception of continuous change was found to be strongly influenced by sound intensity. They investigated spectral flatness (which captures the tonalness vs. noisiness of a sound) as an acoustic feature that could potentially influence listeners' perception because it is affected by every spectral component and thus provides a global description of timbre. They also attempted to model spectral centroid as a predictive feature for valence and arousal (Dean & Bailes, 2011). Both spectral centroid and flatness, however, did not appear to have direct overall influence on listeners' perceptions of valence and arousal for the piece they studied.

Thompson and Balkwill's (2010) cue-redundancy model proposes that listeners across different cultures can appreciate affective qualities of unfamiliar music by attending to commonalities, whereas listeners who are familiar with a musical style should find it easier to decode emotional meaning in that music because they can draw from both culture-specific and common psychophysical cues. Egermann and colleagues (2015) found that listeners from different cultures (Pygmy and Western) had more similar arousal responses when presented with Western and Pygmy music but differed in their valence responses. They suggested that "music-induced arousal responses appeared to be based on rather universal, culturally independent response mechanisms" (p. 8). Similarly, global valence responses are also found to be influenced more by enculturation and explicit music training (Heng & McAdams, 2024).

It is likely that a listener's cultural background and experience play an important part in their comprehension of music because appraising these musical events is dependent on a whole contextual framework of social, structural, source, and identity roles (Thompson et al., 2022). As many previous studies mainly have explored global responses to a single musical stimulus, the type and the amount of difference in continuous responses is an interesting area for investigation, and is addressed by the current study.

With contextual information playing an important role in influencing listener responses, it is therefore crucial in this study to have a lengthy piece of music that

allows enough time for potential affective content to change and evolve. In addition to highly controlled musical materials for behavioral studies, it is also important to incorporate real pieces of music to demonstrate the generalizability of research findings. With these factors in mind, a Chinese orchestral work composed by Wang Chenwei (2017) was selected for this study. This piece of music is written for the Chinese orchestra, but with influences from Arabic music. The version that participants heard was a premier of the work by the Taipei Chinese orchestra.

Our approach to functional data analysis involves averaging over groups of participants to produce normative affect profiles that can be predicted using functional data analysis. To explore listeners' musical background in their perception of affective intentions, we split our analyses according to the musical background of the participants; in particular, we compare trained Chinese musicians, trained Western musicians, and nonmusicians.

Specifically, we test the following hypotheses:

H1. Listeners with different musical backgrounds, but a similar non-Western cultural environment, will have different response profiles of their perceived affect based on a similar pattern of responses to global judgments of perceived affect in music in previous work (Heng & McAdams, 2024). This hypothesis is evaluated with functional data analysis, comparing response curves across the three scales.

H2. Similar to previous studies with listeners from different cultures (e.g., Egermann et al., 2015), valence responses will diverge more between different listener groups than will arousal responses. This hypothesis is also tested with functional data analysis.

H3. A listener's perceived emotional intensity will be related to arousal rather than to the distance away from the origin in the two-dimensional valence-arousal model. Schubert (2001) has proposed that emotional intensity is primarily related to arousal rather than to valence (p. 401). We test that this is the case against more extreme emotions along both dimensions near the perimeter of the 2D emotion space.

H4. If H2 is supported, surface acoustic features within the musical sound will relate more directly to perceived arousal than perceived valence based on results that demonstrate that arousal responses show less variability across culture and training than valence (Egermann et al., 2015). This hypothesis is tested through time series analysis examining acoustic variables as contributors to perceptual responses through impulse-response techniques.

H5. Being more culturally bound, musical cues that are learned, such as modal structures, metrical relations, and so on, will exert a greater influence on listeners' perceived valence ratings than on their arousal ratings. This hypothesis will also be tested through time series analyses.

There have not been many studies examining listeners' continuous responses to music, and even less delving into differences in experience with different musical traditions. Therefore, much of this study is also exploratory, attempting to probe into which and how acoustic and musical features and listeners' musical backgrounds influence their continuous responses over the course of a lengthy piece of music.

Method

PARTICIPANTS

Three groups of listeners with different musical backgrounds were recruited from Singapore. By having only participants from Singapore, effects of differences in language and other socio-cultural factors can be held relatively constant so that any differences can be attributed more directly to music training. The first group were musicians trained in the Chinese music tradition (henceforth CHM) ($n = 30$; M years of music training = 11.2 years; SD music training = 5.49; 15 females and 15 males; M age = 35.5 years, SD age = 8.88). The second group were musicians trained in the Western music tradition (henceforth WM) ($n = 30$; M years of music training = 11.9 years; SD music training = 5.92; 17 females, 11 males, and 2 unspecified; M age = 27.4 years, SD age = 10.23). The final group were nonmusicians (henceforth NM) ($n = 30$; M years of music training = 0.16 years; SD music training = 0.29; 15 females and 15 males; M age = 35.5 years, SD age = 12.30). All participants reported having no problems with hearing.

STIMULI

The Chinese orchestra is a relatively modern invention that is less than a century old (e.g., Chan, 2003). It utilizes traditional Chinese instruments and combines a composition aesthetic derived from the rich folk and traditional musics around China, as well as composition and orchestration techniques borrowed from Western orchestral music. A Chinese orchestral work composed by Wang Chenwei (2017) was used for the listening

experiment.[1] The work utilizes Chinese instruments but was written with modes and metrical concepts borrowed from Arabic music and a programmatic content based on Aleppo and the Syrian war. This piece was selected for several reasons. First, it has a rich programmatic and narrative content as defined by the composer, so that affective intentions might be more easily elicited. CHM are expected to be more familiar with the instrumental timbres as compared to the other two listener groups. If familiarity with instrumental timbres and orchestration styles influences perception of affective intentions, we would expect to observe a difference between CHM and the other two groups. With formal music training, CHM and WM are likely to have internalized a repertoire of formalized musical rules and commonly used tropes. Even though this piece of music uses modes and meters that are less familiar to both groups of musician listeners, they might interpret it with respect to the musical traditions they are familiar with. This would set them apart from the nonmusicians who would likely respond more to the surface acoustic features of the music as compared to the musicians. Instead of selecting a piece of Chinese orchestral music that uses more conventional Chinese musical forms and elements, using one with Arabic musical features will ensure that it does not bias any particular musician group with regards to knowledge of particular formal musical rules. A third reason for selecting this piece of music is that it is relatively less well-known to the listeners in Singapore as it was commissioned and premiered by the Taipei Chinese Orchestra and was only published in a recording by the orchestra in December 2021. The recording was provided by the composer with a sample rate and bit depth of 48 kHz/24-bit (wav, stereo). The entire piece of music is 930 seconds in duration.

PROCEDURE

Participants were each given a unique link for the experiment where they could access the experimental task online. They were asked to use headphones for the listening task and to complete the experiment on a computer in a quiet space. They first had to click on a box to acknowledge having read and accepted a written informed consent. Participants were then presented with three clips of music corresponding to the lowest, medium, and highest dynamic levels that they might encounter in the music and asked to adjust their volume level so that it was comfortable, and they could still hear the softest clip clearly. They then proceeded to the instructions. First, the difference between perceived and induced affect was explained and participants had to respond to questions to ensure they understood the difference. They were then reminded that they were supposed to only think about the affective intentions they perceived the music to be conveying and not how the music made them feel in this experiment. Next, the experimental interface was introduced with a unidimensional scale for emotional intensity and a two-dimensional scale for valence and arousal. Emotional intensity was described to the participants as the "intensity or force of the emotion the music is trying to communicate". Valence was described to participants as the pleasantness/unpleasantness, or positive/negative aspects of an emotion. Arousal was described as indicating the "excitedness" or the energy level of an emotion. Examples of emotions on the extreme ends of these scales were given. It was also explained to participants that the extreme emotions would be toward the sides of the two-dimensional space, whereas emotions that are more neutral would be close to the centre.

For the unidimensional emotional intensity ratings, participants moved the mouse left and right continuously to indicate the changes in perceived emotional intensity they thought the music was trying to express as it progressed (from "low intensity" on the left to "high intensity" on the right). It did not matter where on the screen they placed their cursors vertically—only the horizontal values were mapped. This also ensured that there would not be any discontinuity in their responses. Participants were able to move their cursors throughout the entire space for the two-dimensional valence (positive-negative) and arousal (low-high) ratings, and a face emoji evolved to indicate the perceived emotion (Figure 1). This interface was adapted from Nagel and colleagues' (2007) EMuJoy interface, which was developed to collect continuous real-time data of perceived emotions in music. The horizontal and vertical positions were continuously captured throughout the entire piece of music. A practice excerpt of another piece of Chinese orchestral music by the same composer, lasting 39 seconds, was presented for participants to get used to the interface and the tasks. The interface for emotional intensity ratings and that for valence and arousal ratings were presented separately during this instruction phase, and participants had to perform a practice trial for each of these interfaces with detailed explanations to ensure they did not confuse the unidimensional emotional intensity ratings with the two-dimensional valence and arousal ratings. Valence and

[1] https://www.youtube.com/watch?v=pReUYEKtLR0&ab_channel=wangchenwei *Note.* This is not the recording that was used in the experiment.

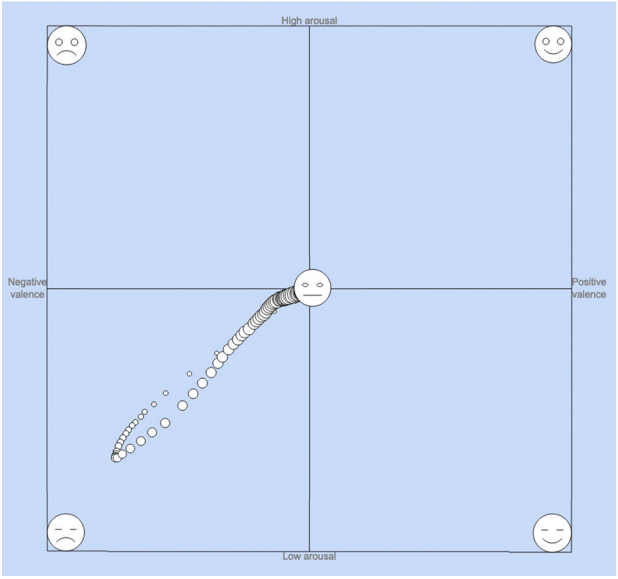

FIGURE 1. Interface for valence and arousal ratings. Face emojis are overlaid on the four extreme corners here to show how it changes according to its position. Only one face emoji was present at the position of the cursor on the interface in the experiment.

arousal ratings were further clarified with the changing emoji face—the eyes narrow when a participant moves their cursor below the maximum y-value; similarly, the smile decreases when the cursor is moved leftwards from the maximum x-value. This way, the participant always had immediate feedback as to the valence and arousal rating they were giving at a particular moment.

The music was then presented for the actual experimental task. All participants performed the emotional intensity rating task first and the two-dimensional valence-arousal rating task after that. The rationale was that rating on a unidimensional scale requires less cognitive load, and with participants not having heard the piece before, responses could be affected if they were presented with the more complex two-dimensional rating task first. This order also served to familiarize participants with the music and allowed them to more easily focus on subtle differences on the two dimensions of valence and arousal they would have to rate upon a second hearing. Participants could choose to take a break in between the two tasks. After completing the experimental tasks, participants filled out a demographic questionnaire gathering information on their age, sex, years and type of music training, and amount of time spent on musical activities. This was then followed by a short debrief, and participants were also encouraged to email the experimenters if they wished to further clarify anything else. Each participant was given monetary compensation for their participation, which took from 40 minutes to an hour. The study was reviewed for ethical compliance by the McGill University Research Ethics Board II.

DATA ANALYSIS

Listeners' continuous responses were collected at a rate of 2 Hz over the entire 930 s of the music. This gave 1,861 observation points per participant for each scale of emotional intensity, valence, and arousal. This sampling rate was used because listener responses appeared to lag behind causal musical events by between around one to three seconds and based on Nyquist's sampling theorem, the sampling rate should be at least twice that of the highest expected frequency (Schubert, 2010). Emotional intensity was coded on a continuous scale from 1 to 9, and the valence and arousal scales were coded in a range from –1 to +1. Statistical analyses were performed in R (Version 4.3.1; R Core Team, 2023). The *fda* package was used for functional data analyses and the *tseries*, *TSA*, *lmtest*, and *vars* packages were used for the time series analyses.

Functional Data Analysis

The first aim of this study was to explore whether different musical backgrounds contributed to differences in emotional intensity, valence, and arousal responses. To answer this question, functional data analysis (FDA) was applied to the data. FDA assumes listeners' continuous responses "reflect a smooth variation [. . .] that could be assessed, in principle, as often as desired, and is therefore a [. . .] *function*" (Ramsay et al., 2009). The collection of ordered data points for each participant were fitted onto a single curve. In other words, instead of considering each datapoint as a discrete observation, FDA treats each curve as arising from a single process. This allows for conventional statistical methods such as regression analyses to be carried out. A brief description of how functional data analysis is applied to the data will be presented.

This technique is a process of constructing continuous functions with parameters that are easy to estimate and that can accommodate nearly any curve feature. The basic functional building blocks called basis functions are combined linearly, similarly to how polynomial functions are linear combinations of monomial functions with specific coefficients. There are several types of basis functions that can be used, but as b-splines are used for this dataset, the details of how the data are prepared and functional data analysis is applied to them will focus on b-splines.

We list the steps taken to obtain the functional data object—the set of functions for the data. The sequence of observations (sampled continuous responses) is divided into intervals—the boundaries between intervals are called break points. At each break point, neighboring polynomials are constrained to have a certain number of derivatives that match. A straight line will only have the function value matching because its derivative is a constant, so the function is continuous, but the slope is not. A linear spline is of order 2 (degree 1 + constant 1), and so on. Therefore, the number of derivatives is the order of the polynomial minus 2, so that the joints at each break point will be smooth with splines of order greater than 2. In this study, the first and second derivatives are explored with cubic splines, so order 6 splines (degree 5 + constant 1) are used.

The number of basis functions is specified as the number of intervals plus the order number of the splines. The optimal number of intervals is obtained by finding the dip in cross-validation error. This is done using leave-one-out cross-validation, where each data point is iteratively used as a validation set, and the model is trained on the remaining data points. Several numbers of basis functions that are used to fit the model for the training set are computed and then used to predict the validation set. The number that minimizes prediction error on the validation set is selected (42 in this case). After defining the basis, coefficients that define the parameters for the basis functions are derived. The more basis functions there are, the better the curve will fit the data. However, as each observation contains intrinsic errors, there is a risk of overfitting. Such functions will also be very rough because they attempt to run through every point. To reduce this, roughness penalties are applied. This means that complex functions are penalized so that there is a balance between fitting the data well and creating a smooth function that represents a continuous underlying process. The curvature of a function is the square of the second derivative of that function. Therefore, the integrated squared second derivative can be used to measure a function's total curvature—the lower it is, the smoother a function is.

The roughness penalty is defined from a functional parameter object, which consists of a basis object, a derivative of order m to be penalized, and a smoothing parameter λ. The optimal λ is chosen by finding the minimal value using generalized cross-validation (GVC) (Craven & Wahba, 1979). This is an extension of leave-one-out cross-validation. In classic linear regression, the best fit is defined by

$$y_i = x(t_i) + \varepsilon_i,\ i = 1, 2, \ldots n,$$

and data fitting is determined by minimizing the sum of squared errors

$$SSE(x) = \sum_{i}^{n} [y_i - x(t_i)]^2$$

A roughness penalty applied will prevent the function from overfitting and avoid singularity problems. Since the integrated squared second derivative measures the function's curvature, minimizing

$$\frac{1}{n}\sum_{i=1}^{n} (f(t_i) - y_i)^2 + \lambda \int_0^t (f^m(u))^2 du$$

controls the tradeoff between roughness and fit of the solution.

Instead of iteratively leaving out each observation and refitting the model, GCV calculates a measure of the mean squared error in regression analysis and takes the form

$$GCV(\lambda) = \left(\frac{n}{n - df(\lambda)}\right)\left(\frac{SSE}{n - df(\lambda)}\right).$$

The estimate of GCV is adjusted firstly by subtracting $df(\lambda)$ from n and by multiplying by $\frac{n}{n-df(\lambda)}$.

The data are fitted by applying this functional parameter with optimal λ to the functional data object.

Initial Data Processing. Participants' responses were captured for the entire 930 s of the music. However, only responses from 20 s to 930 s were included in the analysis. This restriction prevents the initial cursor movement from influencing the results. The cursor always started from the middle of the interface, and its position might not accurately reflect the perceived affective intention at the start of the music. Schubert (2013) has also cautioned about treatment of the initial segment of a continuous response as listeners require a certain amount of time to orient themselves at the start. There is not much contrast in the first 20 s of this piece of music, and the exclusion of the first 20 s simply removed the initial dip of participants moving the cursor from its starting position to where they wanted their responses to be for the three scales of emotional intensity, valence, and arousal. Schubert has also explored what he calls the "afterglow effect," in which participants displayed a great deal of disagreement in ratings after the conclusion of the music. In the present study, participants' responses were collected only up to the point where the music ended, and the position at which their cursor left off was assumed to be the perceived affect at the conclusion of the piece. For functional data analysis the data were

smoothed using b-spline functions of order 6 and a roughness penalty of 0.001 that was obtained using generalized cross-validation to prevent problems of overfitting. As listeners might have different response speeds to the same audio stimulus, phase variation might occur between their curves. Given that the variation between listeners' responses to the music itself is of interest, it is critical that phase variation be removed so that differences between listener responses at the same point in the music can be more accurately explored. Rather than using only specific points for alignment, values throughout the curve are registered. The rationale here is that if the curves differ only in terms of amplitude variation, then their values will be proportional throughout time t. Each point on the registered curve plotted against the target curve should therefore approach a straight line that tends to pass through the origin. This time-warping function was therefore fitted onto the curves after smoothing, and analyses were performed on the registered curves. Figures 2 to 4 show curves before and after registration.

An examination of the random effects from a mixed effects functional data regression reveals only small standard deviations of random intercepts for valence and arousal (all around 0.1 on a scale range of 2). The residual standard deviations for valence and arousal are 0.04 and 0.05, respectively. It appears therefore that individual variations of participants are relatively small compared with between-group variability, and that using an average of participant responses to compare differences between groups should not cause any unexpected issues. The random effects intercepts of emotional intensity, however, have a standard deviation of 0.67 and very large residual standard deviation of 3.47 (on a scale range of 8). As such, interpretation of the emotional intensity curves will be taken with caution.

Acoustic and Musical Features

Using the Timbre Toolbox (Peeters et al., 2011, revised by Kazazis et al., 2021) implemented in the MATLAB environment (The MathWorks Inc., 2022), the entire piece of music was analyzed for several time-varying acoustic features. A short-term Fourier transform was used to obtain a power spectrum from the audio file. A Hann window of length 50 ms with a 50% overlap between successive frames was used. All the acoustic descriptors derived from this input representation that were available in the Timbre Toolbox were selected for analysis. These included spectral centroid, spectral crest, spectral decrease, spectral flatness, spectral flux, spectral kurtosis, spectral roll-off, spectral skewness, spectral slope, spectral spread, and spectral variation. In addition, RMS energy and zero-crossing rate, calculated from the audio signal representation, were also included. Each acoustic descriptor characterizes an aspect of the sound, and as this study is exploratory in terms of determining the features that might influence perceived affective intentions in music, no descriptors were excluded a priori. Table 1 lists these acoustic descriptors and a brief definition of each of them.

Two other features were computed with the MIRToolbox (Lartillot, 2022). A recent study by Harrison and Pearce (2020) found that the MIRToolbox's *mirroughness* measure is the best candidate for estimating consonance when only audio information is available. Its *mirmode* function estimates the modal strength of the music in terms of "majorness" and "minorness." Finally, the onset detection function from a VAMP plugin created by the Centre for Digital Music in Queen Mary, University of London, for the SonicVisualiser (Cannam et al., 2010) was used to estimate note density.

These features are compared with the emotional intensity, valence, and arousal curves to explore whether there were any consistent variations between them to suggest an influence of particular features. They are also used to model impulse response functions as described in the following section.

Time Series Analysis

With this set of data, which are taken sequentially in time, and which are likely autoregressive in nature, time series analysis proves to be another very useful technique to explore the relationships between acoustic and musical features with listener responses. Impulse response functions represent the reactions of endogenous variables to "shocks" hitting the system (Lütkepohl, 2008) and are useful for studying the impact a variable has on a system. Vector autoregressive (VAR) models are used in forecasting by capturing the co-evolution of the variables (Lütkepohl, 2005). To ascertain which set of impulse responses are actually responsible for impacting a system, structural vector autoregressive (SVAR) models are used. These apply restrictions that allow the identification of impacts to a system, and unique impulse responses can be derived for each shock to a system. This is done by modeling the instantaneous relations between the observable variables directly so that a model with uncorrelated residuals can be found. This is necessary so that when impulse responses are constructed, only those that actually reflect what is happening in the system can be highlighted. Impulse response functions are constructed on these SVAR models. The impulse response functions allow us to test the fourth and fifth hypotheses, that

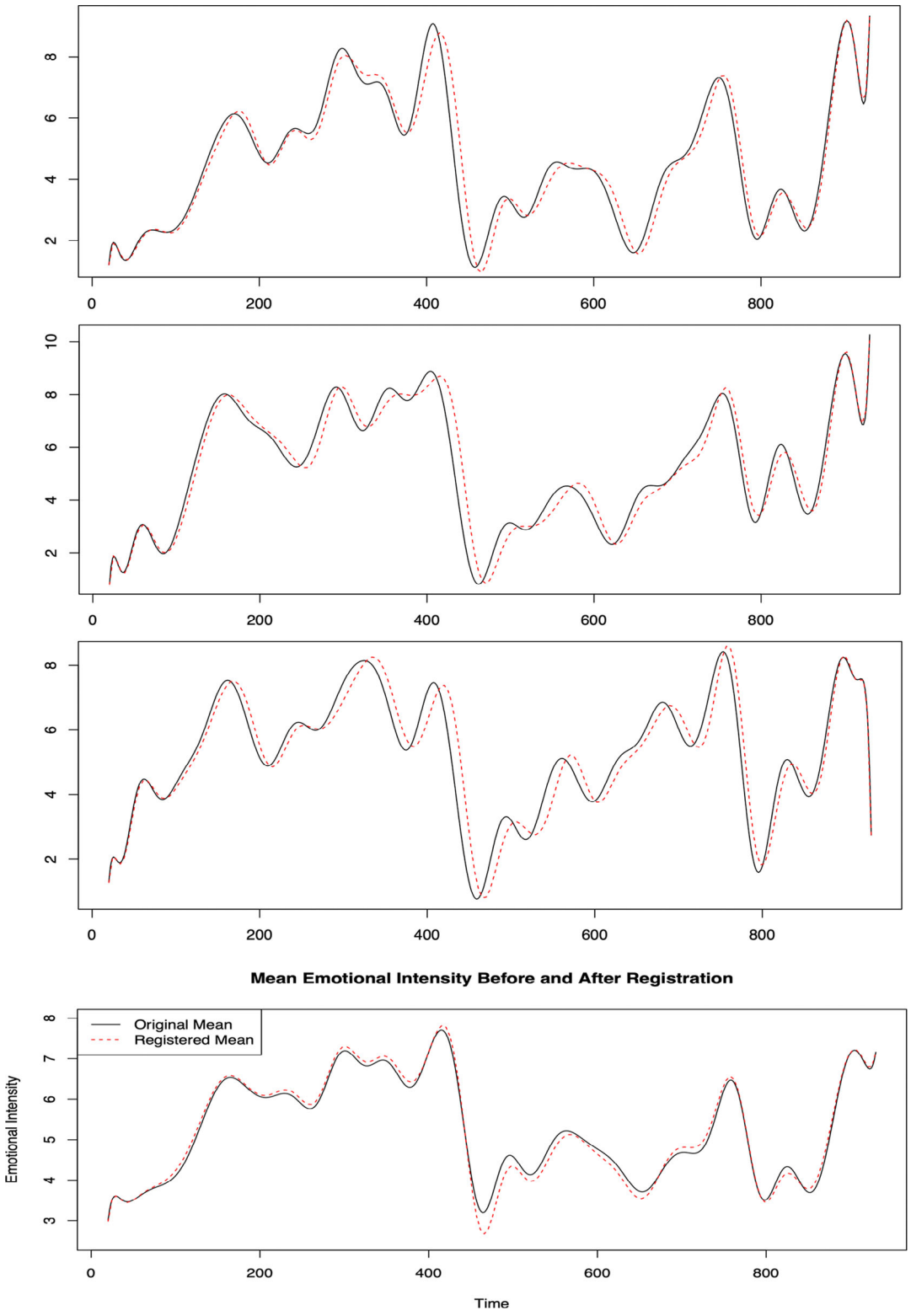

FIGURE 2. Emotional intensity curves before and after registration. The top three panels show examples of individual participant's pre-registered (solid black line) and registered (dashed red line) curves. The fourth panel shows the mean of pre-registered curves (solid black line) and mean of registered curves (dashed red line) across all participants.

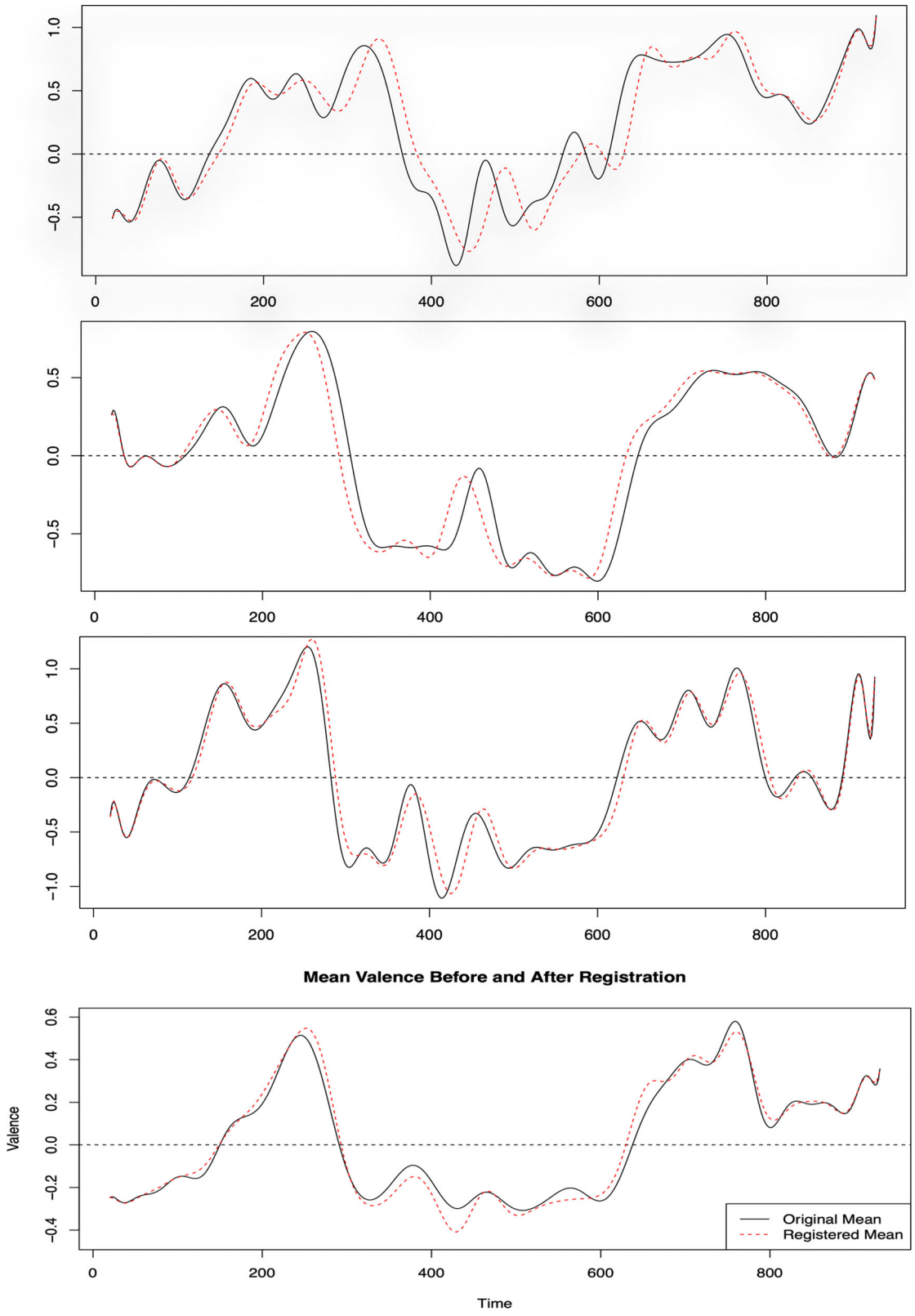

FIGURE 3. Valence curves before and after registration. See Figure 2 caption.

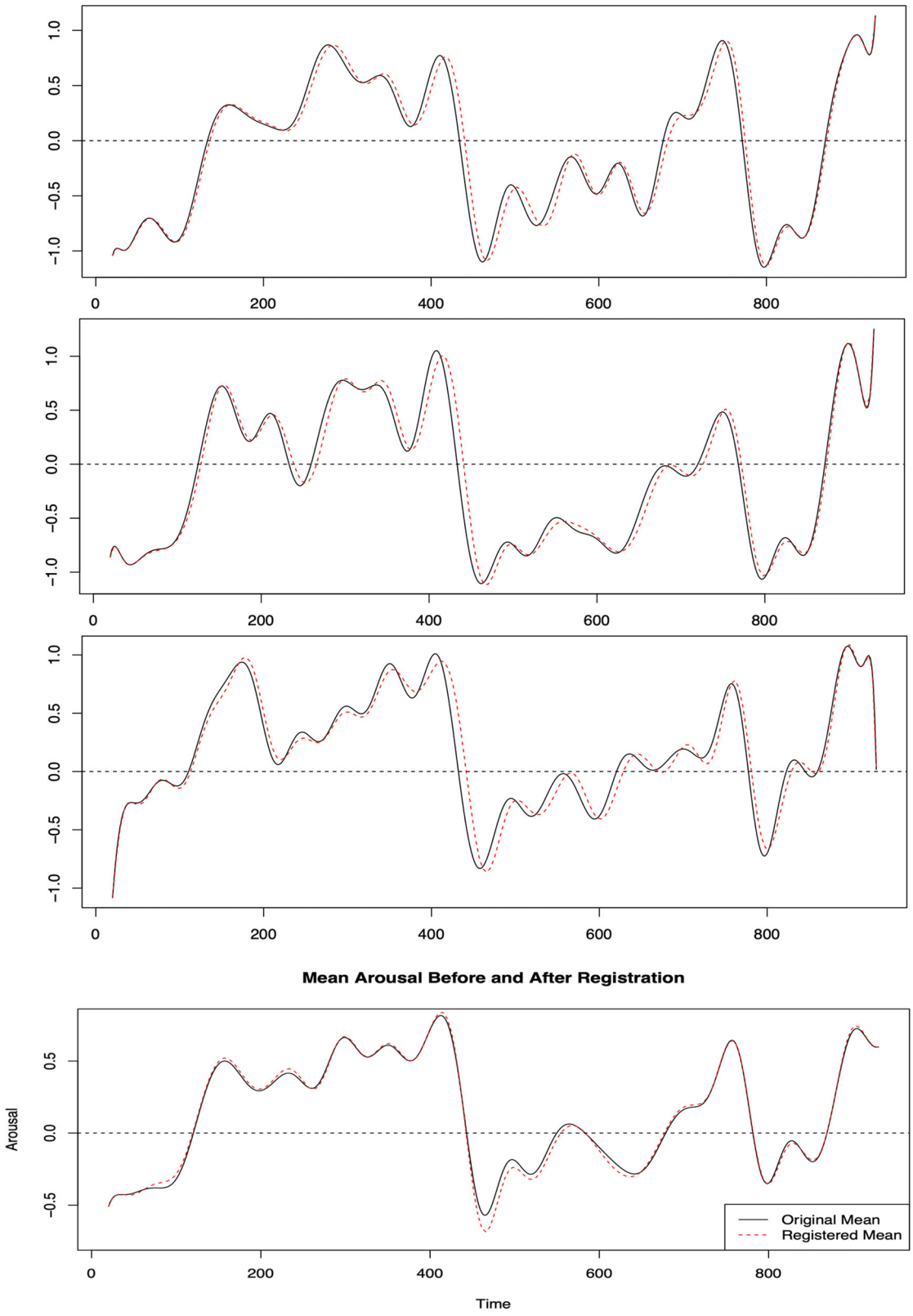

FIGURE 4. Arousal curves before and after registration. See Figure 2 caption.

TABLE 1. *Acoustic Descriptors Used in the Analysis*

Acoustic descriptors	Definition
Spectral centroid	A measure of the spectral center of gravity, an indication of auditory brightness.
Spectral crest	Measures the ratio of the maximum of the spectrum to the arithmetic mean of the spectrum. Spectral crest is an indication of the peakiness of the spectrum. A higher spectral crest indicates more tonality, while a lower spectral crest indicates more noise.
Spectral decrease	Represents the amount of decrease of the spectrum, while emphasizing the slopes of the lower frequencies, and features in instrument recognition.
Spectral flatness	Measures the ratio of the geometric mean to the arithmetic mean of the spectrum. Gives an indication of the "peakiness" of the spectrum—sinusoidal components produce a peakier spectrum, and white noise gives a flat spectrum.
Spectral flux	A measure of how much the power spectrum of a signal changes (arithmetic difference).
Spectral kurtosis	Measures the flatness of the spectrum around its centroid.
Spectral roll-off	The frequency below which 95% of the signal energy is contained.
Spectral skewness	A measure of the asymmetry of the spectrum around its mean value.
Spectral slope	Measures the amount of decrease of the spectrum over frequency and is an important aspect of timbre; the slope is directly related to the resonant characteristics of a sound and may be implicated in recognition. Spectral slope is most pronounced when the energy in the lower frequencies is much greater than the energy in the higher frequencies.
Spectral spread	The standard deviation around the spectral centroid.
Spectral variation	Similar to spectral flux, this is also a time-varying descriptor representing the amount of changes of the spectrum over time (geometric difference).
RMS energy	Measures the root mean square of energy of the signal in each time frame, an indication of the intensity level of the music.
Zero-crossing rate	Shows the number of times the signal crosses the zero axis in a given time window and tends to be small for periodic sounds and large for noisy ones.
Roughness	Uses Sethares's (1998) roughness model in which the frequency of each component is multiplied by its amplitude, and its dissonance with all other peaks is computed based on Sethares's (1993) dissonance curve. The average of the dissonances of all pairs of peaks is taken as a measure of roughness, indicating dissonance perception.
Mirmode	Uses Gómez's (2006) key profile matrix, which utilizes a combination of machine learning with the key estimation algorithm proposed by Krumhansl and Kessler (1982) and returns a value between –1 and +1: the closer it is to –1, the more minor it is, and the closer it is to +1, the more major.
Onset detection	Calculates an onset likelihood function for each spectral frame, and picks peaks in a smoothed version of this function to estimate the onset times of notes. It provides information about the general textural density of the music, measuring both the number of instruments sounding together at the same time, as well as the amount of rhythmic activity that is happening.

surface acoustic features will relate more directly to perceived arousal than valence, and that learned musical cues influence listeners' perceived valence more than their perceived arousal.

Listeners' responses were collected at a rate of 2 Hz, and the measured values of the acoustic and musical features were averaged over successive 0.5-s windows. Responses were all standardized before transformation. The process of transforming these variables in a way that is appropriate for time series analysis followed the procedure outlined by Dean and Bailes (2010). Acoustic and musical features are likely to influence listeners' perceptions but not the other way round and so perceptual responses were taken as the endogenous variables and acoustic and musical features as exogenous ones. Listener responses are likely nonstationary in each moment—the process does not remain in statistical equilibrium and the probabilistic processes change over time (Box et al., 2016). The perceptual response is thus likely to be influenced by preceding moments. Therefore, the endogenous variables first had to be made stationary.

Responses were averaged over each listener group, giving a mean set of observations for each of the CHM, WM, and NM groups for emotional intensity, valence, and arousal. Autocorrelation and partial autocorrelation functions for each of these endogenous series were first determined in order to set the lag amount for stationarizing. The stationarized series were then tested with the Augmented Dickey-Fuller Generalized Least-Squares test and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test for stationarity. All endogenous series showed stationarity after one differencing step and a lag of 4 s. The exogenous series were then transformed in the same way as the endogenous series. Granger causality tests determined plausible exogenous predictors.

With these, ARIMAX (autoregressive integrated moving average procedure with an exogenous variable) models were built, and the best model was selected based on minimizing the Akaike Information Criterion (AIC). After the VAR model was built, structural restrictions were imposed by specifying matrices that capture the contemporaneous relationships among variables. One matrix captures the contemporaneous relationships between the endogenous variables. The diagonals are set to 1, meaning the variables are uncorrelated and off-diagonal elements are to be estimated. The second matrix captures the contemporaneous effects of the structural shocks on the endogenous variables.

Impulse response functions were constructed with these SVARs. Each of the acoustic features was modeled as input and the listeners' emotional intensity, valence, and arousal responses were taken as reactions to shocks in the system. With this approach, we can see what the lag response in the endogenous variable is with respect to an input from the exogenous variable. Although listeners might be familiar with particular musical tropes and styles and anticipate certain responses before a particular stimulus is presented, it is difficult to untangle the complex contributions from anticipatory responses and reactionary responses. In these analyses, we will therefore only focus on responses that occur after the presentation of the sound, and so a restriction that each impulse response starts from zero was applied. As we are more interested in the immediate reaction engendered by particular acoustic and musical features on listeners' responses, rather than the smaller background contributions they might have, the additional restriction that the impulse response converges to zero after a period of time was also imposed. To restrict the model so that each impulse starts at zero, the appropriate elements in the matrix are set to 0. Setting all immediate responses to shocks in a variable to 0 ensures that each impulse response function starts from 0. In addition, a dynamically stable model would have the impulse response functions converge to zero over time.

Valence responses and arousal responses were also modeled against each other to examine whether there were any systematic relationships between them. If valence is related to arousal responses or vice versa, it could imply that listeners' perceptions of affect on the two dimensions are not independent, or that the 2D interface for rating perceived affect might create unwanted artifacts in the response data. However, this analysis did not reveal any significant influences between the two measures. None of the 95% confidence intervals of the impulse response functions for these analyses breached the *x*-axis.

Results

GROUP DIFFERENCES WITH FUNCTIONAL *F*-TESTS

The means of the responses of each listener group on emotional intensity, valence, and arousal were plotted from 20 s to 930 s (Figures 5 to 7, respectively). A few observations can be made from these mean curves. First, it appears that the shapes of the mean emotional intensity curves are very similar to those of arousal. Listeners likely interpreted the strength of emotional response in a similar way to that of the arousal response as suggested by Schubert (2001).

The mean curves for valence showed more divergence between the listener groups than those of emotional intensity and arousal. This was especially prominent in the middle of the piece, from around 300 s to 500 s. Compared to emotional intensity or arousal, valence responses of CHM were more extreme with greater swings above and below those of the other two listener groups. Even though the three groups of listeners diverged significantly in their perception of affective intentions for certain temporal regions, their divergence was mostly in terms of quantity rather than quality, except for a region in the middle of the piece (355–409 s, Figure 7) with high-arousal/negative-valence. For the other divergences, responses for the three groups of listeners remained on the same side of the space (upper or lower for arousal, left or right for valence) and differed only in magnitude.

To examine differences between listener groups, a functional regression model is estimated, after which an *F* statistic is computed for the regression to explore the differences among the three listener groups for emotional intensity, valence, and arousal. It is difficult to derive a null distribution for functional statistics, so a permutation test method was used to obtain the significance value. This is done by randomly pairing the response and covariates. Theoretically, if there is no systematic relationship in the variables, there should be no difference when we permute and pair response and covariates. Functional regression equations were constructed, and *F* ratios obtained by comparing the predictive relationship of the model with the random one. The *p* values for the test are derived from the proportion (of several hundreds of random permutations) of permutation *F* values that are larger than the *F* ratio for the observed data. Pointwise critical *F* values are obtained using the equation:

$$F(t) = \frac{Var[\hat{y}(t)]}{\frac{1}{n}\sum \left(y_i(t) - \hat{y}_i(t)\right)^2}$$

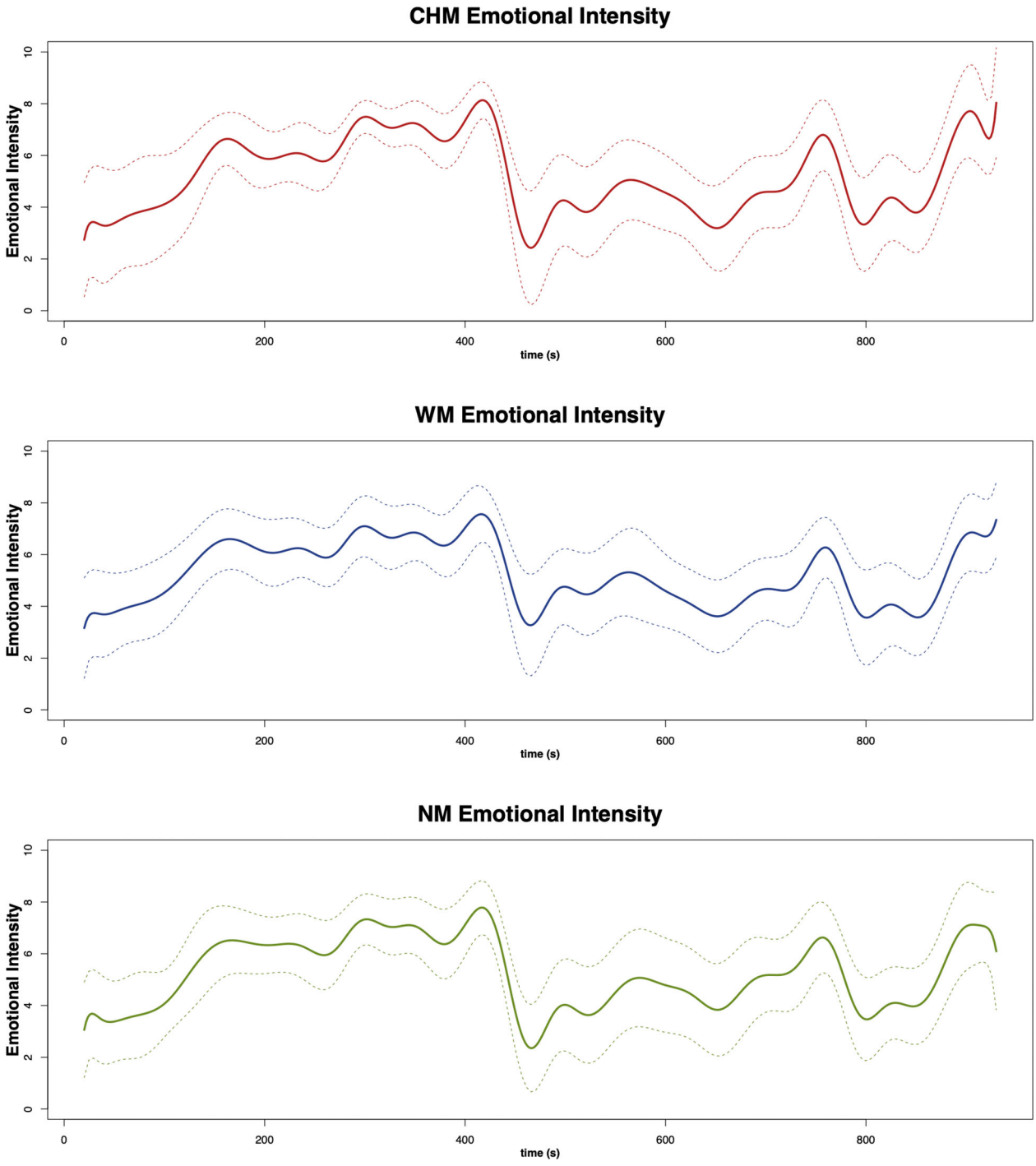

FIGURE 5. Mean emotional intensity for each listener group with dashed lines indicating the standard deviation.

where $\hat{y}$ is the vector of predicted responses. A maximal F ratio is also calculated by randomly shuffling and assigning the curves to the three groups and finding the maximum F ratio with these randomized curves.

The functional F statistic is used to indicate the temporal regions in which the listener groups diverged significantly in their responses. This directly tests the first three research hypotheses: 1) listeners with

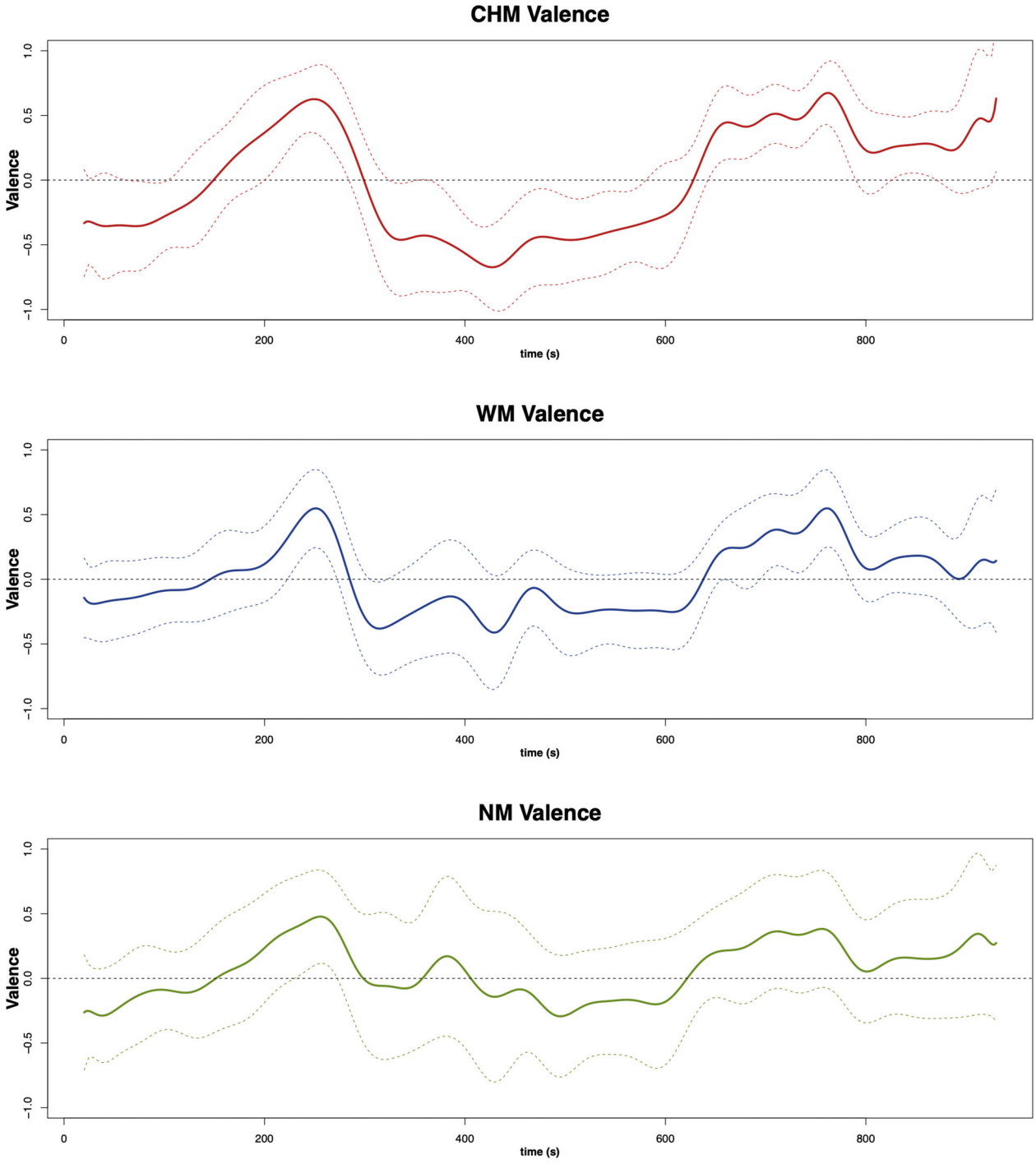

FIGURE 6. Mean valence for each listener group. See Figure 5 caption.

different musical backgrounds respond differently in their perceived affect of a piece of music, 2) valence responses will diverge more than the arousal responses between the three listener groups, and 3) a listener's perceived emotional intensity reflects more of their perception of the distance away from the origin in the two-dimensional valence-arousal space than its relation to the single dimension of arousal.

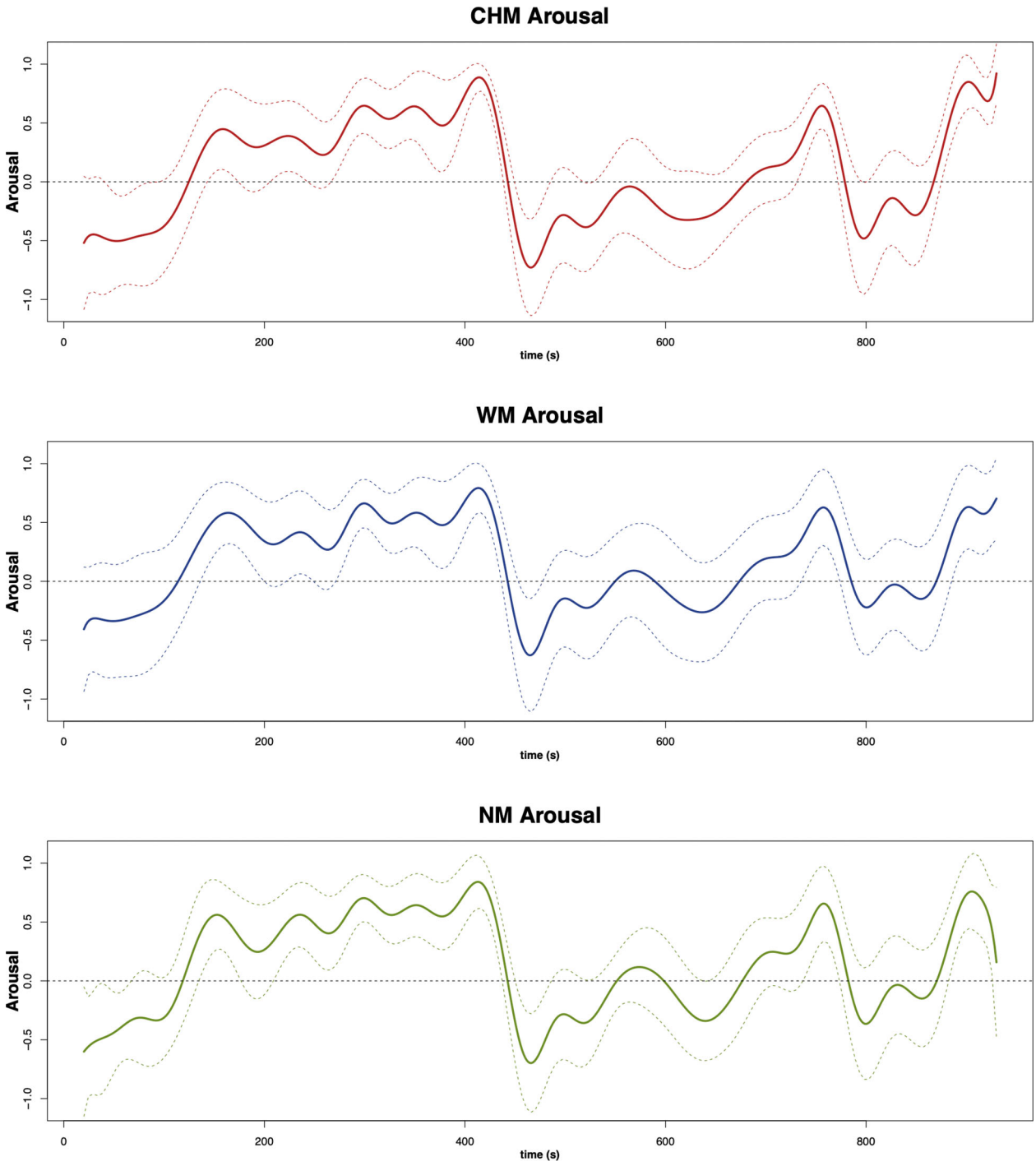

FIGURE 7. Mean arousal for each listener group. See Figure 5 caption.

Figures 8–10 show the functional *F* statistic for the three scales. The pointwise critical *F* value for a significance level of .05, seen as a blue dotted curve, was calculated for participants' responses over the 910 s included in the analysis, and the maximal critical *F* value is seen as a straight blue dashed line over the entire function. The red line indicates the functional *F* statistic across time. The pointwise critical *F* value will

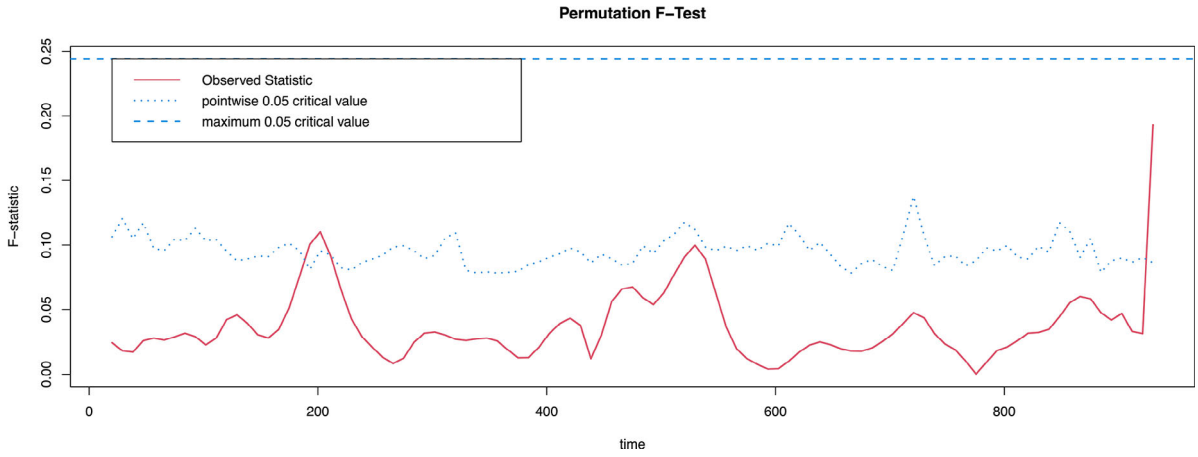

FIGURE 8. Functional F-test for emotional intensity. The blue dotted curve shows the pointwise critical F-value of .05. The straight blue dashed line shows the maximal critical F-value. The red line indicates the functional F statistic across time.

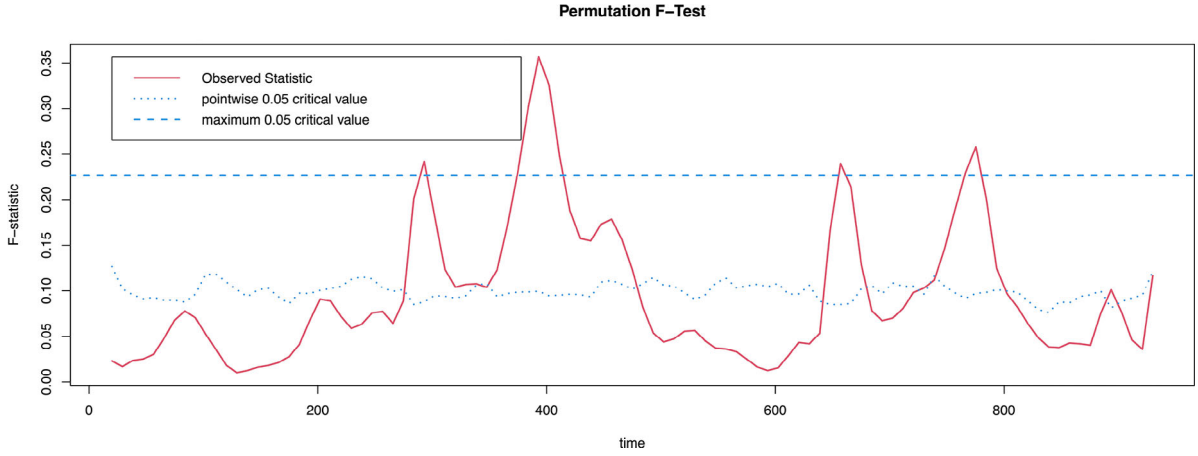

FIGURE 9. Functional F-test for valence. See Figure 8 caption.

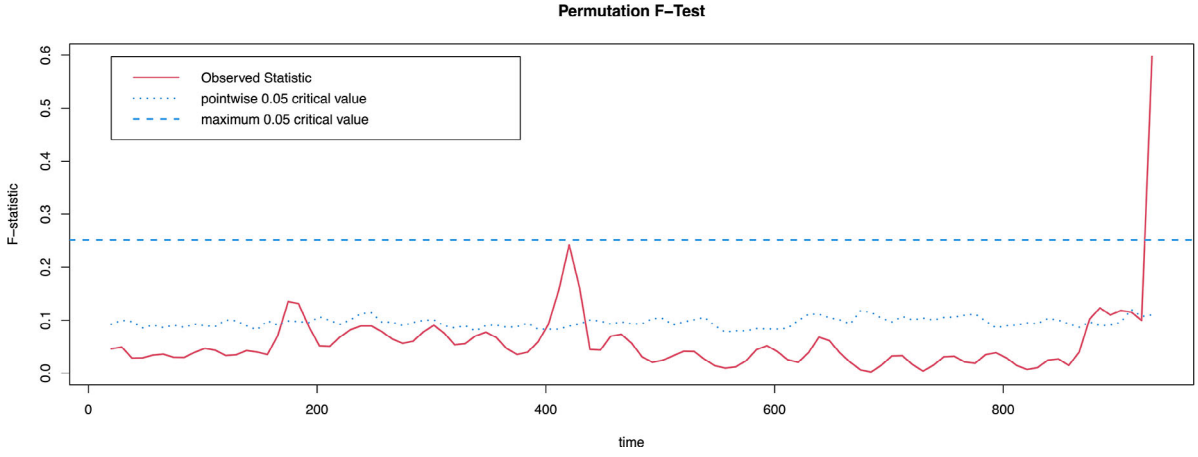

FIGURE 10. Functional F-test for arousal. See Figure 8 caption.

TABLE 2. *Spearman Correlations of Emotional Intensity with Arousal, Emotional Intensity with Valence, and Arousal with Valence for Each Listener Group*

CHM	WM	NM
Emotional Intensity with Arousal		
$\rho(28) = .79, p < .01$	$\rho(28) = .42, p < .01$	$\rho(28) = .52, p < .01$
Emotional Intensity with distance from the origin		
$\rho(28) = .20, p = .29$	$\rho(28) = .28, p = .13$	$\rho(28) = .16, p = .40$
Emotional Intensity with Valence		
$\rho(28) = .32, p = .15$	$\rho(28) = -.08, p = .66$	$\rho(28) = .18, p = .35$
Arousal with Valence		
$\rho(28) = .23, p = .22$	$\rho(28) = -.29, p = .11$	$\rho(28) = .22, p = .24$

be used as a significance criterion here because listeners' responses vary according to the context of the acoustic and musical parameters, and because this study investigates time-varying changes that are influenced by varying contextual cues rather than a single unitary response for the entire piece of music. So temporal regions in which the F curve exceeds the pointwise critical curve are taken as significantly different among listener groups. These functional F-tests confirm the pattern observed in the mean curves—there are less regions that are significantly different between the three listener groups in emotional intensity and arousal, and many more regions of difference in valence. Although the shapes of the mean emotional intensity and arousal curves are very similar, they do not share the exact same regions in which listener responses diverged significantly between groups.

Differences in the mean valence curves for the three listener groups confirms the second hypothesis, that valence responses between the listener groups diverge more than arousal responses, at least in certain parts of the piece. They also provide partial support for the first hypothesis, that listeners with different musical backgrounds respond differently in their perceived affective responses. These differences occur mostly in valence responses and not as much in arousal or emotional intensity responses.

A visual examination revealed close similarities between the shapes of the mean emotional intensity and arousal curves and no clearly observable correspondences with valence curves. The regions where arousal was close to zero corresponded to regions where the emotional intensity was around 5, the midpoint of the emotional intensity scale. The peaks, troughs, and directions of the slopes have a close visual match for both sets of curves. A significant functional canonical correlation (Ramsay et al., 2009) is found between the emotional intensity curves and the arousal curves, $\rho(88) = .62$, $p < .001$. No significant correlations are found between the emotional intensity and valence curves, $\rho(88) = .19$, $p = .07$, or between the arousal and valence curves, $\rho(88) = .07$, $p = .50$. Although the correlations between emotional intensity and both the distance from the origin and valence are marginally significant, those measures explain only about 4% of the variance in emotional intensity in both cases. The correlations for each listener group can be seen in Table 2. Correlations between emotional intensity and absolute valence values were also computed to test whether extreme positive and negative values of valence were both associated with high emotional intensity. The lack of a significant correlation suggests that rather than emotional intensity reflecting the distance from the origin, or the intensity of affective intentions in general, emotional intensity relates more directly to the level of arousal. A low arousal, rated as –1 on the arousal scale is considered as having a low emotional intensity of 1, and a high arousal, rated as +1 is considered as having a high emotional intensity of 9. This supports the third hypothesis—listeners' perceived emotional intensity is not related to the distance away from the origin in the two-dimensional valence-arousal model, but rather is closely related to the single dimension of perceived arousal as previously proposed by Schubert (2001).

The standard deviations (Figures S1–S3 in the Supplementary Materials accompanying this paper at online.ucpress.edu/mp) showed an increase at the end

of the piece, which could be related to the afterglow effect. On the other hand, it might also be a unique artifact of this piece of music. Towards the end, there is a drop in dynamic level, a decrease in the number of instruments, and a reduction in the overall note density, before a sudden dramatic build up to an intense, loud climactic conclusion.

AFFECTIVE RESPONSES WITH ACOUSTIC AND MUSICAL FEATURES

Mean curves for each listener group are correlated with the curves of all of the acoustic and musical features over the entire piece of music. Time series observations are often autocorrelated, and in order to prevent an overestimation of the correlation size, the degrees of freedom are adjusted using this formula (Bartlett, 1946; Box et al., 2016):

$$EDF \approx \frac{n}{1 + 2\sum_{k-1}^{P} \rho_k}$$

where ρ_k is the autocorrelation at lag k, p is the maximum lag considered for the calculation, and n is the total number of observations.

As the shapes of the emotional intensity curves are very similar to those of arousal, Figure 11 shows the mean arousal curves of the three listener groups together with only the features that are significantly correlated with them: RMS energy, $r(237.6) = .30$, $p < .001$, roughness, $r(221.7) = .21$, $p < .005$, note onset density, $r(242.9) = .19$, $p < .005$, spectral slope, $r(248.8) = -.21$, $p < .001$, and zero-crossing, $r(316.5) = .14$, $p < .01$. Note that although the magnitudes of the changes do not correspond exactly, the peaks and troughs of these curves match very closely, especially for RMS energy, roughness, and note density. Spectral slope varies negatively with emotional intensity and arousal responses, the troughs of the spectral slope curve matching the peaks of the emotional intensity and arousal curves, and vice versa.

To quantify the effects of these acoustic and musical features more clearly, impulse response functions were modeled to study their contributions to listeners' valence and arousal responses. Figure 12 shows an example of the impulse response plot for note density on arousal responses for CHM, WM, and NM. The full set of impulse response functions can be found in Supplementary Figures S4 to S12. The *y*-axis reflects the amount of unit change in response to each acoustic or musical feature, and the *x*-axis shows the number of lags taken for the reaction to reach a certain level. Each lag is 0.5 s. The red dotted lines show the 95% confidence interval for each impulse response. Only impulse responses with the 95% confidence intervals breaching the *x*-axis were considered to have an influence on listeners' responses.

Table 3 shows the acoustic and musical features that influence each group of listeners in their perceived emotional intensity, arousal, and valence responses across the entire piece of music, together with the corresponding coefficients obtained from the SVAR models. The SVAR coefficients reflect steady-state relationships and are crucial for understanding the baseline interactions among variables. Impulse response functions illustrate how shocks disturb the system's equilibrium and how the disturbances spread over time, providing insights into the dynamic responses and transmission mechanisms in the system. All three groups of listeners appear to utilize similar features in their perception of emotional intensity, although the two musician groups show greater magnitude than the nonmusicians in their impulse response curves.

Zero-crossing rate indicates noisiness, a higher level of which also indicates a higher perceived emotional intensity and arousal. Roughness, as a measure of sensory dissonance, is seen to be more directly related to the perception of emotional intensity and arousal than to that of valence. A higher RMS energy relates to higher emotional intensity and arousal as well. Spectral slope also influences listeners' perceptions of emotional intensity and arousal. These features collectively point towards a more even distribution of energy across partials—richer, noisier, and louder sounds with greater harmonic dissonance and a denser texture seem to induce higher emotional intensity ratings. The three listener groups also appear to use similar features for emotional intensity and arousal ratings (spectral slope, RMS energy, zero-crossing rate, roughness, and onset density). WM and NM groups utilize these features for both scales.

In addition to all these features, CHM's arousal ratings also feature spectral flatness, spectral flux, and spectral kurtosis. In an orchestral setting, a dense texture with instruments spanning many different registers and complex chordal structures, spectral flatness could be higher, as compared to a thinner texture and a less complex harmonic structure. Changes in instruments and instrumental combinations, or changes in the playing techniques of the instruments could create a high spectral flux, compared to a slow passage with less changes and contrasts in sound. A higher spectral kurtosis might indicate a section of music where the pitches of the instruments are clustered more closely together, or it might indicate a thinner texture in which only very few instruments are playing a very clear melodic line.

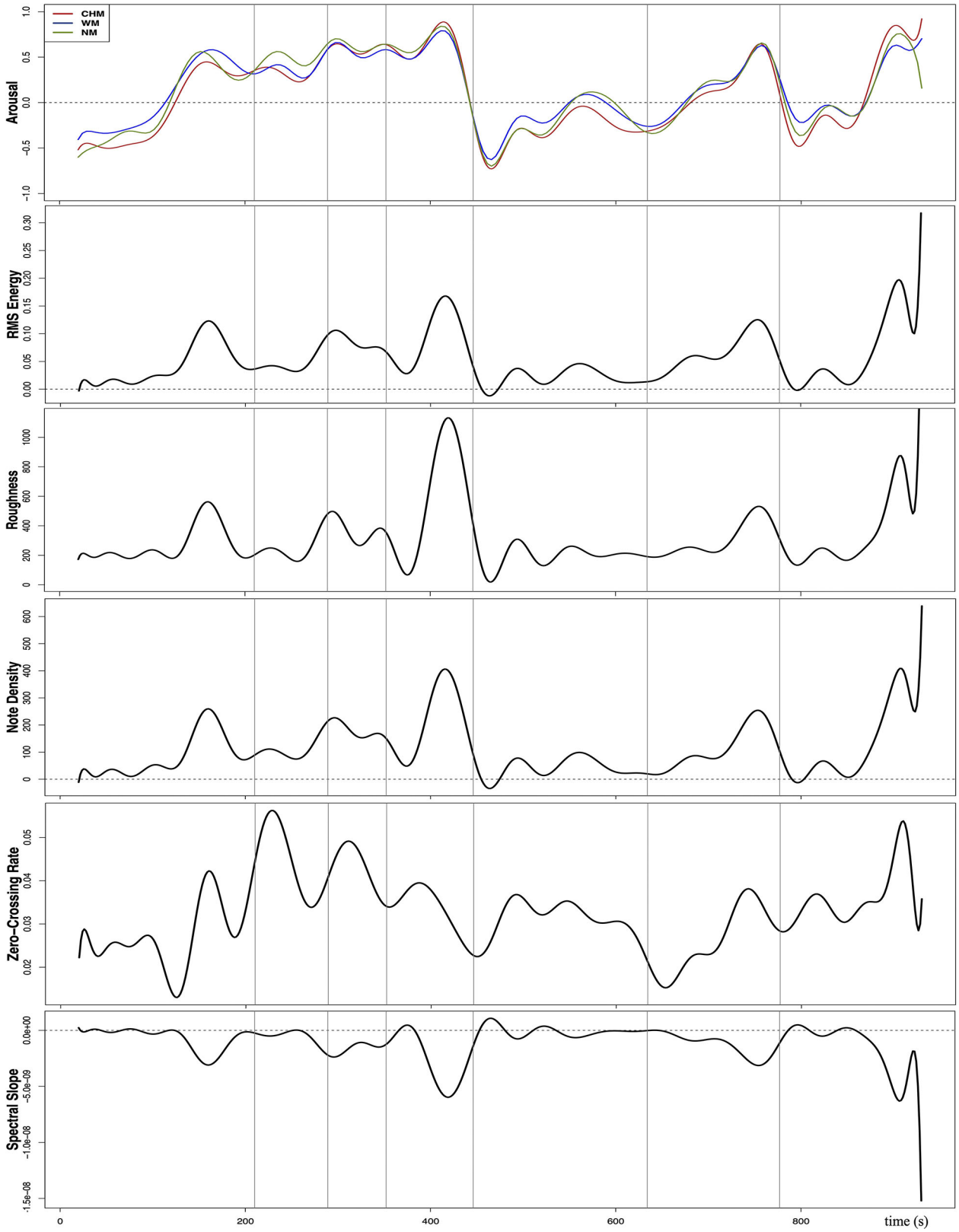

FIGURE 11. Mean arousal curves with those for RMS energy, roughness, note density, spectral slope, and zero-crossing rate.

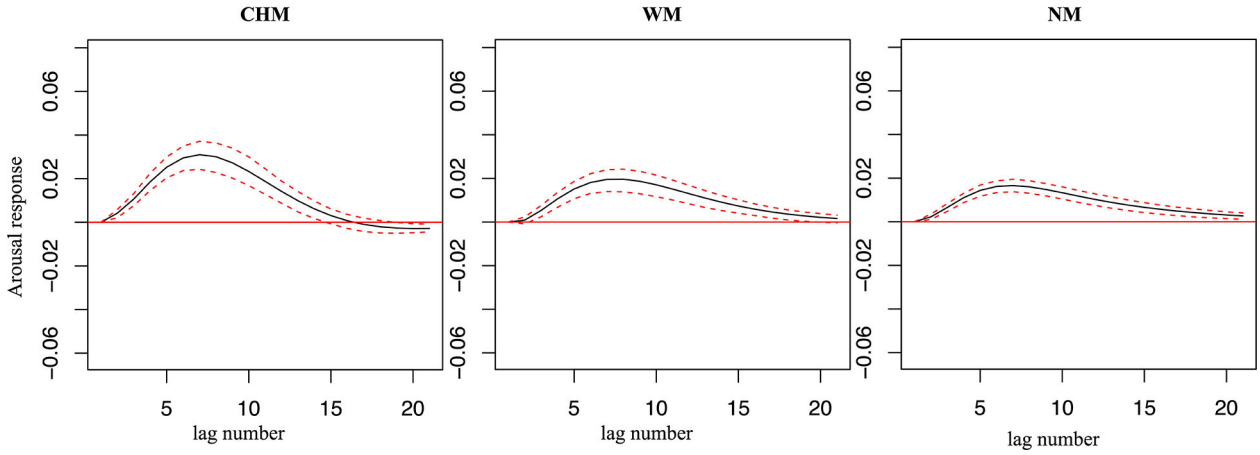

FIGURE 12. Impulse-response functions of note density with arousal responses of Chinese musicians, Western musicians, and nonmusicians.

TABLE 3. *Acoustic and Musical Features Influencing Listeners' Perceptions of Emotional Intensity, Arousal, and Valence with Standardized Coefficients of SVARs in Parentheses*

CHM	WM	NM
	Emotional intensity	
Spectral Slope (–1.2568)	Spectral Slope (–1.3896808)	Spectral Slope (–0.9935)
RMS Energy (1.3897)	RMS Energy (1.2567911)	RMS Energy (1.3897)
Zero-Crossing Rate (0.7383)	Zero-Crossing Rate (0.5888855)	Zero-Crossing Rate (0.8199)
Roughness (0.8199)	Roughness (0.9934935)	Roughness (1.1228)
Density (1.1228)	Density (0.8198525)	Density (1.2568)
	Arousal	
Spectral Slope (–0.73031)	Spectral Slope (–1.50903819)	Spectral Slope (–0.56339)
RMS Energy (1.50904)	RMS Energy (1.43183864)	RMS Energy (1.50904)
Zero-Crossing Rate (0.95794)	Zero-Crossing Rate (0.75424147)	Zero-Crossing Rate (0.71462)
Roughness (1.01979)	Roughness (0.95794399)	Roughness (0.95794)
Density (1.43184)	Density (1.01978709)	Density (1.43184)
Spectral Flatness (0.71462)		
Spectral Flux (0.75424)		
Spectral Kurtosis (-0.56339)		
	Valence	
Spectral Centroid (–1.19072)	Spectral Slope (1.49397307)	Zero-Crossing Rate (1.49397)
Zero-Crossing Rate (0.77032)	RMS Energy (–2.00591387)	Density (2.00591)

Note. Only features whose 95% C.I. of impulse response functions breach 0 are listed. A positive sign indicates that when values of a feature increase, listener's ratings on the particular affective scale rise whereas a negative sign indicates an increase in the value of a feature leading to a decrease in ratings.

A comparison of the mean valence curves with the acoustic and musical feature curves did not yield any clear correspondences—none of the correlations were significant. It appears to be more difficult to attribute any clear and strong influence of any of the features with listeners' perceived valence responses. Impulse response functions corroborate this, showing not only that valence responses appear to have the lowest number of acoustic and musical features that influence its perception, but also the very small contributions of these features. CHM and WM do not share any features in the perception of valence, but zero-crossing rate appears to influence valence for both CHM and NM. For CHM a slightly darker but noisier sound points towards a more positive valence. More tonal, darker, and softer sounds indicate a more positive valence for

WM, and noisier sounds with denser textures are perceived as more positively valenced for NM.

Clear indication of the association of these acoustic features with perceived emotional intensity and arousal responses, and a lack of strong relationships between any features with perceived valence, provides support for the fourth hypothesis, that surface acoustic features relate more directly to perceived arousal than to perceived valence.

MUSICAL AND NARRATIVE ANALYSIS

An understanding of the music, both in terms of its affective and narrative content, includes both acoustic features and the musical context. Although acoustic features show certain relationships with listener responses, they do not account for the totality of a listener's reactions to the music. It is therefore worthwhile to briefly explore certain aspects of the musical composition and discuss how these aspects might contribute to listeners' understanding of the affective and narrative content.

The composer clearly delineated the piece into seven sections and explicitly detailed the narrative content for each section. The piece opens with a quiet and solemn section, "Hymn from the Ancients." A grand and majestic section depicting an ancient and prosperous city follows, aptly named "Citadel on the Hill." The next section "Bustle of the Souq" represents a joyous and busy Arab marketplace, contrasting with the subsequent highly tensed section, "Flames of Resentment." A subsequent intense section is titled "Crossfire of the Armies." The intensity decreases and a desolate "Tears of the Rubble" follows. The piece concludes hopefully, the final section being labeled "Path to the Future."

In their introduction to their book on Arabic music, Farraj and Shumays (2019) remind readers that Arabic music encompasses a wide range of musical traditions and genres. The composer of this piece of music has also made a detailed study of the modes and meters used and has detailed his usage very explicitly, both within the score and in his explanatory notes accompanying the score. A very brief explanation about *maqāmāt* and *iqā'āt* will follow here in order to explain the musical analyses and descriptions presented below.

The Arabic *maqām* is a "system of scales, habitual melodic phrases, modulation possibilities, ornamentation norms, and aesthetic conventions" (Farraj & Shumays, 2019, p. 4). A *jins* (plural *ajnās*) is a scale fragment of a *maqām* (plural *maqāmāt*). They are the basic melodic unit, and the *maqām* can be understood as a pathway between many *ajnās* (Farraj, 2018). Each *jins* conveys a distinct mood or character and interestingly, Farrag and Shumays (2019) state that the "mood of each *jins* is subjective, and no study in ethnomusicology has found any evidence that it is perceived in the same way by all (or even majority of) listeners" (p. 193). So although the *jins* possess affective content, what this affective content is might mean something very different to different listeners. Although Arabic scales are rooted in the Pythagorean tuning system, there are many intervals that cannot be described using simple harmonic ratios. Farraj and Shumays believe that whereas harmonic relationships are based on physical reality, it is in fact the aesthetic and cultural choices of Arabic musicians throughout history that established the framework used in Arabic scales instead of it being driven by purely mathematical rationale. For the purposes of this work, however, the composer has specified that a quarter-tone interval is sufficient. While this may not be an accurate reflection of Arabic musical practices, as the Cairo Congress of Arabic Music of 1932 (Congrès international de musique Arabe, 1934) attempted to rationalize it, the intention of the work is not an imitation, but an original composition inspired by Arabic musical elements.

Iqā' (plural *iqā'āt*) is used to describe a rhythmic cycle. *Iqā'āt* are made up of two different basic building blocks, the dum and tak, onomatopoeias derived from the sound produced on membranophones such as the darabuka. Similarly to the *maqāmāt* and *ajnās*, although the composer has incorporated some specific *iqā'āt* in the work, they are not meant to be a true indication of, but rather an inspiration from, Arabic musical practices.

Valence responses from 270 s to 480 s showed significant differences in the *F*-test among the three listener groups (Figure 10). From 270 s to 315 s, CHM rated valence significantly more positively than WM. From 315 s to 480 s, CHM's valence responses continued becoming more negative, whereas WM and NM either stopped decreasing or started to increase in their valence ratings. For the rest of this section, CHM had the lowest valence ratings, NM the highest, and WM fell in the middle. The composer has clearly delineated the region at 289 s, separating it both sonically with a loud punctuated chord and narratively by labeling it "Flames of Resentment," a stark contrast to the previous section "Bustle of the Souq." Musically, the section leading to this region has dense unisons on the plucked and bowed strings over the rhythmic pattern from the *iqā'* Yuruk Semai (Figure S13 in Supplementary Materials). This *iqā'* has a more active, stable, and cyclical pulse, which provides a contrast to "Flames of Resentment" that opens with dotted rhythms and triplets, and loud fanfare-like unisons in winds utilizing various *ajnās* including Sabā Zamzam, Nikriz, Ḥijāzkar, Kurd, ʿAjam,

and Nahāwand. The instrument density then drops off, leaving only a solo *sanxian* (lute) over the cello, double bass, and bass *sheng* (mouth organ) in the low register. These fanfare-like and quieter sections alternate two more times, oscillating between *jins* Ḥijāzkar and the other *ajnās*. Following that, an ascending scalar pattern coupled in thirds and built from the *ajnās* Ḥijāz and Sabā Zamzam creates a noisy and almost dissonant effect.

Although there was a decrease in the valence ratings for all three listener groups, WM showed the greatest decrease. CHM started off this section with a higher valence than WM and NM, but their ratings also very quickly decreased. NM also showed a general decrease in valence ratings, but their rate of change was the least among the three groups. From about 350 s onwards, the valence ratings of both musician groups were in the negative region, whereas NM's valence ratings became positive.

Between 360 s to 420 s, NM rated their perceived valence significantly higher than both CHM and WM, and WM rated theirs significantly higher than CHM. *Maqām* Nahāwand is used here and it might be heard as a minor mode to listeners more familiar with Western music than Arabic music. The instrumentation also creates a dark timbre. However, it appears that modal and timbral content did not contribute as much to communicating a negative valence to NM. Musically at this section, it is likely the increasing dissonances, timbral quality, and the expectation of a negative narrative content led CHM to continue decreasing in their valence ratings, but this did not occur for the other two groups.

The next region where the three groups diverged significantly happens from 650 s to 680 s. CHM perceived the valence to be significantly higher than WM and NM. Mudawwar Halabī is a rhythmic pattern of 12 beats (Figure S13 in Supplementary Materials), and the composer gave the first, final, then middle four beats of this 12-beat pattern to the hand-drum while having the orchestra play a rhythmic pattern in groups of three beats. This bi-metric rhythm in the hand-drum versus the melody introduces an interesting change to the general mood compared to what happened before. The pulse is stable, but not static because of the slight tension created by the overlay of three and four. The instruments used also create a clearer, brighter sound. These features possibly contribute to the perception of a more positive valence in CHM.

The final region where valence responses of the listeners were significantly different occurs from 740 s to 780 s. Here NM rated the perceived valence significantly lower than that of CHM and WM. There is a large swell into dense unisons and stable melodic tendencies, which fall on the fifth and sixth degrees of the scale. This stability is likely sensed by musician listeners much more than by nonmusicians.

An understanding of a tonal schema with its associations to happiness and sadness has been consistently found to influence listeners who have grown up in a culture of Western music. Because of this, we wanted to explore whether there were any influences of tonality on listeners' perception of valence. Although Arabic modes are used in this piece, we hypothesized that the listeners are familiar with the concept of major and minor tonality but not with Arabic modes, and therefore might base their judgments on how close the music sounds to a major or minor tonality, rather than on the Arabic modes and the affect they might elicit. There were two general regions where an increase in "majorness" corresponded with an increase in listeners' valence ratings, one around 200–250 s, and another bigger region from 600 s to the end (Figure 13). There are no clear trends that can be observed where the slopes of the curves move in the same direction; the correspondences present here are simply the side of the horizontal axis the curves fall on. In other words, it does not appear that listeners perceive a more positive valence with a stronger "majorness," simply that a characteristic of being major indicates in general a positive valence. There are, however, many regions in which this correspondence does not apply, and no general distinguishing characteristics can be discerned in the regions where this correspondence occurs compared to regions in which it does not.

Discussion

These in-depth explorations of various aspects of acoustic and musical features, together with an analysis of the music and the composer's narrative intentions provide some insight into how listeners' perceptions of affective intentions evolve over the course of a lengthy piece of music. This study set out to examine the following hypotheses: H1) listeners from a similar non-Western cultural environment but with different musical backgrounds will respond differently in their perceived affect; H2) valence responses between different listener groups will diverge more than arousal responses; H3) a listener's perceived emotional intensity ratings will be related to the single dimension of arousal, rather than the distance away from the origin in the two-dimensional valence-arousal model; H4) surface acoustic factors within the musical sound will relate more directly to perceived arousal than to perceived

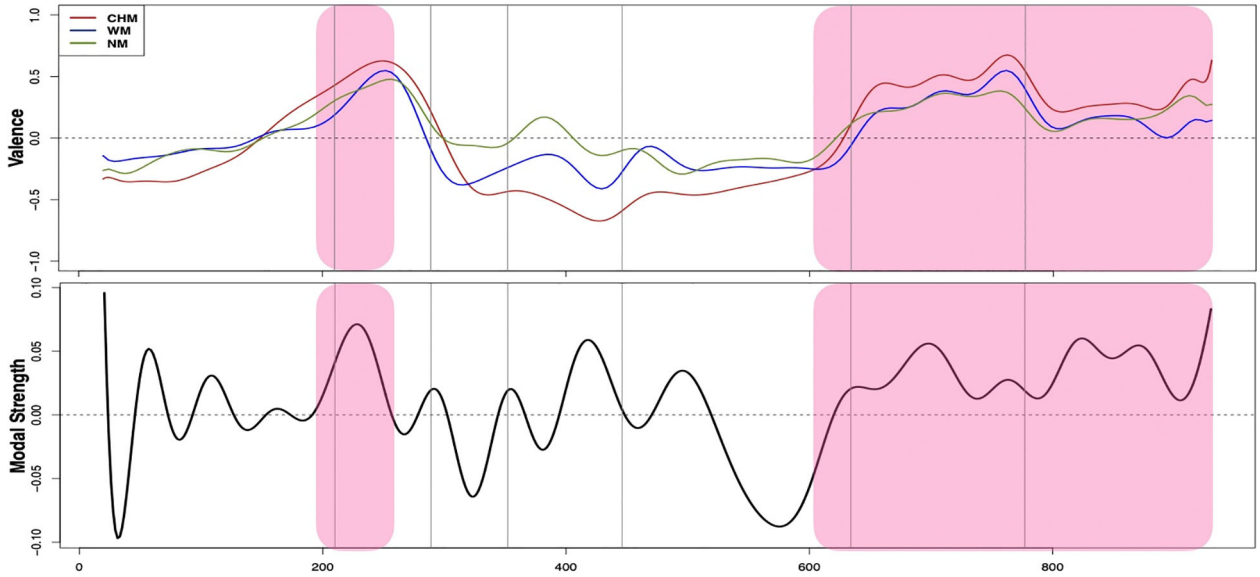

FIGURE 13. Valence responses with modal strength. Highlighted areas show regions where an increase in "majorness" corresponds to more positive-valence ratings and vice versa.

valence; and H5) musical cues that are learned, such as modal structures, metrical relations, and so on, exert a greater influence on listeners' perceived valence ratings than on their arousal ratings. Results support the first two hypotheses, as well as the third, with certain qualifications, as the mean curves for emotional intensity shared many similarities with the mean arousal curves but not with the mean absolute valence curves or the curves plotting distance from the origin of the emotion space. Furthermore, the acoustic features that influence the perception of emotional intensity were found to also similarly influence the perception of arousal. The results also appear to support the fourth hypothesis, with direct influences of particular acoustic features on perceived arousal. Evidence for the fifth hypothesis is somewhat mixed; valence responses appear to be influenced by type and amount of music training more than arousal responses, but the aspects that are learned are less clear. There is no clear correspondence of learned musical cues with perceived affective responses, but there are likely complex interactions with other musical and acoustic features, affective intentions, and the narrative content of the music. Finally, this study was exploratory, attempting to probe into which acoustic and musical features are important, as well as their directions of influence on listeners' continuous responses over the course of a lengthy piece of real music. A musical analysis, together with a close reading of the narrative content intended by the composer, was performed on selected sections, highlighting the ways in which listeners' perception of affective intentions could be related to the musical structure and context. We will discuss the results in light of the hypotheses.

H1. LISTENERS WITH DIFFERENT MUSICAL BACKGROUNDS RESPOND DIFFERENTLY

The mean curves for valence showed more divergence between the listener groups than those of emotional intensity and arousal. In general, valence responses from CHM had a greater magnitude than the other two listener groups compared to emotional intensity or arousal. This could be due to CHM being more confident about their valence responses in general, due to the timbre and performance practice of a Chinese orchestra being the most familiar to this group of listeners.

The composer clearly delineated the work into seven sections, each having a distinct programmatic content and narrative idea. These narratives relate closely to specific affective intentions. The piece opens with a slow, somber section titled "Hymn of the Ancients" with low-arousal and negative-to-neutral valence. "Citadel on the Hill" follows, a grand and majestic depiction of the height of political stability and prosperity, carrying with it a high arousal and relatively neutral valence. The third section titled "Bustle of the Souq" portrays a busy marketplace with plenty of activity going on and conveys a happy, joyous atmosphere, which can easily be related to a high-arousal/positive-valence affective intention. The fourth and fifth sections, "Flames of Resentment" and "Crossfire of the Armies," contrast the

previous high-arousal/positive-valence section with high arousal and negative valence. The penultimate section is titled "Tears of the Rubble," a low-arousal negatively valenced section, and the final section "Path to the Future" generally has a low-arousal/positive-valence affect before making a dramatic close with high arousal and positive valence. Each of these sections relate closely to specific affective intentions. Although this study did not set out to investigate whether participants' perceived affective intentions accurately corresponded to the affective intentions the composer intended to express in the music, it does appear that this communication was quite successful when listeners' responses were compared to the intended affect—the valence and arousal curves (Figures 3 and 4) corresponded largely with the composer's intentions.

Listeners rated the emotional intensity on their first presentation of the piece of music. With no information and expectation of what the music would be, all three groups might have been equally uncertain as to what to expect. Their responses likely reflect their surface interpretation of the acoustic and musical features in the music, and all three groups utilized a similar set of features to inform them of the perceived emotional intensity of the music. Features related to timbre, loudness, and rhythm—spectral slope, RMS energy, zero-crossing rate, roughness, and note onset density—can be seen to relate directly to listeners' perceived emotional intensity responses, and so these could be surface features that are able to communicate particular types of information quickly and effectively. These features do not depend as much on prior musical experience or expertise in a particular style of music.

In the second presentation of the music, all listeners had to rate their perceived valence and arousal. By this second presentation, listeners may already have formed expectations, which might have influenced them to focus on particular aspects of the musical sound and ignore others. The aspects of the sound a listener focuses on may be related to their musical background. The Brunswikian lens model adapted by Juslin (2000) for musical communication suggests that multiple probabilistic cues are available. Thompson and Balkwill (2010) propose common and culture-specific cues listeners have when listening to music from various cultures. As can be seen from the impulse response functions, CHM have a number of timbral cues that are related to their arousal responses: spectral flatness, spectral flux, and spectral kurtosis, in addition to those that were also found to relate to emotional intensity responses. The arousal responses of WM and NM, on the other hand, were influenced by the same set of features as their emotional intensity responses. CHM, who are more experienced with the instruments and compositional techniques used in Chinese orchestral music, might have had an idea of which features figure more prominently in the communication of particular intentions, and therefore would have more information available for their judgments. This additional information might work in tandem with the expectations they had formed, thereby giving them a clearer idea of the affective intentions in the music.

Familiarity with a musical tradition also plays a role in tempering listener responses in other ways. For instance, the common trope of a dramatic ending that is expected by musician listeners may be less expected for the nonmusician listeners. When the sudden drop to a *pianissimo* occurred towards the ending of the piece, the perceived arousal responses of CHM and WM dropped slightly but rose again immediately to end on a high arousal. These two groups of listeners appear to have anticipated a return to a loud and majestic close and therefore kept their arousal responses higher than those of the NM.

H2. VALENCE RESPONSES BETWEEN LISTENER GROUPS DIVERGE MORE THAN AROUSAL RESPONSES

The functional *F*-tests show more regions of difference in valence between listener groups than in emotional intensity or arousal, indicating a greater cultural specificity for perceived valence. The amount of divergence was not constant throughout the music but varied in regions with different affective intentions. The regions with low-arousal/negative-valence were the most agreed upon in the quality of perceived affective intention for all three listener groups (0–109 s and 446–534 s), suggesting that this affective intention can be perceived accurately by all listeners regardless of musical background. Differences were greatest for the region with high-arousal/negative-valence (355–409 s), differing in both magnitude and quality of affect. CHM's perceived affective intentions were accurate, and NM's responses diverged the most from the composer's intended affect for the section. This compares similarly to previous work examining the differences in perceived affective intentions for more limited musical notes, measures, and phrases performed with different affective intentions (Heng & McAdams, 2024). CHM were the most accurate in both studies with regard to the high-arousal/negative-valence intention.

In all the regions where listener responses diverged significantly for arousal ratings, the divergence of all three listener groups were in terms of magnitude but not in terms of quality. This means that even though

they were significantly different, the responses still fell on the same side of the arousal space (high or low in arousal). This is different in the case of valence ratings. In the region between 355–409 s, the three groups of listeners not only differed significantly in the magnitude of their valence ratings, NM also rated valence positively, whereas the valence ratings for CHM and WM were negative. This similarity in arousal responses but divergence in valence responses between listener groups suggests that valence responses are likely learned from the music training. Perceived arousal and emotional intensity responses, on the other hand, are likely based more on universal, culturally independent response mechanisms and therefore show a greater agreement among the listener groups. This is consistent with Egermann and colleagues' (2015) findings that arousal responses are more universal, and valence responses are more culturally learned.

Emotional intensity and arousal responses were also elicited by a similar group of acoustic and musical features in all three listener groups. The features of spectral slope, RMS energy, zero-crossing rate, roughness, and note onset density appear to communicate particular straightforward and simple types of information quickly and effectively and did not depend as much on prior musical experience or expertise with a particular style of music. They might also be less dependent on any narratives formed by the listeners.

H3. HIGHLY SIMILAR EMOTIONAL INTENSITY AND AROUSAL RESPONSES

A listener's perceived emotional intensity does not appear to be related to the distance away from the origin (a "neutral" emotion) in the two-dimensional valence-arousal space, but rather, is more related to the single dimension of arousal. A high degree of similarity can be observed between the shapes of the mean emotional intensity and arousal curves, but not between the shapes of the mean emotional intensity and valence curves (Figures 5, 6, 7). There is also considerable overlap in the features influencing listeners' perception of emotional intensity and arousal and much less overlap in the features that influence emotional intensity and those that influence valence, again pointing towards the similarity in the dimensions of perceived emotional intensity and arousal. Listeners appear to think about emotional intensity in terms of how much arousal the music is expressing; very intense positive or negative affect, however, does not play a part in contributing to emotionally intense responses.

It should be noted that the mixed effects functional data regression for emotional intensity found random effects intercepts with a standard deviation of 0.67 and residuals of standard deviation of 3.47. It could be that there are more individual differences in participants' conceptualizations of what emotional intensity means—some take it to mean that emotional intensity is closely related to arousal, and others take it to mean the distance away from the origin so more extreme emotions. It is also possible that because participants are presented with the emotional intensity rating first, they may not know what to expect. Upon the second presentation in which they were to rate valence and arousal, they have already formed their expectations of the music, and so the ratings converge more between participants.

H4. SURFACE ACOUSTIC FEATURES RELATE MORE DIRECTLY TO PERCEIVED AROUSAL THAN VALENCE

The impulse response functions demonstrate several acoustic and musical features that relate directly to emotional intensity and arousal responses. They also tend to show a greater influence on these responses than do the features that relate to valence responses. Dean and Bailes (2010) showed that sound intensity influences listeners' perception of change in music. Sound intensity is also directly represented by RMS energy, which can be seen to influence perceived emotional intensity and arousal in this study. As opposed to the more direct influence on perceived arousal from surface acoustic cues, cues that influence valence responses function in more complex ways. Even though listeners are relatively successful in their perceived valence responses when these are compared with the composer's intentions, there were very few surface acoustic cues that can be seen to directly influence these responses. The impulse response functions of the cues that relate to valence responses were much smaller in peak value compared to the cues that influence emotional intensity and arousal responses.

However, musical features can also be seen to interact in complex ways with perceived affective intentions. Modal strength demonstrates only a very slight relationship with perceived valence responses. This might occur for a few reasons. The fact that the piece is written with modes inspired from the Arabic musical traditions could pose a problem with the tonality measure in the MIRToolbox. Even though the computation of the frequency of each note with regard to a major or minor mode might be accurate, this might not be a true reflection of how listeners perceive the modality as the Arabic modes are fundamentally very different from the major/

minor modes. In addition, listeners might also have been aware of the very different modal framework used in this piece and therefore were not consciously basing their judgments on how "major" or "minor" it sounded. It is interesting to observe that correspondences of major/minor to positive/negative valence ratings only occur in the "major" portions of the piece with positive valence ratings, but not "minor" with negative valence ratings. Listing the most dominant *ajnās* used within the piece (Figure S14 in Supplementary Materials), it can be observed that they either begin with a minor second or have a minor third in the first three notes. With the most commonly occurring *ajnās* of this piece very clearly not major sounding, it is likely that listeners understood the context and did not use a "minor" sound to indicate a negative valence. Listeners respond to the musical context and modify their responses from previously learned musical knowledge accordingly. The Arabic modes used in the piece have also been labeled as they occur over the piece of music, but it does not appear that there is any consistent relationship between the modes and listener responses—the same *maqāmāt* and *ajnās* occur without any discernible patterns in places with different valence and arousal content. The unfamiliarity of the Arabic musical tradition to these listeners meant that there were no learned implications of affective intentions of the Arabic modes available to them.

In a previous study exploring listeners' global perceived affect on a short excerpt of music performed with different affective intentions on Chinese and Western instruments and presented as single notes, measures, or phrases, the same acoustic feature could map onto affect perception in different directions depending on other features it interacts with (Heng & McAdams, 2024). Extending from that, it is likely that features could also map onto affect perception in different directions depending on the context. This would have to be further explored with functional principal component analyses, functional derivatives, and analysis of sections of the piece separately.

H5. LEARNED CUES EXERT INFLUENCES ON BOTH PERCEIVED VALENCE AND AROUSAL

Acoustic and musical features influencing listeners' perception of affective intentions were related to familiarity with a musical tradition, indicating that learning plays a role in the cues available to listeners. Impulse response functions show a larger number of features influencing CHM responses to arousal than was the case for WM or NM. Being familiar with the timbre of the instruments used and the stylistic features of this genre of music allowed this group of listeners to be more sensitive to the various nuances that contribute to the expression of particular affective intentions and therefore gave them more access to the culture-specific cues for decoding the intentions.

Along with learned cues that provide specific information regarding perceived affective intentions, learning likely also plays a role in narrative formation as "particular mappings between sound pattern and story are highly dependent on enculturation" (Margulis et al., 2019, p. 6). It can be seen here that there are complex interactions between acoustic features and affective, musical, and narrative content.

The three listener groups diverged in their affective responses in certain regions of the music. As mentioned previously, their divergences were more in terms of quantity than quality, but with CHM providing responses more toward the extremes of the scale followed by WM, and lastly NM (e.g., 152–352 s; 446–593 s; 634–777 s). These more extreme responses suggest that CHM are more confident in their responses as they have more information available to them for making judgments on affective intentions. WM, with knowledge in the Western classical music tradition, also had available to them more in-depth knowledge of musical structures, orchestration techniques, and so on, as compared to NM. This knowledge also provided them with a greater amount of information as compared to NM for making affective judgments.

The perception of affective intentions in music listening is therefore a complex process that is influenced by the degree of familiarity listeners have with a musical tradition, the narrative content implicated in the music, and the complex sonic environment created by the composer's work and the musicians' interpretation in performance. The perception of emotional intensity, arousal, and valence taps into different aspects of commonalities and culture-specific understanding of listeners. Several acoustic features for instance, influence the perception of emotional intensity and arousal more directly and similarly across all three listener groups, whereas others influence only CHM and not the other two listener groups. The more extreme and accurate valence responses of CHM when compared with the composer's intentions also point towards the implications of culture-specific understanding.

Conclusion

Music listening involves not only numerous rapid and complex mechanisms in processing sonic

materials; these real-time processes also usually occur and evolve over a certain period of time. This study tracked listeners' responses and their changes over the course of a lengthy piece of music and explores some acoustic and musical features that might be implicated in the perception of affective intentions. We used a Chinese orchestral composition for the stimulus. The piece was not specially composed for the study, which means that it is difficult to control for the changes in acoustic and musical features and the interaction of these features to a listener's perception. On the other hand, it provides a more naturalistic listening experience, replicating as much as possible music listening in real life.

From the results, it can be seen that although there are several acoustic and musical features that appear to influence perceived affective intentions, many others do not show any direct relationships. Valence responses also appear to be less directly related to the acoustic and musical features explored in this study. However, when narrative and musical contexts are taken into consideration, a different pattern emerges. Listeners appear to take into account the affective and narrative content of what is happening in the music and respond in different ways to the accompanying acoustic features. With enculturation in a musical tradition, listeners are likely to be more sensitive to changes in the music that communicate particular affective intentions. However, even without having formal training in a particular musical tradition, there are enough shared cues that provide information for decoding. As espoused by the cue-redundancy model (Thompson & Balkwill, 2010), learning and experience in a musical tradition provide listeners with more available cues for musical communication. Perceived affective intentions in music are therefore influenced by many factors including listeners' experience of a musical tradition, the acoustic features present in the music, musical content such as harmony, tonality, rhythm, etc., and the narrative content of the piece. More importantly, these factors do not implicate the perception of affective intentions independently; complex interactions exist between these features, the musical and narrative context, and prior knowledge and experience that communicate nuanced affective intentions to a listener.

Emotional intensity and arousal responses appear to be elicited from more universal, common cues than valence responses. When listeners are presented with an unfamiliar piece of music for the first time, the narrative content is still in the process of being formed. They might therefore focus on acoustic and musical features that convey information for them to make decisions to the expressed affective intentions more directly. With an additional presentation of the music, there is likely to be a clearer expectation not only of musical aspects of the sound, but the narrative content of the music may also be more clearly formed. This may be even more prominent for music in which the narrative and programmatic content is especially emphasized, as in the case of the music used in this study. Valence especially appears to be a complex dimension. The narrative content likely influences how listeners perceive the valence of the music to a large extent.

LIMITATIONS AND FUTURE DIRECTIONS

This study examined a single piece of music, so although results appear to show that listeners do respond differently based on their musical backgrounds and that certain acoustic and musical features relate to perceived affective intentions, it is difficult to generalize to all music listening. However, musical experience is specific to a given piece and not general to all possible pieces. It is therefore important to complement studies on controlled proto-musical materials with those on real music to determine if the results still hold in more realistic music-listening scenarios. Therefore, a few different pieces of music, in different musical styles and traditions will have to be studied as well to see if these responses to acoustic and musical features can be generalized, although one might reasonably presume each piece of music to have its own specificities in terms of acoustic and musical features it employs. Again, listeners do not hear music in general, but specific pieces of music!

The number of participants may be underpowered for a between-groups study, with 30 participants for each group for a total of 90 participants. With a smaller number of participants than is optimal for between-group studies, according to Brysbaert (2019) clear inferences can be more difficult if the data are variable. We acknowledged this and explored standard deviations of the data to investigate it. Emotional intensity, for instance, had high variability between participants and, therefore, interpretation of patterns observed in the emotional intensity curves were largely withheld. The standard deviations were much smaller for valence and arousal ratings. Future work could attempt to replicate the findings with a much larger sample in a different cultural setting, although getting a large number of participants per group for our sample, as suggested by Brysbaert (2019), would seem unlikely given the constraints on each group type within the population. We

feel nonetheless that our exploration of the acoustic features that are associated with the different responses between groups makes a valuable contribution to the literature.

Although functional data analysis allows for curve estimation from discrete datapoints, as well as derivatives of these functional data, and therefore has the potential to study changes over time, it is more difficult to infer correlations between a set of curves (participants' responses) in response to a single continuous variable (each acoustic feature). Therefore, after detailed visual exploration revealed correspondences between acoustic and musical features and perceived arousal, correlation analyses were conducted for the acoustic and musical features with changes in perceived affective intentions over the course of the music, taking into account the effective degrees of freedom. Time series analysis was used to supplement this, and impulse response functions were used to visualize the influences of these acoustic and musical features. Although this approach sheds light on how listeners' perceptions react to the various acoustic and musical features, it provides a generalization over the entire length of music and is not able to show changes over the course of the music, such as where the importance of a particular feature might change in different parts of the music, as has been explored by Dubnov et al. (2006). Other useful techniques of functional data analysis, such as functional principal component analysis, can also be used to further explore the complex relationships that could underlie the perception of affective intentions in continuous music listening (e.g., Almansa & Delicado, 2009).

This piece of music is intended to communicate a wide range of different affective intentions and narrative content. This means that an examination of how features might influence affect perception by taking the entire piece of music as a whole might not yield any clear results. A related paper will examine separate sections of the music in a more detailed interrupted time-series analysis. We will also examine derivatives of the response curves with respect to separate sections of the music to estimate the contribution of different musical and acoustic features to the affective velocity and acceleration of responses (Vines et al., 2005). These analyses would further illuminate how acoustic and musical features might contribute to particular affective intentions and demonstrate how listeners respond in these different sections, helping to further understand this complex process of music perception.

Music listening is a common human activity, but at the same time is a very complex cognitive behavior. Increasing our understanding of how listeners perceive complex, lengthy, continuous sounds is important in untangling auditory perception and cognition and how acoustic and musical features are used. It also illuminates how learning plays a role in influencing the way listeners process sounds and shapes perception. In addition to understanding how humans process musical sounds, music research can also illuminate the ways in which everyday sounds and language are processed, as well as how aspects of the sounds can convey affective intentions in human communication.

Published online: September 24, 2025

Author Note

We thank composer Wang Chenwei for providing the music used for the experimental stimulus. We also thank Bennett K. Smith for programming the experimental interface.

This research was supported by grants to SMc from the Canadian Social Sciences and Humanities Research Council (895-2018-1023, 435-2021-0224) and the Fonds de recherche du Québec—Société et culture (2017-SE-205667), as well as a Canada Research Chair (950-223484).

LH is now at the Department of Psychology, St. Francis Xavier University. CW is now at the Department of Psychology, Queen's University.

Correspondence concerning this article should be addressed to Lena Heng (lena.heng@mail.mcgill.ca), St. Francis Xavier University, Department of Psychology, StFX, 4130 University Ave, Antigonish, NS, Canada B2G 2W5.

References

Almansa, J., & Delicado, P. (2009). Analysing musical performance through functional data analysis: Rhythmic structure in Schumann's *Träumerei*. *Connection Science*, *21*(2–3), 207–225. https://doi.org/10.1080/09540090902733848

Bartlett, M. S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *8*(1), 27–41. https://doi.org/10.2307/2983611

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2016). *Time series analysis: Forecasting and control* (5th ed.). John Wiley and Sons, Inc.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), 16. https://doi.org/10.5334/joc.72

Cannam, C., Landone, C., & Sandler, M. (2010). Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. *Proceedings of the ACM Multimedia 2010 International Conference* (pp. 1467–1468). ACM Digital Library. https://dl.acm.org/doi/pdf/10.1145/1873951.1874248

Chan, M. C. (2003). Chinese orchestral music is better because of you—An appreciation of Chinese orchestral music (Vol. 1). Joint Publishing H. K. Pte Ltd.

Congrès international de musique Arabe. (1934). *Recueil des travaux du Congrès de Musique Arabe* [Proceedings of the works of the Congress of Arabic Music]. Imprimerie Nationale Boulac (Cairo).

Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, *31*(4), 377–403. https://doi.org/10.1007/bf01404567

Dean, R. T., & Bailes, F. (2010). Time series analysis as a method to examine acoustical influences on real-time perception of music. *Empirical Musicology Review*, *5*(4), 152–175. https://doi.org/10.18061/1811/48550

Dean, R. T., & Bailes, F. (2011). Modelling perception of structure and affect in music: Spectral centroid and Wishart's *Red Bird*. *Empirical Musicology Review*, *6*(2). https://doi.org/10.18061/1811/51217

Dubnov, S., McAdams, S., & Reynolds, R. (2006). Structural and affective aspects of music from statistical audio signal analysis. *Journal of the American Society for Information Science and Technology*, *57*(11), 1526–1536. https://doi.org/10.1002/Asi.20429

Eerola, T., Lartillot, O., & Toiviainen, P. (2009). Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. *Proceedings of the International Conference on Music Information Retrieval* (pp. 621–626). ICMIR.

Egermann, H., Fernando, N., Chuen, L., & McAdams, S. (2015). Music induces universal emotion-related psychophysiological responses: Comparing Canadian listeners to Congolese Pygmies. *Frontiers in Psychology*, *5*, 1341. https://doi.org/10.3389/fpsyg.2014.01341

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Farraj, J. (2018). *Arabic Maqam World*. Maqam World. http://www.maqamworld.com/en/index.php

Farraj, J., & Shumays, S. A. (2019). *Inside Arabic music: Arabic maqam performance and theory in the 20th century*. Oxford University Press.

Gabrielsson, A., & Lindström, E. (2010). The role of structure in the musical expression of emotions. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 368–402). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199230143.003.0014

Gómez, E. (2006). Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing*, *18*(3), 294–304. https://doi.org/10.1287/ijoc.1040.0126

Harrison, P., & Pearce, M. T. (2020). Simultaneous consonance in music perception and composition. *Psychological Review*, *127*(2), 216. https://doi.org/10.1037/rev0000169

Heng, L., & McAdams, S. (2024). The function of timbre in the perception of affective intentions: Effect of enculturation in different musical traditions. *Musicae Scientiae*, *28*(4), 675–702. https://doi.org/10.1177/10298649241237775

Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(6), 1797–1812. https://doi.org/10.1037/0096-1523.26.6.1797

Juslin, P. N. (2013a). From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions. *Physics of Life Reviews*, *10*(3), 235–266. https://doi.org/10.1016/j.plrev.2013.05.008

Juslin, P. N. (2013b). What does music express? Basic emotions and beyond. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00596

Kazazis, S., Depalle, P., & McAdams, S. (2021). *The Timbre Toolbox version R2021a, user's manual*. https://github.com/MPCL-McGill/TimbreToolbox-R2021a

Krumhansl, C. L. (1998). Topic in music: An empirical study of memorability, openness, and emotion in Mozart's String Quintet in C Major and Beethoven's String Quartet in A Minor. *Music Perception*, *16*(1), 119–134. https://doi.org/10.2307/40285781

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, *89*(4), 334. https://doi.org/10.1037/0033-295x.89.4.334

Lartillot, O. (2022). *MIRtoolbox. MATLAB Central File Exchange*. [Computer software]. https://www.mathworks.com/matlabcentral/fileexchange/24583-mirtoolbox

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science and Business Media.

Lütkepohl, H. (2008). Impulse response function. In S. N. Durlauf & L. Blume (Eds.), *The new Palgrave dictionary of economics* (Vol. 11). Palgrave Macmillan. https://hdl.handle.net/1814/17636

Margulis, E. H., Wong, P. C. M., Simchy-Gross, R., & McAuley, J. D. (2019). What the music said: Narrative listening across cultures. *Palgrave Communications*, *5*(1). https://doi.org/10.1057/s41599-019-0363-1

McAdams, S., Vines, B. W., Vieillard, S., Smith, B. K., & Reynolds, R. (2004). Influences of large-scale form on continuous ratings in response to a contemporary piece in a live concert setting. *Music Perception*, *22*(2), 297–350. https://doi.org/10.1525/Mp.2004.22.2.297

Nagel, F., Kopiez, R., Grewe, O., & Altenmüller, E. (2007). EMuJoy: Software for continuous measurement of perceived emotions in music. *Behavior Research Methods*, *39*(2), 283–290. https://doi.org/10.3758/bf03193159

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. *Journal of the Acoustical Society of America*, *130*(5), 2902–2916. https://doi.org/10.1121/1.3642604

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ramsay, J., Hooker, G., & Graves, S. (2009). Introduction to functional data analysis. In *Functional data analysis with R and MATLAB* (pp. 1–19). Springer.

Reisenzein, R. (1994). Pleasure-arousal theory and the intensity of emotions. *Journal of Personality and Social Psychology*, *67*(3), 525–539. https://doi.org/10.1037/0022-3514.67.3.525

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Schubert, E. (2001). Continuous measurement of self-report emotional response to music. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and emotion: Theory and research* (pp. 393–414). Oxford University Press.

Schubert, E. (2010). Continuous self-report methods. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 223–254). Oxford University Press.

Schubert, E. (2013). Reliability issues regarding the beginning, middle and end of continuous emotion ratings to music. *Psychology of Music*, *41*(3), 350–371. https://doi.org/10.1177/0305735611430079

Sethares, W. A. (1993). Local consonance and the relationship between timbre and scale. *Journal of the Acoustical Society of America*, *94*(3), 1218–1228. https://doi.org/10.1121/1.408175

Sethares, W. A. (1998). Consonance-based spectral mappings. *Computer Music Journal*, *22*(1), 56–72. https://doi.org/10.2307/3681045

Sloboda, J. A., & Lehmann, A. C. (2001). Tracking performance correlates of changes in perceived intensity of emotion during different interpretations of a Chopin piano prelude. *Music Perception*, *19*(1), 87–120.

The MathWorks Inc. (2022). *MATLAB. Statistics and machine learning toolbox*. The MathWorks Inc. https://www.mathworks.com/help/stats/index.html

Thompson, W. F., & Balkwill, L. (2010). Cross-cultural similarities and differences. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 755–788). Oxford University Press.

Thompson, W. F., Bullot, N. J., & Margulis, E. H. (2022). The psychological basis of music appreciation: Structure, self, source. *Psychological Review*, *130*(1), 260–284. https://doi.org/10.1037/rev0000364

Vines, B. W., Nuzzo, R. L., & Levitin, D. J. (2005). Analyzing temporal dynamics in music: Differential calculus, physics, and functional data analysis techniques. *Music Perception*, *23*(2), 137–152. https://doi.org/10.1525/mp.2005.23.2.137

Wang, C. (2017). *Aleppo* [Chinese orchestral composition].