SPECTRAL CORRELATES IN EMOTION LABELING OF SUSTAINED MUSICAL INSTRUMENT TONES

Bin Wu[1], **Simon Wun**[1], **Chung Lee**[2], **Andrew Horner**[1]
[1]Department of Computer Science and Engineering,
Hong Kong University of Science and Technology, Hong Kong
[2]The Information Systems Technology and Design Pillar,
Singapore University of Technology and Design, 20 Dover Drive, Singapore 138682
{bwuaa, simonwun}@cse.ust.hk, im.lee.chung@gmail.com, horner@cse.ust.hk

ABSTRACT

Music is one of the strongest inducers of emotion in humans. Melody, rhythm, and harmony provide the primary triggers, but what about timbre? Do the musical instruments have underlying emotional characters? For example, is the well-known melancholy sound of the English horn due to its timbre or to how composers use it? Though music emotion recognition has received a lot of attention, researchers have only recently begun considering the relationship between emotion and timbre. To this end, we devised a listening test to compare representative tones from eight different wind and string instruments. The goal was to determine if some tones were consistently perceived as being happier or sadder in pairwise comparisons. A total of eight emotions were tested in the study. The results showed strong underlying emotional characters for each instrument. The emotions Happy, Joyful, Heroic, and Comic were strongly correlated with one another. The violin, trumpet, and clarinet best represented these emotions. Sad and Depressed were also strongly correlated. These two emotions were best represented by the horn and flute. Scary was the emotional outlier of the group, while the oboe had the most emotionally neutral timbre. Also, we found that emotional judgment correlates significantly with average spectral centroid for the more distinctive emotions, including Happy, Joyful, Sad, Depressed, and Shy. These results can provide insights in orchestration, and lay the groundwork for future studies on emotion and timbre.

1. INTRODUCTION

Music is one of the most effective forms of media for conveying emotion. A lot of work has been done on emotion recognition in music, considering such factors as melody [3], rhythm [18], and lyrics [10]. However, little attention has been given to the relationship between emotion and timbre. Does the timbre of a particular musical instrument arouse specific emotions? For example, the English horn often plays a sad and melancholy character in orchestral music – is the melancholy character due to the instrument's timbre, the melody composers feel inspired to write for the instrument, or both? If listeners just heard an isolated tone from the English horn without a melodic context, would it sound more melancholy than other instruments? This paper addresses this fundamental question.

Some previous studies have shown that emotion is closely related to timbre. Scherer and Oshinsky found that timbre is a salient factor in the rating of synthetic tones [17]. Peretz *et al.* showed timbre speeds up the discrimination of emotion categories [15]. Bigand *et al.* reported similar results from their study of emotional similarities between one-second musical excerpts [4]. It was also found that timbre is essential to musical genre recognition and discrimination [2, 19].

Little attention has been given to the direct connection between emotion and timbre, though a study by Eerola *et al.* was an excellent start [6]. They carried out listening tests to investigate the correlations of emotions with temporal and spectral sound features. The study confirmed strong correlations between some features, especially attack time and brightness, and the emotional dimensions valence and arousal for one-second isolated instrument tones.

Valence and arousal refer to how positive and energetic a music stimulus sounds [21]. Despite the widespread use of these emotional dimensions in music research, composers may find them vague and difficult to interpret for composition and arrangement purposes. In our study, to make the results intuitive for composers, the listening test subjects compared sounds in terms of emotional categories such as Happy and Sad. The use of emotional categories has been shown to be generally congruent with results obtained using the dimensional model in music [7].

Moreover, we equalized the attacks and decays of the stimuli so that the temporal features attack time and decay time would not be factors in the subjects' judgment. This modification allowed us to isolate the effects of spectral features, such as the average spectral centroid, which strongly correlates with the perceptual brightness of sound.

The next section describes the listening test in detail. We report the listening test results in Section 3. Section 4

discusses applications of the results and questions arising from our study.

2. EXPERIMENTAL METHOD

2.1 Stimuli

The stimuli consisted of eight sustained tones from several wind and bowed string instruments: the bassoon (Bs), clarinet (Cl), flute (Fl), horn (Hn), oboe (Ob), saxophone (Sx), trumpet (Tp), and violin (Vn). They were obtained from the McGill and Prosonus sample libraries, except for the trumpet tone, which had been recorded at the University of Illinois at Urbana-Champaign School of Music. All the tones were used in a discrimination test carried out by Horner *et al.* [9], and six of them were also used by McAdams *et al.* [14].

The tones were used in their entirety, including the full attack, sustain, and decay sections. They were nearly harmonic and had fundamental frequencies close to 311.1 Hz (Eb4). The original fundamental frequencies deviated by up to 1 Hz (6 cents), and were synthesized by additive synthesis at 311.1 Hz. They were stored in the format of 16-bit samples at a 22050- or 44010-Hz sampling rate (depending on the number of harmonics with significant amplitudes).

Duration, loudness, and harmonic frequency deviations were equalized so that these factors would not influence the results. Furthermore, to isolate the effects of spectral features, the attacks and decays of the tone were equalized by time-stretching the actual attacks and decays. As a result, the stimuli were standardized to last for 2 seconds with attacks and decays 0.05 seconds long.

2.2 Subjects

32 subjects without hearing problems were hired to take the test. These undergraduate students ranged in age from 19 to 24. Half of them had music training (that is, at least five years of practice on an instrument).

2.3 Emotion Categories

The subjects compared the stimuli in terms of eight emotion categories: Happy, Sad, Heroic, Scary, Comic, Shy, Joyful, and Depressed. These terms were selected for their relevance to composition and arrangement by one of the authors, who had received formal composition education. Their ratings according to the Affective Norms for English Words [5] are shown in Figure 1 using the Valence-Arousal model. It is worth noting that Happy, Joyful, Comic, and Heroic form one cluster, and that Sad and Depressed form another cluster.

2.4 Listening Test

Every subject made pairwise comparisons of all the eight instruments. During each trial, the subjects heard a pair of tones from different instruments and were prompted to choose which tone more strongly aroused a given emotion. Figure 2 shows a screenshot of the listening test program.

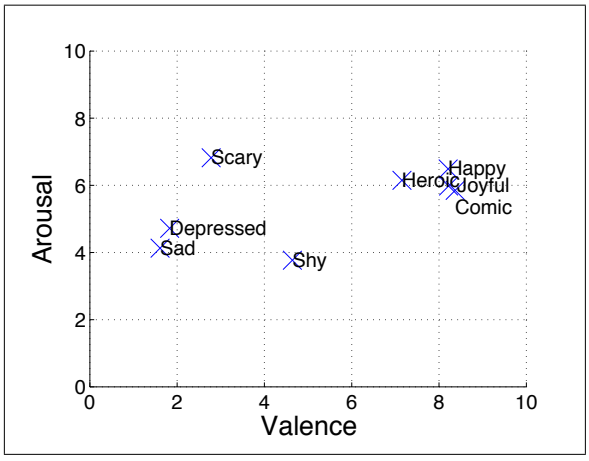

Figure 1. Russel's Valence-Arousal emotion model. Valence refers to how positive an emotion is. Arousal refers to how energetic an emotion is.

Each combination of two different instruments was presented in four trials for each emotion, and the listening test totaled $C_2^8 \times 4 \times 8 = 896$ trials. The overall trial presentation order was randomized.

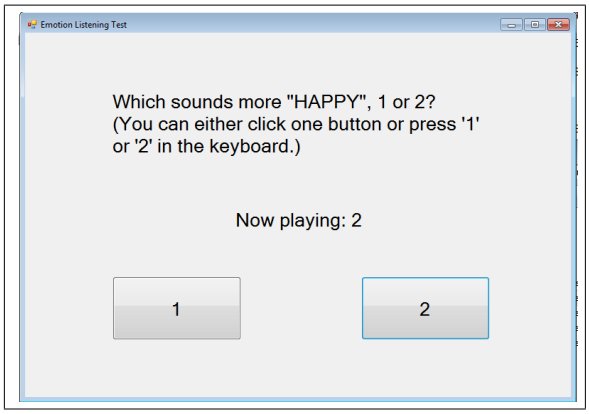

Figure 2. Listening test interface.

Before the first trial, the subjects read online definitions of the emotion categories from the Cambridge Academic Content Dictionary [1]. The listening test took about 1.5 hours, with breaks every 30 minutes.

The subjects were seated in a "quiet room" with less than 40 dB SPL background noise level. Residual noise was mostly due to computers and air conditioning. The noise level was reduced further with headphones. Sound signals were converted to analog by a Sound Blaster X-Fi Xtreme Audio sound card, and then presented through Sony MDR-7506 headphones at a level of approximately 78 dB SPL, as measured with a sound-level meter. The Sound Blaster DAC utilized 24 bits with a maximum sampling rate of 96 kHz and a 108 dB S/N ratio.

3. RESULTS

3.1 Quality of Responses

The subjects' responses were first screened for inconsistencies. A subject's consistency was defined in the four comparisons of a pair of instruments A and B for a particular emotion as follows:

$$consistency_{A,B} = \frac{max(v_A, v_B)}{4} \quad (1)$$

where v_A and v_B are the number of votes the subject gave to the two instruments respectively. A value of $consistency = 1$ represents perfect discrimination, whereas 0.5 represents random guessing. The mean of the average consistency of all subjects was 0.79.

Predictably the subjects were only fairly consistent because of the emotional ambiguities in the stimuli. We assessed the quality of subject responses further using a probabilistic approach. One probabilistic model, successful for image labeling, was adapted to suit this study [20]. The original model took the difficulty of labeling and the ambiguities in image categories into account. This was done to estimate the annotators' expertise and the quality of their responses. Those who made low-quality responses were unable to discriminate between image categories and were considered random pickers. In our study, we verified that the two least consistent subjects made responses of the lowest quality. They were excluded from the results.

We measured the level of agreement among the remaining subjects with an overall Fleiss' Kappa statistic. It was calculated at 0.22, which can be interpreted as indicating fair agreement [12].

3.2 Emotional Judgment

Figure 3 depicts the emotion "voting" results on a gray scale. Each row shows the percentage of positive votes an instrument received when compared to the other instruments for one particular emotion. The lighter the color of a cell, the more positive votes its "row instrument" received when compared to its "column instrument". For example, the bassoon was nearly always judged happier than the horn but usually not as happy as the clarinet.

The subjects gave clear-cut votes to distinctive emotions such as Sad and Depressed, for which the voting patterns have a lot of contrast. On the other hand, there were considerable ambiguities in the emotion comparisons for Scary.

The voting patterns for Sad, Depressed, and Shy were similar, suggesting that these emotions correlate with each other. Likewise, Happy and Joyful form another group of correlated emotions, which apparently includes Heroic and Comic.

We ranked the instruments in order of the number of positive votes they received for each emotion, and derived scale values using the Bradley-Terry-Luce (BTL) model [11]. Table 1 lists the rankings of the instruments, which can be visualized on a BTL scale in Figure 4. The instrument rankings for correlated emotions are similar. The horn and flute ranked high for the sad emotions, whereas the violin, trumpet, and clarinet ranked high for the happy emotions. Note that the oboe nearly always ranked in the middle.

The comparisons between instruments close in rank (e.g., the trumpet and the clarinet) were generally difficult. The votes received by a pair of such instruments could be close, corresponding to the grayest areas in Figure 3. By contrast, instruments ranking in different extremes (e.g., the horn and clarinet) could receive quite different numbers of votes, which correspond to bright and dark areas in Figure 3.

Table 2 shows the spectral characteristics of the test tones such as average spectral centroid. Table 3 shows close ties between test tone ranking and the average spectral centroid, as measured by the Spearman correlation coefficient. Emotional judgment correlated significantly with average spectral centroid, except for the less distinctive emotion Scary. For example, a high-centroid instrument is likely to sound happier, and a low-centroid instrument sadder. Emotional judgment did not have statistically significant correlations with the other spectral characteristics that we tested.

4. DISCUSSION

The pairwise correlations between emotions in our listening test are basically consistent with the pairwise relations between emotional words in the Valence-Arousal model in Figure 1. Both show Scary as the biggest outlier. Both show Happy, Joyful, Heroic, and Comic in a one cluster, and Sad and Depressed in another group. The biggest difference is that Shy is included in the sad group in this study, but it is emotionally-neutral in the Valence-Arousal model.

The results were consistent with those of Eerola's Valence-Arousal results for musical instrument tones [6]. Both show that musical instrument timbres carry cues about emotional expression that are easily and consistently recognized by listeners. Both show that spectral centroid/brightness is a significant component in music emotion.

The main application motivating this study is to provide guidelines for composers/arrangers in orchestration, especially for computer games, film, and stage music that needs to reflect characterization and dramatic action. The results provide some clear guidelines in achieving the desired emotional impact. Composers/arrangers can choose timbres that reinforce the desired emotion of their melody to achieve the strongest emotional impact. For example, composers could combine a joyful melody with instruments that have inherently joyful timbres. On the other hand, instrumental music, like opera arias, often include mixed emotions representing the complexity of characters and dramatic situations. The results of this study also provide a good starting point for achieving mixed emotions. For example, composers/arrangers might create a feeling of overall joyfulness mixed with an undertone of sadness by skillfully combining an otherwise joyful melody with the normally sad horn for a sophisticated mix of emotions.

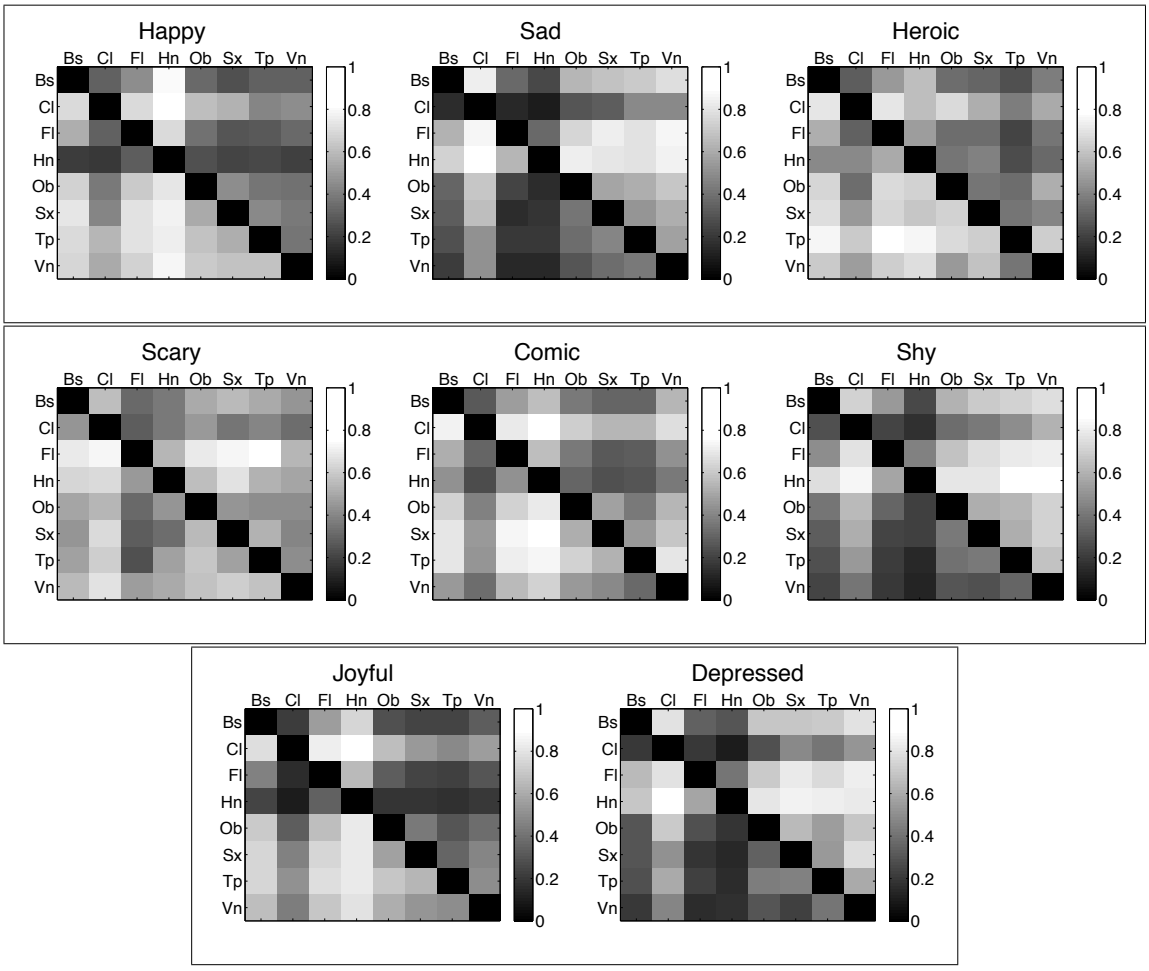

Figure 3. Comparison between instruments for each emotion. The lighter the color of a cell, the more positive votes its "row instrument" received when compared to its "column instrument".

In addition, instrument rankings for the different emotions may serve as a reference for combining instruments that blend together or contrast with each other emotionally.

Moreover, there are many other useful emotional labels to test. For example, one listener commented that the horn was not so much sad as mysterious. Is it the most mysterious timbre? What makes the instrument sound mysterious? Is it due to the reverberant reflections it makes? We could test the effect of reverberations by using tones recorded in an anechoic chamber and comparing them to those recorded in halls with varying reverberation times. In general, how does reverberation influence emotion? Does it make the sound more mysterious, majestic, or heroic? What is the effect of reverberation time and reverberation amount on emotion?

From an audio engineering and sound reproduction point of view, how do spectral alterations influence emotion? For example, how does the spectral distortion of playback equipment influence emotion? How about the effect of audio coding on emotion? For example, past work has shown that MP-3 compression with low bit-rates can cause a large reduction in brightness in the saxophone without affecting other instruments [13]. Is the emotional impact changed in a corresponding way? We expect it to be less joyful/happy – is it?

For future work, it will be also fascinating to see how emotion varies with different pitches, dynamic levels, brightness, and articulations. Do these parameters change perceived emotion in a consistent way, or does it vary from instrument to instrument? For example, we know that increased brightness makes a tone more dramatic (more happy or more angry), but is the effect more pronounced in some instruments and less so in others? For example, if the violin, which is the happiest instrument, is played softly with less brightness, is it still happier than the horn, which is the saddest instrument, if the horn is played loudly with maximum brightness? At what point are they equally happy? Can we normalize the instruments to equal happiness by simply adjusting brightness or other attributes? In the same way that we can normalize brightness by filtering or spectral tilting, can we normalize happiness by filtering or spectral tilting (or pitch or dynamic level)? How do the

Ranking \ Emotion	Happy	Sad	Heroic	Scary	Comic	Shy	Joyful	Depressed
1	Vn (5.23)	Hn (8.80)	Tp (3.38)	Fl (2.95)	Cl (3.44)	Hn (8.09)	Cl (7.30)	Hn (8.70)
2	Tp (4.53)	Fl (6.30)	Cl (2.23)	Hn (2.05)	Tp (3.23)	Fl (4.95)	Tp (6.94)	Fl (6.11)
3	Cl (4.42)	Bs (3.75)	Sx (1.93)	Vn (2.02)	Sx (2.84)	Bs (3.85)	Vn (5.35)	Bs (4.08)
4	Sx (3.91)	Ob (2.20)	Vn (1.95)	Tp (1.43)	Ob (2.20)	Ob (2.55)	Sx (5.15)	Ob (2.37)
5	Ob (3.10)	Sx (1.70)	Ob (1.65)	Bs (1.40)	Vn (1.63)	Sx (2.01)	Ob (3.79)	Sx (1.68)
6	Fl (2.02)	Tp (1.47)	Hn (1.05)	Sx (1.34)	Bs (1.37)	Tp (1.53)	Bs (2.17)	Tp (1.64)
7	Bs (2)	Vn (1.08)	Bs (1.03)	Ob (1.28)	Fl (1.30)	Cl (1.49)	Fl (1.84)	Cl (1.20)
8	Hn (1)	Cl (1)	Fl (1)	Cl (1)	Hn(1)	Vn (1)	Hn (1)	Vn (1)

Table 1. Rankings of instruments for each emotion from strongest to weakest with Bradley-Terry-Luce scale values in brackets.

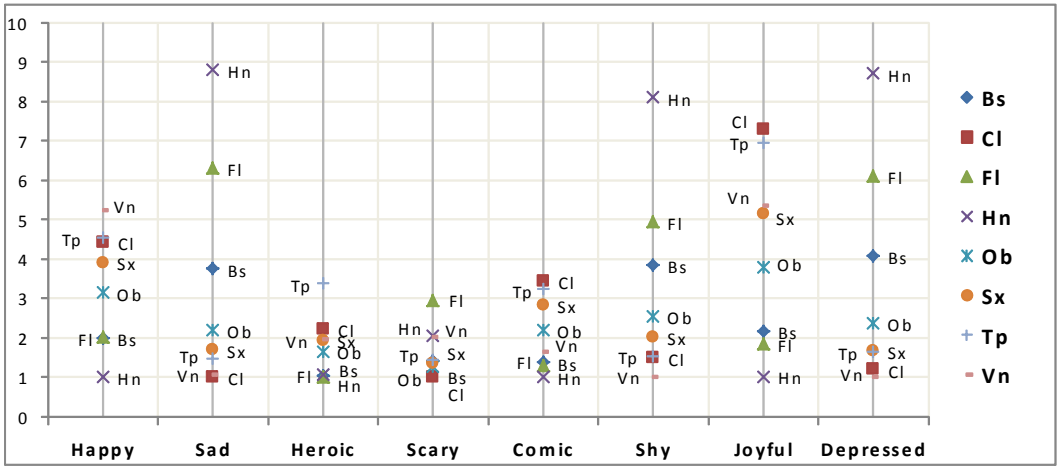

Figure 4. Bradley-Terry-Luce scale values of the instruments for each emotion.

happy spaces of the violin overlap with other instruments in terms of pitch, dynamic level, brightness, and articulation? In general, how does timbre space relate to emotional space?

Emotion gives us a fresh perspective on timbre, helping us to get a handle on its perceived dimensions. It gives us a focus for exploring its many aspects. Just as timbre is a multidimensional perceived space, emotion is an even higher-level multidimensional perceived space deeper inside the listener.

5. ACKNOWLEDGMENTS

This work has been supported by Hong Kong Research Grants Council grants HKUST613111 and HKUST613112.

6. REFERENCES

[1] "happy, sad, heroic, scary, comic, shy, joyful and depressed". *Cambridge Academic Content Dictionary*. Online: http://goo.gl/v5xJZ (17 Feb 2013).

[2] Jean-Julien Aucouturier, François Pachet, and Mark Sandler. The way it sounds: timbre models for analysis and retrieval of music signals. *IEEE Transactions on Multimedia*, 7(6):1028–1035, 2005.

[3] Laura-Lee Balkwill and William Forde Thompson. A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music perception*, pages 43–64, 1999.

[4] Emmanuel Bigand, Sandrine Vieillard, François Madurell, Jeremy Marozeau, and A Dacquet. Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition & Emotion*, 19(8):1113–1139, 2005.

[5] Margaret M Bradley and Peter J Lang. Affective norms for english words (ANEW): Instruction manual and affective ratings. *Psychology*, (C-1):1–45, 1999.

[6] Tuomas Eerola, Rafael Ferrer, and Vinoo Alluri. Timbre and affect dimensions: Evidence from affect and similarity ratings and acoustic correlates of isolated instrument sounds. *Music Perception: An Interdisciplinary Journal*, 30(1):49–70, 2012.

[7] Tuomas Eerola and Jonna K Vuoskoski. A comparison

Features \ Instrument	Bs	Cl	Fl	Hn	Ob	Sx	Tp	Vn
Average Spectral Centroid	3.3806	6.3843	3.4909	2.4229	4.2713	4.1218	4.2561	4.3502
Spectral Flux	3.9274	5.8697	12.659	7.0653	6.1665	7.7108	4.4892	12.187
Spectral Entropy	0.48388	0.53137	0.57382	0.38468	0.44436	0.49527	0.57133	0.54972
Spectral Spread	12.628	14.121	12.518	6.4839	7.4289	11.429	8.1665	10.464
Irregularity	0.17081	1.083	0.44738	0.18566	0.63774	0.85751	0.058612	0.5557
Roughness	0.5379	0.54442	0.017542	0.30809	0.11009	0.69387	0.033896	8.9796

Table 2. Spectral characteristics of the instrument test tones.

Timbre \ Emotion	Happy	Sad	Heoric	Scary	Comic	Shy	Joyful	Depressed
Average Spectral Centroids	**0.8333****	**-0.9048****	**0.6429***	-0.5714	**0.7619****	**-0.8810****	**0.8571****	**-0.8810****
Spectral Flux	0.0714	0.1667	-0.3095	0.5238	-0.381	0.0714	-0.2857	0.0714
Spectral Entropy	0.5714	-0.381	0.1905	0.3095	0.2619	-0.4048	0.4048	-0.4048
Spectral Spread	0.119	-0.3571	-0.0238	-0.3095	0.3095	-0.2619	0.3333	-0.2619
Irregularity	0.2619	-0.4524	0.2381	-0.5952	0.4286	-0.381	0.3571	-0.381
Roughness	0.381	-0.5238	0.3095	-0.3571	0.2381	-0.5714	0.381	-0.5714

Table 3. Spearman correlation between emotions and spectral features. **: $p < 0.05$; * : $0.05 < p < 0.1$.

of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2011.

[8] Cardillo G. Fleiss' kappa: compute the fleiss' kappa for multiple raters. `http://goo.gl/0DFKV`, 2007.

[9] Andrew Horner, James Beauchamp, and Richard So. Detection of random alterations to time-varying musical instrument spectra. *The Journal of the Acoustical Society of America*, 116:1800–1810, 2004.

[10] Yajie Hu, Xiaoou Chen, and Deshun Yang. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *Proceedings of ISMIR*, volume 10, 2009.

[11] David R Hunter. Mm algorithms for generalized bradley-terry models. *Annals of Statistics*, pages 384–406, 2004.

[12] Fleiss L Joseph. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382, 1971.

[13] Chung Lee, Andrew Horner, and James Beauchamp. Impact of mp3-compression on timbre space of sustained musical instrument tones. *The Journal of the Acoustical Society of America*, 131(4):3433–3433, 2012.

[14] Stephen McAdams, James W Beauchamp, and Suzanna Meneguzzi. Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. *The Journal of the Acoustical Society of America*, 105:882, 1999.

[15] Isabelle Peretz, Lise Gagnon, and Bernard Bouchard. Music and emotion: perceptual determinants, immediacy, and isolation after brain damage. *Cognition*, 68(2):111–141, 1998.

[16] James A Russell. Evidence of convergent validity on the dimensions of affect. *Journal of personality and social psychology*, 36(10):1152, 1978.

[17] Klaus R Scherer and James S Oshinsky. Cue utilization in emotion attribution from auditory stimuli. *Motivation and emotion*, 1(4):331–346, 1977.

[18] Janto Skowronek, Martin McKinney, and Steven Van De Par. A demonstrator for automatic music mood estimation. In *Proceedings of the International Conference on Music Information Retrieval, Vienna, Austria*, 2007.

[19] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on*, 10(5):293–302, 2002.

[20] Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in Neural Information Processing Systems*, 22(2035-2043):7–13, 2009.

[21] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H. Chen. A regression approach to music emotion recognition. *IEEE TASLP*, 16(2):448–457, 2008.