OPEN ACCESS Freely available online
Journal of the Audio Engineering Society
Vol. 65, No. 4, April 2017 (© 2017)
DOI: https://doi.org/10.17743/jaes.2017.0001

An Analysis of Low-Arousal Piano Music Ratings to Uncover What Makes Calm and Sad Music So Difficult to Distinguish in Music Emotion Recognition

YU HONG, CHUCK-JEE CHAU, AND ANDREW HORNER, *AES Member*
(yhongag@cse.ust.hk) (chuckjee@cse.ust.hk) (horner@cse.ust.hk)

Department of Computer Science and Engineering Hong Kong University of Science and Technology,
Clear Water Bay, Kowloon, Hong Kong

Music emotion recognition and recommendation systems often use a simplified 4-quadrant model with categories such as Happy, Sad, Angry, and Calm. Previous research has shown that both listeners and automated systems often have difficulty distinguishing low-arousal categories such as Calm and Sad. This paper seeks to explore what makes the categories Calm and Sad so difficult to distinguish. We used 300 low-arousal excerpts from the classical piano repertoire to determine the coverage of the categories Calm and Sad in the low-arousal space, their overlap, and their balance to one another. Our results show that Calm was 40% bigger in terms of coverage than Sad, but that on average Sad excerpts were significantly more negative in mood than Calm excerpts were positive. Calm and Sad overlapped in nearly 20% of the excerpts, meaning 20% of the excerpts were about equally Calm and Sad. Calm and Sad covered about 92% of the low-arousal space, where 8% of the space were holes that were not-at-all Calm or Sad. The largest holes were for excerpts considered Mysterious and Doubtful, but there were smaller holes among positive excerpts as well. Due to the holes in the coverage, the overlaps, and imbalances the Calm-Sad model adds about 6% more errors when compared to asking users directly whether the mood of the music is positive or negative. Nevertheless, the Calm-Sad model is still useful and appropriate for applications in music emotion recognition and recommendation such as when a simple and intuitive interface is preferred or when categorization is more important than precise differentiation.

0 INTRODUCTION

Previous research has made good progress on the problem of music emotion recognition [1–44], with a wide variety of musical applications [73–78]. Some music emotion recognition systems have used dimensional models, most commonly describing the valence or positiveness of the music in one dimension and its arousal or energy-level in a second dimension [1–12, 44–46, 72]. Other systems have used categorical models, using adjectives to describe the character expressed by the music or the experienced emotion of the listener [13–34], or simply dividing the valence-arousal plane usually by quadrants [6, 17–18, 35–39, 44].

A particularly popular categorical model for music emotion recognition is the 4-quadrant model [6, 13–15, 17–30, 35–39, 44]. It simplifies the valence-arousal plane into four distinct quadrants with labels such as Happy, Sad, Angry, and Calm (see Figure 1). Alternative category names are also common, such as Scary or Fearful instead of (or in addition to) Angry [23, 47], and Peaceful, Relaxed, or Tender instead of (or in addition to) Calm [23–25, 32, 47–48]. In any case, a big advantage of this model is its simplicity—the four categories are natural and intuitive dimensions of the valence-arousal plane. They are universally understood opposites, and they are concrete representations of the abstract valence-arousal plane.

Researchers have frequently used the 4-quadrant model as a basic way of categorizing music. They have also used it in music emotion recognition systems and have compared the responses of listeners to the predicted categorization of these systems.

Many researchers have noted that automated 4-quadrant models generally do a very good job in distinguishing high and low arousal music and, therefore, do well distinguishing category pairs such as Happy-Calm and Angry-Sad. The systems also usually do well distinguishing Happy-Angry. The most difficult case is Calm-Sad. This case usually accounts for the largest errors in 4-quadrant music emotion

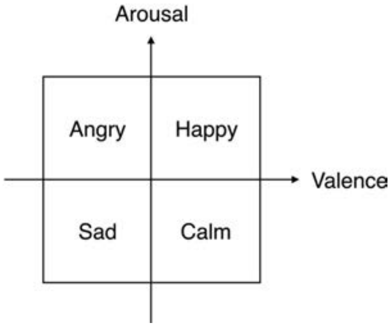


Fig. 1. Simplified 4-quadrant categorical model.

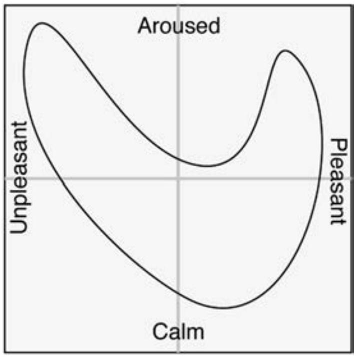


Fig. 2. An arousal-valence emotional space. Adapted from IAPS (Lang et al., 1988) and IADS (Bradley and Lang, 1991) by Dietz [53].

recognition systems [1–2, 14, 17, 19–20, 23–28, 32, 35–38, 44, 47–51]. Avoiding the problem, some researchers have used 4-category models, but actually used three high-arousal categories and only one low-arousal category [15–16].

So, what makes the categories Calm and Sad so difficult to distinguish? Several previous researchers have noted that valence is harder to distinguish than arousal [1–4, 16, 25, 44]. And while many previous researchers have identified the distinguishability of Calm and Sad as a problem, only a few of them have indicated why it is a problem. Bradley [52] conducted an experiment to determine the valence and arousal value of many English words and found that the distribution followed a parabolic shape (see Figure 2). That is, in the low arousal region, the mood was normally distributed, while in the high arousal region, the mood was either very positive or very negative. Dietz [53] also found a similar result. Naji [23] suspected that the confusion might be due to mixed feelings of the listeners. Pouyanfar [24] suggested that the confusion was mostly due to similarity of the low arousal classes.

This paper takes a close look at the Calm-Sad case for low-arousal excerpts. The difficulty in distinguishing Calm and Sad is one of the biggest bottlenecks in music emotion recognition. Though several studies have identified the bottleneck [1–2, 14, 17, 19–20, 23–28, 32, 35–38, 44, 47–51] and addressed it briefly [1–2, 16, 23–24, 52–53], it deserves a careful and detailed study. If we can understand it better, we can start to address the bottleneck in a systematic way.

Taking a fresh look at the problem, there are several possible reasons why the categories Calm and Sad are so hard to distinguish. First, perhaps Calm and Sad do not fully cover their respective quadrants, leaving the extremities and boundaries uncovered. For example, even if the music is low-arousal and negative in mood, its character may be suspensefully Scary and not-at-all Sad. If listeners are forced to choose a category in the 4-quadrant model, some may choose Angry as the closest match, some may choose Sad, and some may even choose Calm thinking of it as “the calm before the storm.” Second, perhaps the categories Calm and Sad overlap, as Naji suggests [23], making it difficult to determine which is more dominant. It is not difficult to imagine music that is “Sad, but Calm” or “Calm, but Sad,” so we probably don’t have to worry about a gap between Calm and Sad, but there could certainly be music where both are expressed about equally. Third, and this is the subtlest possibility, the categories Calm and Sad may not form a well-balanced pair. For example, Sad might be more negative than Calm is positive, perhaps so much so that Sad covers the most negative third of the low-arousal space, Calm the middle third, and some other more positive word such as Peaceful the most positive third.

In this paper we explore these questions using low-arousal musical excerpts drawn from a representative cross-section of the classical piano standard repertoire. What is the extent of Calm and Sad’s coverage of their respective low-arousal quadrants? How much overlap exists between Calm and Sad? And do Calm-Sad form a well-balanced pair? We conducted a series of listening tests to address the issues above. The answers to these questions will help us understand why Calm and Sad are so easily confused by both listeners and music emotion recognition systems. Hopefully, they will also suggest solutions to these issues and improve the accuracy of music emotion recognition systems. In turn, they may also help in related applications such as music recommendation based on the listener’s mood or previous music preference [14–15].

1 METHOD

1.0 Overview

In order to better understand why listeners and music recognition systems often confuse the emotional categories Calm and Sad we designed a series of listening tests to evaluate the coverage, overlap, and balance of Calm and Sad.

We chose the genre of classical piano music for our study, in part because it minimizes the effect of timbre which is particularly simple within this genre with only one instrument.

We selected 100 low-arousal classical piano pieces/movements that gave a reasonably balanced distribution across the stylistic periods. Our focus on exclusively low-arousal excerpts narrowed the choice considerably. The pieces we chose are listed in the

Appendix. We picked pieces by well-known piano composers from the Baroque, Classical, Early Romantic, Late Romantic, and Early 20th Century periods. The balance between the periods is important because it determines the harmonic language included in the pool of excerpts. It also determines the proportion of excerpts that are distinctly positive or negative compared to those that are more ambiguous or neutral. For example, Baroque and Classical pieces might be more distinctly positive or negative, while Late Romantic and Early 20th Century pieces might be more often ambiguous in mood. In order to achieve a good stylistic balance, we picked 5 Baroque pieces, 16 Classical pieces, 21 Early Romantic pieces, 32 Late Romantic pieces, and 26 Early 20th Century pieces. Combining the Baroque and Classical groups, there were a similar number of pieces from each period but with a little extra weight on the Late Romantic and Early 20th Century periods. This balance between the periods seemed appropriate.

Then, we selected 3 contrasting 10-second low-arousal excerpts from each piece. We wanted to avoid repetitions of the same phrase from being selected twice, so we picked contrasting excerpts that avoided repetitions or near-repetitions (e.g., an octave higher in the repeat). We did not try to select the excerpts based on how Calm or Sad they were or how positive or negative they were.

We selected excerpts where the character of the music was generally maintained over its duration and tried to avoid phrase boundaries that included the end of one phrase and the beginning of the next. Then we added one-second fade-ins and fade-outs followed by one-second of silence and normalized the amplitude levels of the excerpts by the maximum-energy of the sound in a 0.5 second window. The normalization factor is given by:

$$Normalization\ Factor = \sqrt{\frac{mean\{\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{300}\}}{\alpha_i}}$$

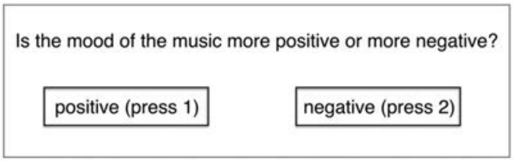
where

$$i = current\ excerpt\ number$$

$$\alpha_i = max\{Total\ energy\ over\ every\ continuous\ 0.5s\ segment\ of\ excerpt\ i\}.$$

We listened to the normalized excerpts and verified that they sounded at about the same loudness level.

The 300 excerpts were presented in a different random order for each listening test and listener. The length of each test was about 50 minutes (10-second excerpts × 300 excerpts). To avoid fatigue, about half way through each test, we had subjects take a forced short 5-minute break before resuming the test. Also, listeners could take breaks between examples whenever desired. But, listeners could not rewind, fast forward, or repeat excerpts. Listeners could not modify their selection once made; we wanted their first reaction. The computer listening test program would not accept an answer until the entire excerpt had been played. Once subjects had selected an answer, the next excerpt was played automatically. We adjusted the volume on all the listening test computers to the same moderate level before the test and asked listeners not to adjust the level during the test. They did the listening test individually with basic-level professional headphones (Sony MDR-7506).



Is the mood of the music more positive or more negative?

Fig. 3. Computerized graphical user interface for our first listening test.

Subjects were undergraduate students from the Hong Kong University of Science and Technology, mostly ranging in age from 19 to 23 with a mean of 21.0 and standard deviation of 1.8. We asked subjects about any hearing issues before the test, and none of them reported any hearing problems. They were not music school students or professional musicians but average attentive listeners. Most of them played a musical instrument, and about half of them had played the instrument for at least five years. We aimed to include about 30 subjects in each of our listening tests, though the exact number varied somewhat and is given in the subsections below in the description of each particular test. Also, we checked subjects' responses and excluded a few subjects (about 10%) who were obviously not focusing on the test and giving spam responses. We made the determination based on their keystrokes and overall outlier responses.

1.1 First Test: Positive and Negative Mood

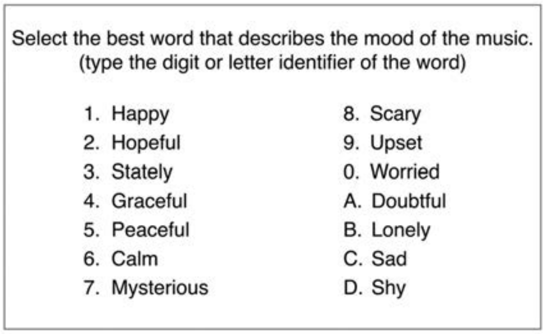
In our first listening test, subjects were asked "Is the mood of the music more positive or more negative?" The number of subjects was 26 after excluding spammers. Figure 3 shows the computerized listening test interface. Listeners could respond by typing "1" or "2" from the normal keys or the numeric keypad, or by clicking with the mouse. There were 6 training examples at the beginning of the test for listeners to become accustomed to the test environment. The training examples were picked randomly from the 300 excerpts and the responses were not used in the results. The 6 excerpts were also included in the regular test where listeners' responses were used.

The purpose of this test was to determine the valence values for each of the 300 excerpts. Positive replies were taken as 1 and negative replies as 0 and the average over all listeners determined the valence for each excerpt. If all listeners were positive, the excerpt would be maximally positive. If all listeners were negative, the excerpt would be maximally negative. If half of the listeners were positive and half negative, the excerpt would be neutral.

The advantage of this comparison is its simplicity. Listeners only need to make a simple binary forced choice decision for each excerpt. It is a less complex task than asking listeners to judge gradations in valence directly.

Table 1. 14 categories and dictionary definitions shown to subjects before taking our second listening test.

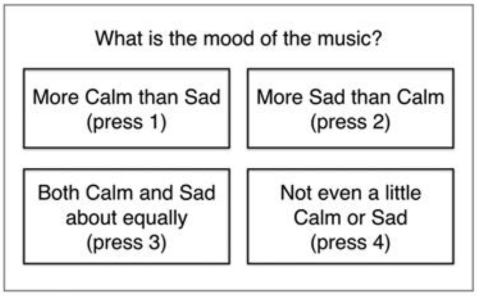
HAPPY: glad, pleased
HOPEFUL: believing that what you hope for is likely to happen
STATELY: formal, slow, and having a style and appearance that causes admiration
GRACEFUL: moving in a smooth and attractive way
PEACEFUL: quiet and calm; not worried or disturbed in any way
CALM: relaxed and quiet, not angry, nervous, or upset
MYSTERIOUS: exciting wonder, curiosity, or surprise while baffling efforts to comprehend or identify
SCARY: causing fright
UPSET: unhappy because something unpleasant or disappointing has happened
WORRIED: unhappy because you keep thinking about something bad that might happen
DOUBTFUL: unlikely to be successful
LONELY: unhappy because you are alone or do not have anyone to talk to
SAD: affected with or expressive of grief or unhappiness
SHY: nervous and embarrassed about meeting and speaking to other people



Select the best word that describes the mood of the music.
(type the digit or letter identifier of the word)

1. Happy
2. Hopeful
3. Stately
4. Graceful
5. Peaceful
6. Calm
7. Mysterious
8. Scary
9. Upset
0. Worried
A. Doubtful
B. Lonely
C. Sad
D. Shy

Fig. 4. Interface for our second listening test.



What is the mood of the music?

More Calm than Sad
(press 1)

More Sad than Calm
(press 2)

Both Calm and Sad
about equally
(press 3)

Not even a little
Calm or Sad
(press 4)

Fig. 5. Interface for our third listening test.

1.2 Second Test: Best Word from 14 Categories

In our second test 31 subjects (about a third were the same as those who took our first listening test) were asked to select the best word that described the mood of the music from a list of 14 categories. Table 1 shows the dictionary definitions of the 14 categories that were presented to listeners just before taking the test.

Figure 4 shows the listening test interface. We selected the 14 words based on an informal free-choice pre-test. The informal pre-test was done by the 3 authors. The excerpts in the pre-test and the excerpts in the regular test were the same. Each of us tried to describe the excerpt in a single one-word adjective, and the 14 words we used most frequently were selected as the 14 categories. It turns out that about half of the 14 words were used in our previous related studies [79–96] and most of the others were used in studies by other researchers [9, 20, 21, 35, 47]. All the categories included in the 4-quadrant model in Figure 1 appear in Figure 4 except Angry. Why? When we took the pre-test, Angry seemed too strong for the low-arousal excerpts and we more often picked the milder word Upset, so Angry seemed unnecessary for this test.

The purpose of this test was to see which words would be most frequently chosen and whether listeners would consistently map negative words such as Scary and Lonely to Sad and positive words to Calm in our later tests, where they were forced to choose between Calm and Sad. We were particularly interested to see what the valence values would be for more neutral words such as Mysterious and Shy that can take on shades that are either positive (sagely-Mysterious, playfully-Shy) or negative (diabolically-Mysterious, fearfully-Shy).

We were also interested to see which words were picked most frequently. In particular, we wanted to see which words were picked most frequently among positive, negative, and neutral excerpts separately.

1.3 Third Test: Calm, Sad, Both, Neither

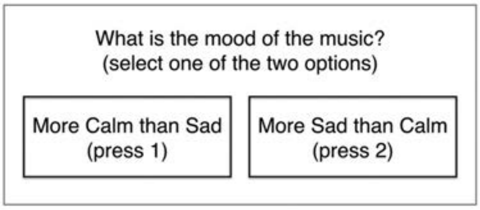
In our third test 31 subjects (the same subjects as in our second listening test) were asked to select one of four alternative categories to describe the mood of the music. Figure 5 shows the listening test interface.

The purpose of this test was to determine the coverage and overlap of the categories Calm and Sad. The “More Calm than Sad” option allowed listeners to identify excerpts covered by Calm. Similarly, “More Sad than Calm” allowed listeners to identify excerpts covered by Sad. The “Both” option allowed listeners to identify excerpts in the overlap region between Calm and Sad. Together with the valence determined by our first listening test, this allowed us to identify the precise region where overlaps occurred.

The “Neither” option allowed listeners to identify excerpts that were outside the coverage of Calm and Sad. We were particularly interested to see where such holes existed and whether they were very negative, very positive, extremely low-arousal, or borderline high-arousal (i.e., mid-arousal). For listeners who picked “Neither” for a particular excerpt, we were also interested to see which word they picked for the same excerpt in our 14-category second listening test (since the listeners were the same in both tests, we had this possibility available).

1.4 Fourth Test: Forced Calm and Sad

For our fourth test we wanted listeners to categorize all the excerpts as either Calm or Sad. Since subjects had



What is the mood of the music?
(select one of the two options)

Fig. 6. Interface for our fourth listening test.

already categorized some of the excerpts as Calm or Sad in our third test, in our fourth test the same 31 subjects were forced to choose either "More Calm than Sad" or "More Sad than Calm" for excerpts they had previously judged "Both" or "Neither." Listeners did not know that they were only hearing excerpts they had previously categorized as Both or Neither. Since there were so many excerpts (300), and there was a break between the tests, it was impossible to remember their previous responses. We feel the derived Calm-Sad result would have been basically the same if we had asked them to do a two-alternative forced choice Calm-Sad test for all 300 excerpts. Figure 6 shows the listening test interface.

The purpose of this follow-up test was to see how listeners would respond in a 4-quadrant environment where they were forced to choose either Happy, Sad, Angry, or Calm. This allowed us to estimate the error rates for the system with low-arousal excerpts.

It also allowed us to see whether Calm or Sad predominated in examples that were identified as "Both" in our third listening test. We were interested to see whether subjects consistently map negative "Neither" excerpts to Sad and positive "Neither" excerpts to Calm.

2 RESULTS

2.0 Overview

This section describes the results of our four listening tests with low-arousal classical piano excerpts. It evaluates the coverage, overlap, and balance of the emotional categories Calm and Sad. The implications for 4-quadrant music emotion recognition systems are also touched on and considered more fully in the following discussion section.

2.1 First Test: Positive and Negative Mood

For the first test listeners were asked whether the mood of each excerpt was positive or negative. A negative reply was taken as 0 and a positive reply as 1, and the average over all listeners used as the normalized valence for each excerpt. When we selected the excerpts, though we aimed to pick three contrasting excerpts from each piece, we did not try to pick exactly half negative in mood and half positive. But, the results happened to come out about that way with an average valence of 0.493 for all 300 excerpts. Figure 7 shows an almost perfectly symmetric distribution of the excerpts when evenly divided into five classes ranging from very negative to very positive. About twice as many excerpts were in each of the very negative and very positive classes compared to the inner three.

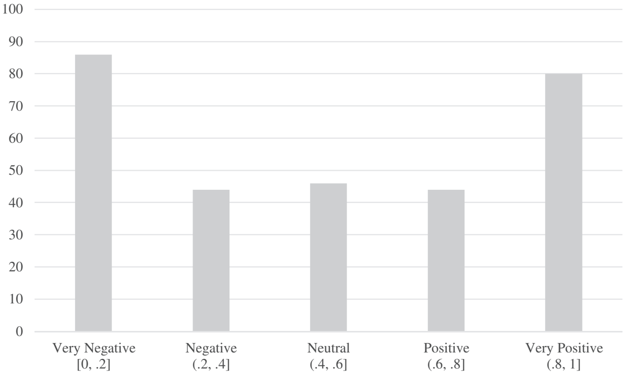


Fig. 7. Among our 300 excerpts the number of excerpts that were very negative, negative, neutral, positive, and very positive based on the average valence (0 means all listeners rated it negative, 1 means all positive, and 0.5 means listeners were exactly evenly divided).

Table 2. Results of second listening test. The percentage each category was selected and its average valence over all listeners and excerpts.

Category	Percentage	Avg. Valence
Calm	14.8%	.60
Peaceful	11.1%	.71
Graceful	8.7%	.73
Sad	8.1%	.31
Worried	8.0%	.27
Lonely	7.6%	.36
Mysterious	7.5%	.34
Stately	6.9%	.61
Doubtful	6.2%	.28
Hopeful	6.0%	.73
Upset	5.7%	.34
Scary	3.8%	.18
Shy	2.9%	.42
Happy	2.7%	.79

2.2 Second Test: Best Word from 14 Categories

For the second test, listeners were asked to select the best word that described the mood of the music from a list of 14 categories. Table 2 shows the percentage that each of the 14 categories were selected averaged over all listeners and excerpts. Figure 8 shows the average valence and the 95% confidence intervals for the 14 categories. The confidence intervals were calculated based on a CDF-based nonparametric method [97, 98].

Calm was the most-frequently chosen category. The second most-frequent category was Peaceful, which is very similar in meaning to Calm but significantly more positive. The third most-frequent was Graceful, another positive valence word. The most-frequent negative valence category was Sad at fourth, slightly more than Worried. Overall, there was about an even balance of positive and negative categories, with six positive, seven negative, and one

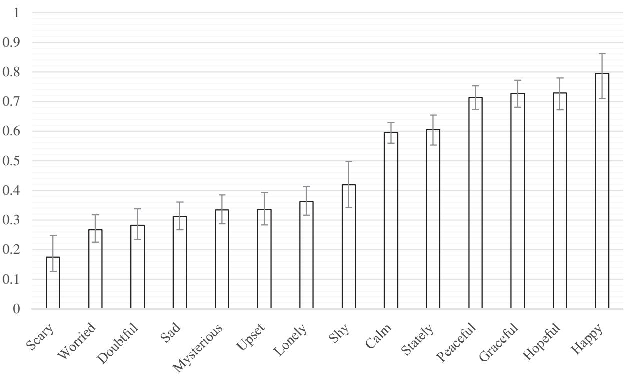


Fig. 8. The average valence and the 95% confidence intervals for the 14 categories.

Table 3. Results of the third listening test: The percentage each category was selected and its average valence over all listeners and excerpts.

Category	Percentage	Avg. Valence
Calm	44.2%	.65
Sad	28.7%	.30
Both	18.9%	.45
Neither	8.2%	.40
All	100%	.49

ambiguous (Shy). Calm and Sad were the most-frequently chosen positive and negative categories respectively.

We also considered whether the order of the words in Figure 4 might have biased listener responses. It turns out that the top-most words Happy and Scary were actually among the least-frequently chosen. Also, the most commonly-chosen words Calm and Peaceful were in the middle near the bottom. Therefore, the word order in Figure 4 did not seem to bias the results in an obvious way.

Among the least-frequently chosen categories were borderline high-arousal words such as Happy, Scary, and Upset. It is not surprising that they would be infrequent since the excerpts were deliberately chosen to exclude high-arousal excerpts. However, we included these categories to take into account low-arousal excerpts that might be serenely-Happy, suspensefully-Scary, and mildly-Upset.

Figure 9 shows a more detailed breakdown of the 14 categories into 5 valence classes: very negative, negative, neutral, positive, and very positive. The 5 classes are formed from equal divisions of the valence values between 0 and 1. So for example, the very positive class contains excerpts that were rated positive by 80–100% of listeners. Specifically, Figure 9 shows the percentage each of the 14 categories was selected for the 5 classes of excerpts. The percentage has been normalized so that the 14 categories sum to 1.0 in each of the 5 classes. Negative categories are shown in the left chart and positive categories on the right chart (plus Shy).

Figure 9 also includes the 95% confidence intervals. When we calculated the confidence interval for each category within each class (e.g., the Calm category within the very positive class), each sample was defined as the percentage of listeners that chose that category for each of the excerpts in the class. We then calculated the mean and standard deviation for all the samples in each category and class. Finally, we calculated the 95% confidence intervals based on the means and standard deviations.

Among the negative category curves in Figure 9, most were fairly similar to one another except Scary, which was much less frequently chosen than the others for several classes (negative to positive). Sad was the most frequently chosen negative category except for the very negative class where Worried was most.

For positive categories there was more variation. Calm was the most frequently chosen positive category, except for the very positive class where its more positive counterpart Peaceful was most. Calm had a wider coverage than the other positive categories and was much more frequently chosen than other positive categories in the neutral and negative classes. With only low-arousal excerpts in the stimuli, Happy was much less frequently chosen even for very positive excerpts. Shy was also much less frequently chosen but it was the most evenly distributed category.

2.3 Third Test: Calm, Sad, Both, Neither

For the third test, listeners were asked to select one of four categories to describe the mood of the music:

(1) More Calm than Sad;
(2) More Sad than Calm;
(3) Both Calm and Sad about equally;
(4) Not even a little Calm or Sad.

Table 3 shows the percentage each category was selected averaged over all listeners and excerpts. The percentage of Calm excerpts was almost as much as the Sad and Both excerpts together. Table 3 also lists the average valence for each category. Calm was slightly less positive than Sad negative, and this negative offset was also present in the Both valence that was slightly negative. The small number of Neither excerpts were even more negative. As we noted in Sec. 2.1 there were about an equal percentage of positive and negative excerpts. The relatively large percentage of Calm excerpts balanced the fact that Sad was more negative than Calm positive. Figure 10 shows the average valence and the 95% confidence intervals [97, 98] for each category.

Figure 11 shows a more detailed breakdown of the four categories into five classes from very negative to very positive. For Calm, Sad, and Neither the curves were nearly linear. As expected, Sad dominated for negative and very negative classes and Calm dominated for positive and very positive classes. But Sad was offset lower: about 75% of the very positive excerpts were rated Calm, while only 50% of the very negative excerpts were rated Sad. A surprisingly large percentage (20%) of the very negative excerpts were rated Calm.

Unlike the other categories, Both reached its maximum in the middle of the curve, though it was arched less than

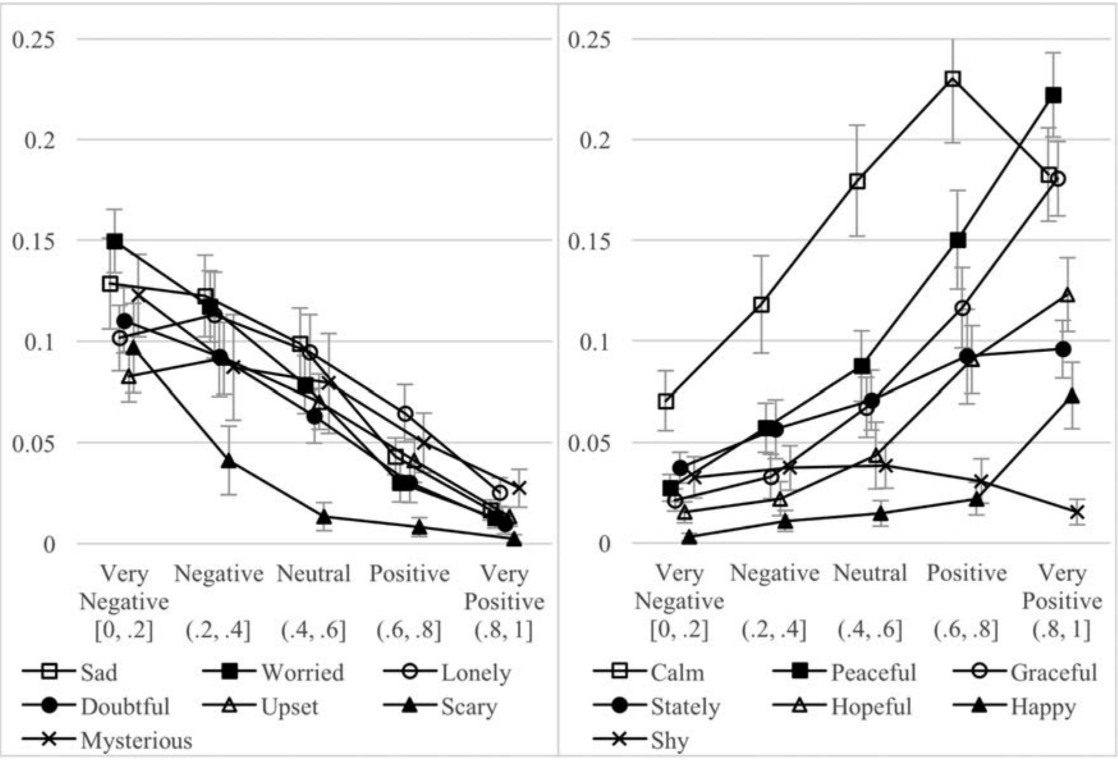


Fig. 9. The percentage each of the 14 categories were selected for very negative to very positive classes of excerpts. The percentages have been normalized so that the 14 categories sum to 1.0 in each of the 5 classes. Negative categories are shown in the left chart and positive categories on the right chart (plus Shy). 95% confidence intervals are also included.

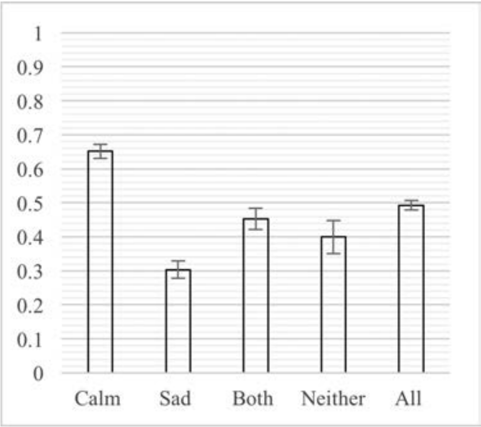


Fig. 10. The average valence and the 95% confidence intervals for each category.

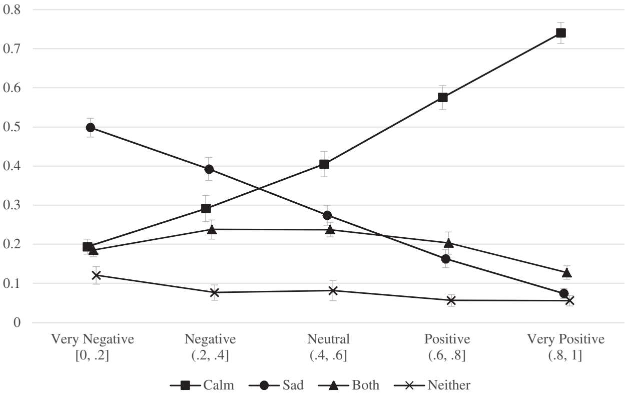


Fig. 11. The percentage of Sad, Calm, Both, Neither choices for very negative, negative, neutral, positive, and very positive classes of excerpts. The percentage have been normalized so that the 4 categories sum to 1.0 in each of the 5 classes. 95% confidence intervals are also included.

one might have expected with a relatively even distribution from very negative to positive. Relatively few excerpts were rated Neither (less than 10%), and its distribution was fairly flat with a slight tilt up on the negative side.

For excerpts where listeners selected “Not even a little Calm or Sad” (Neither), we were interested to know which of the 14 categories they picked when they heard the same excerpt in our second listening test. Table 4 shows the result. The data gives us an idea about which parts of the low-arousal plane were not covered by Calm and Sad. Mysterious topped the list followed by three negative categories: Doubtful, Worried, and Scary. Graceful and Stately were the top positive categories. Figure 12 shows the average valence and the confidence intervals [97, 98] for the 12 categories in Table 4.

2.4 Fourth Test: Forced Calm and Sad

For the fourth test, listeners were forced to choose either “More Calm than Sad” or “More Sad than Calm” for excerpts they previously judged “Both” or “Neither” in our

Table 4. When listeners selected “Not even a little Calm or Sad” for an excerpt in our third listening test, this table shows what they picked for the same excerpt in our second listening test and how often (excluding the small number of inconsistent cases where they picked Calm or Sad).

Category	Percentage	Avg. Valence
Mysterious	13.1%	.29
Doubtful	9.4%	.23
Worried	9.3%	.22
Scary	9.3%	.13
Graceful	9.2%	.68
Stately	7.9%	.52
Lonely	6.9%	.33
Hopeful	6.7%	.73
Peaceful	4.8%	.63
Upset	4.1%	.30
Shy	3.1%	.32
Happy	2.9%	.80

Table 5. The percentage of Calm and Sad and their average valence for the two-alternative forced Calm and Sad test derived from our third and fourth listening tests.

Category	Percentage	Avg. Valence
Calm	58.6%	.62
Sad	41.4%	.31

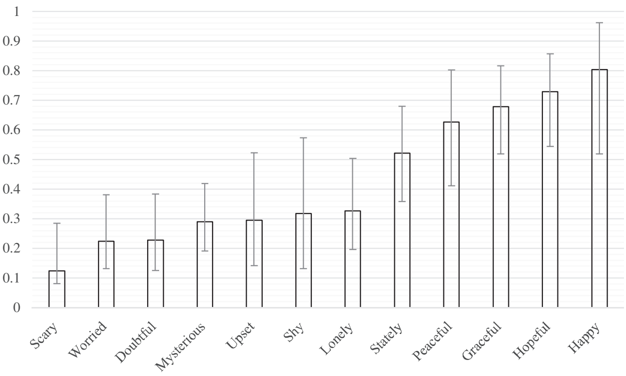


Fig. 12. The average valence and the 95% confidence intervals for the 12 categories in Table 4.

third listening test. This allows us to derive the results of a two-alternative forced Calm and Sad test for all excerpts (by taking the Calm and Sad results from our third listening test and replacing the Both and Neither votes from that same test with the Calm and Sad votes from our forth test). Table 5 shows the results. The coverage of Calm was more than 40% bigger than Sad. The average valence for Calm was slightly less positive than in Table 3 while Sad remained about the same. This further increased the imbalance in Table 3 where Sad was already more negative than Calm positive. Figure 13 shows the average valence and the confidence intervals for Calm and Sad. We performed a non-parametric CDF-based significance test [97] on the categorical data and found that Sad excerpts were significantly more negative in mood than Calm excerpts were positive ($p < 0.001$).

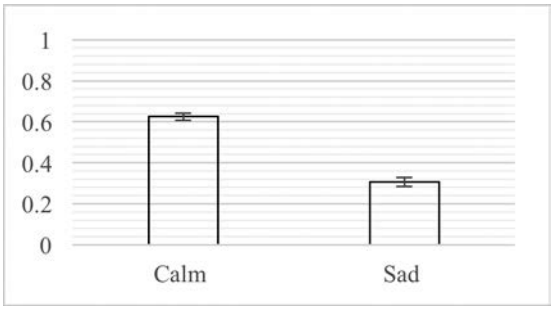


Fig. 13. The average valence and the 95% confidence intervals for the 2-alternative forced Calm and Sad test.

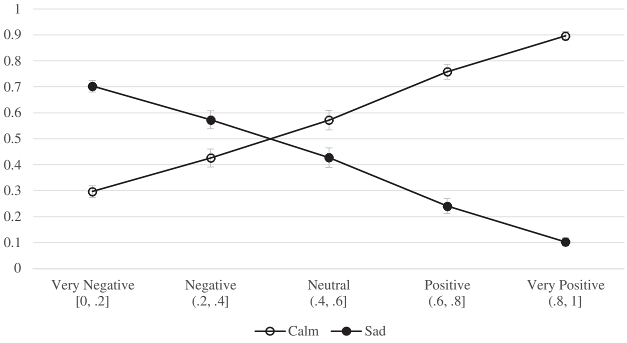


Fig. 14. Derived Calm and Sad percentages divided over 5 valence classes. The percentages have been normalized so that the 2 categories sum to 1.0 in each of the 5 classes.

Figure 14 shows the results divided over five valence classes. The curves in Figure 14 are similar to the curves for Calm and Sad in Figure 11. Most of the reclassified negative excerpts were judged Sad and even more of the reclassified positive excerpts were judged Calm. The balance point where Calm and Sad are equal remained about the same midway between the negative and neutral classes. Again, neutral excerpts were judged Calm more often than Sad. And negative excerpts were judged Calm more often than positive excerpts were judged Sad. Overall, Calm was picked about 20% more frequently than its mirror-image Sad counterpart (e.g., Calm for the very positive class is 90% and Sad for the very negative class is 70%, so the difference is 20%). The difference was relatively constant among the five classes.

For reference, Figure 15 shows the individual results for Both (left chart) and Neither (right chart). For Both and especially Neither, the results are not significantly different from Figure 14, indicating listeners were very consistent about labeling excerpts according to the mood of the music, even when they felt the music was not at all Calm or Sad.

2.5 Failure Rates for Distinguishing Low-Arousal Quadrants

In the previous section we saw in Figure 14 that some positive excerpts were rated Sad and even more negative excerpts were rated Calm. These cases are problematic in

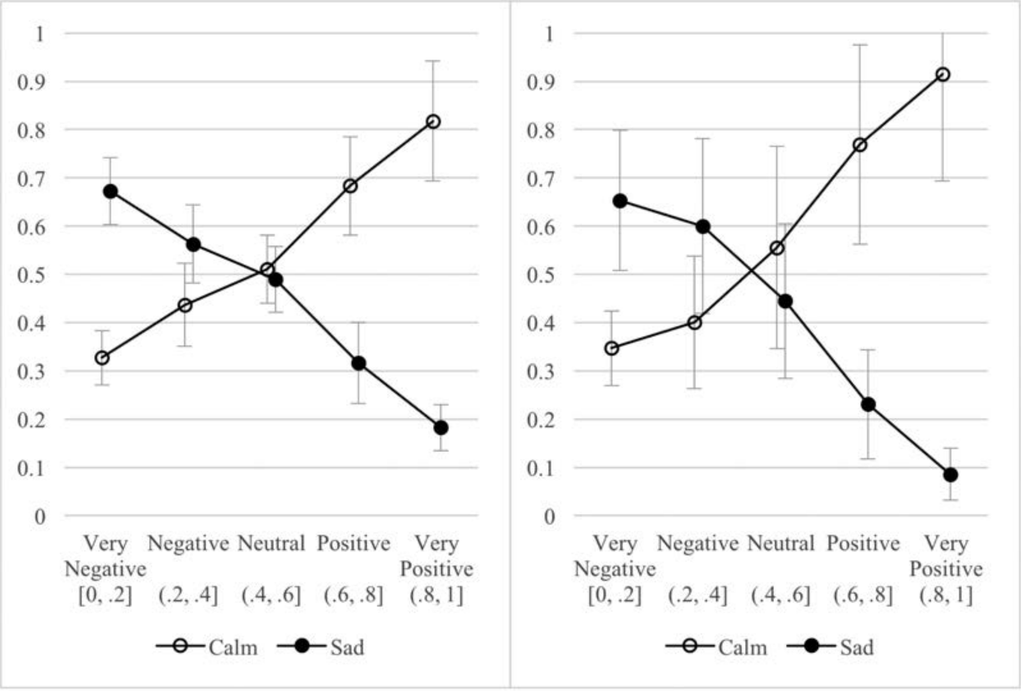


Fig. 15. Results of forced Calm and Sad test for listeners who chose Both (left chart) or Neither (right chart) on the third listening test. 95% confidence intervals are also included.

distinguishing low-arousal quadrants in a music emotion recognition system and can be considered as fail cases. Based on our listening test results, we can find the failure rates for distinguishing the two low-arousal quadrants.

We take listeners' positive-negative judgments in our first listening test as a baseline standard using the majority vote to determine how to classify each excerpt as positive or negative. The baseline itself depends on the exact mix of clear-cut and ambiguous excerpts. If all the excerpts were unanimously judged either positive or negative, the baseline failure rate would be 0%. On the other hand, if all the excerpts were judged positive by half of the listeners and negative by the other half, the failure rate would be 50%. Our set of 300 low-arousal piano excerpts is, of course, a mix. The baseline failure rate was 22% over all excerpts and listeners. This means that 22% of the individual listener judgments about positive and negative were different from the majority judgments averaged over all excerpts and listeners.

Next we calculated the failure rate for Calm and Sad, once again assuming that the majority vote for positive and negative perfectly defines the two low-arousal quadrants. The failure rate was 28% for Calm and Sad. This means that 28% of the individual listener judgments about Calm and Sad were different from the majority judgments for positive and negative, respectively, averaged over all excerpts and listeners. For a music emotion recognition system this means that asking listeners to judge each excerpt as Calm or Sad adds about 6% to the inaccuracy of distinguishing the two low-arousal quadrants compared to asking them to judge whether the mood of the excerpt is positive or negative.

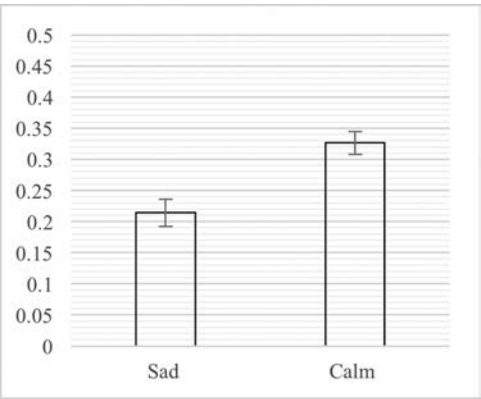


Fig. 16. The failure rates and the 95% confidence intervals for Calm and Sad.

We also considered subsets of the excerpts to see how the failure rate varied over different groups. There was not much difference between positive and negative excerpts (with failure rates of 21.4% for positive excerpts and 21.9% for negative excerpts). There was a larger variation between Calm and Sad excerpts (Calm and Sad as derived from our third and fourth listening tests), with a 33% failure rate for Calm excerpts and 21% for Sad excerpts (see Figure 16). This indicates that there were more negative Calm excerpts than positive Sad excerpts. There was an even larger variation between Calm, Sad, Both, and Neither excerpts with failure rates of 21% for Sad excerpts, 29% for Calm excerpts, 30% for Neither excerpts, and 35% for Both excerpts (using the reclassification of Both and Neither in our

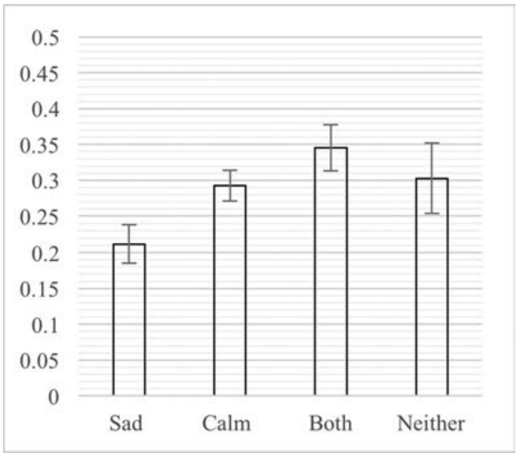


Fig. 17. The failure rates and the 95% confidence intervals for Calm, Sad, Both and Neither.

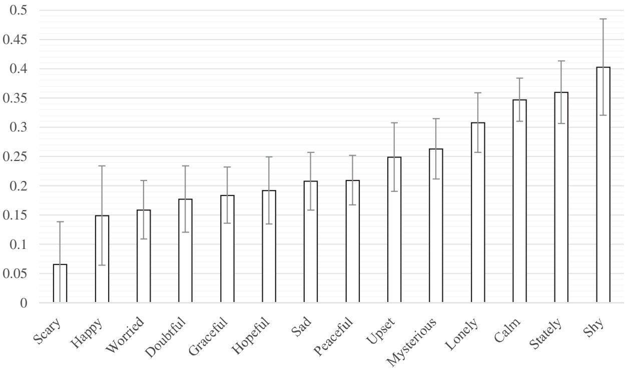


Fig. 18. The failure rates and the 95% confidence intervals for each of the 14 categories of excerpts in our second listening test, ranked lowest to highest.

fourth listening test to Calm and Sad). Figure 17 shows the failure rates and the 95% confidence intervals for each category (Calm, Sad, Both, Neither). It makes sense that Both excerpts would contain a larger percentage of cases where individual listener judgments were different from the majority, since the size of the majority is bound to be smaller for Both excerpts since they are less clear-cut and more ambiguous. Schmidt [6] also noted that music excerpts that were near the quadrant boundaries were more frequently misclassified.

Among the 14 categories in our second listening test, Figure 18 shows the failure rates and the 95% confidence intervals [97] for each category. Scary and Happy were the lowest and Shy was the highest at 40% (we took Mysterious and Shy as negative categories since these gave lower failure rates). These results make sense since Scary and Happy have near-zero failure cases in Figure 9, and Shy is the flattest and most evenly distributed in the same figure.

3 DISCUSSION

3.0 Overview

The main goal of our paper has been to investigate the distribution of the emotional categories Calm and Sad for low-arousal piano excerpts. Our main overall results based on the previous section are the following. Though the excerpts were nearly evenly divided in mood between positive and negative (49.3% positive and 50.7% negative), listeners judged a larger percentage Calm than Sad (59% Calm and 41% Sad). Moreover, while more numerous, the Calm excerpts were significantly less positive in mood than the Sad excerpts were negative. When listeners were allowed to choose between “More Calm than Sad,” “More Sad than Calm,” “Both Calm and Sad about equally,” “Not even a little Calm or Sad” (Neither), the results were 44% Calm, 29% Sad, 19% Both, and 8% Neither.

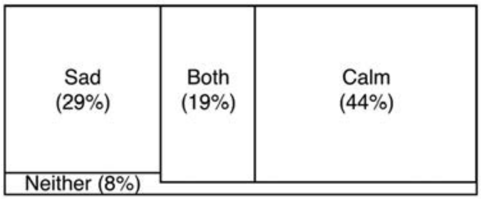


Fig. 19. A more detailed model of Calm and Sad based on the distribution of responses in our listening test.

This section discusses the coverage, overlap, and balance of the emotional categories Calm and Sad for low-arousal piano excerpts in more detail. The implications for 4-quadrant music emotion recognition systems and future work are also discussed.

3.1 Coverage of Calm and Sad

One of our main goals has been to determine the coverage of the emotional categories Calm and Sad compared to their respective low-arousal quadrants in a 4-quadrant music emotion recognition system. We also wanted to determine whether holes exist that are not covered by Calm and Sad and, if so, some idea of their extent.

Calm had the most extensive coverage with 44%, Sad was smaller at 29%, and together with Both they covered 92% of the excerpts. Only 8% of the excerpts were judged as “Not even a little Calm or Sad,” indicating some holes, but not too large. The largest holes were for very negative excerpts with about 3%. The other 5% was roughly equally distributed over negative, neutral, positive, and very positive excerpts. Table 4 indicates that for negative excerpts, the largest holes were for the categories Mysterious, Doubtful, Worried, and Scary. The largest holes for positive excerpts were for the categories Graceful and Stately. Figure 19 shows a graphical representation of the overall distributions and it shows the asymmetries compared to the 4-quadrant model in Figure 1.

Speculating a little, could another positive emotion category such as Peaceful have given a better distribution than Calm (i.e., covering the positive low-arousal quadrant fully with minimal overlap in the negative low-arousal quadrant)? Or, could another negative emotional category such as Worried have given better coverage than Sad (especially since Sad was only chosen for 29% of excerpts compared to Calm at 44%)? Table 2 and Figure 9 both suggest

"Probably not" as the answer to these questions. Calm and Sad were the top-ranked positive and negative categories respectively, indicating they include the largest percentage of positive and negative excerpts. Peaceful and Worried could provide slightly more coverage for very positive and very negative excerpts but at the expense of more neutral excerpts.

These results agree with previous findings [9, 20, 54], where researchers found that the valence value for Peaceful is higher than Calm. Hu [21] also found that Calm and Sad were the largest of 18 categories in the low-arousal quadrants.

3.2 Overlap of Calm and Sad

Though the holes in the coverage of Calm and Sad amounted to 8%, the overlap between them was much larger at 19%, where both Calm and Sad were present in about equal amounts. Predictably, Figure 11 indicates that Both was more frequently chosen for neutral excerpts than for positive or negative excerpts but not by much. Also predictably, Sec. 2.5 showed that listener judgments were more ambiguous and less clear-cut when listeners were forced to categorize both excerpts as either Calm or Sad, even more than Neither excerpts. This agrees with Schmidt's results [6].

Once again speculating, would there be less overlap between Peaceful and Sad? Based on the difference between the very negative, negative, and neutral classes for Calm and Peaceful in Figure 9, it seems like there would be less overlap, but it is hard to know for sure whether it would be a little or a lot without repeating the same types of listening tests for Peaceful and Sad as we ran for Calm and Sad.

3.3 Balance of Calm and Sad

Since there was about an even balance of positive and negative excerpts in our tests, the distribution of 59% of the excerpts as Calm and 41% as Sad already indicates an imbalance between the two categories. Figure 14 gives a more precise picture of the imbalance. It shows that neutral excerpts were judged Calm more often than Sad. It also shows that positive excerpts were judged Calm more often than negative excerpts were judged Sad by about 20%. As another indication of this imbalance, the Calm and Sad curves cross left-of-center between the negative and neutral classes rather than at the neutral class.

At the same time, the smaller number of Sad excerpts were significantly more negative in mood than the larger number of Calm excerpts were positive. Together they formed a weighted balance: the smaller number of more negative Sad excerpts about equally balanced the larger number of less positive Calm excerpts so that the average valence among all excerpts was about equally positive and negative.

These results agree with the results of Hu [21], where they found Calm to be relatively neutral in valence. Han [9] also assumed Calm was relatively neutral in valence. These results contrast with previous findings by Eerola [47], where they found that there was no correlation between valence and Sad.

In this light, could Peaceful provide a better balance to Sad than Calm? Certainly it is a more positive word. Table 2 indicates that the average valence value for Peaceful is about as positive as Sad is negative (likewise for Hopeful and Graceful) and that Calm is less positive. But, Figure 9 suggests that Peaceful is much more frequently very positive than Sad is very negative. A stronger word such as Depressed might be a better mirror image of Peaceful. Also, since Peaceful is so positive, it is more likely to leave holes in the neutral region than Calm (i.e., excerpts that are Calm, but neither Sad nor Peaceful). The tradeoffs between coverage, overlap, and balance are complex.

3.4 Implications for the 4-Quadrant Model and Future work

The 4-quadrant music emotion recognition model is very intuitive and presents users with four clear emotional categories such as Happy, Sad, Angry, and Calm. Yet, previous work has identified difficulties in distinguishing low-arousal categories such as Calm and Sad for listeners and automated systems [1–2, 14, 17, 19–20, 23–28, 32, 35–38, 44, 47–51]. In summary, what do our results tell us about this difficulty?

First, the emotional categories Calm and Sad leave about 8% of the low arousal space uncovered in holes that are neither Calm nor Sad. Second, Calm and Sad overlap equally in about 20% of the low-arousal space. Third, Calm is significantly less positive in mood than Sad is negative—they are not the well-balanced rectangles as they are usually represented in Figure 1 but more like the shapes in Figure 19.

Fourth, the Calm-Sad model results in 6% more errors than the positive-negative model due to holes in the coverage of Calm and Sad, ambiguities in their overlaps, and their asymmetries.

So where do we go from here with the 4-quadrant model? It depends on the application. If accuracy is the main concern, we can break our single 4-category decision into two binary decisions and ask listeners:

"Is the mood of the music positive or negative?"
"Is the energy level of the music high or low?"

This provides a direct determination of the quadrant. On the other hand, accuracy isn't always everything, and the simple intuitive character of four categories such as Happy, Sad, Angry, and Calm might be more desirable even knowing that it will result in higher error rates.

Do we have any other alternatives? Sure, there are many. One option is to consider Peaceful and Depressed as a pair instead. Peaceful and Depressed have some potential for better balance and less overlap. On the other hand, the chance of gaps between Peaceful and Depressed seems larger. Future work can consider these tradeoffs.

More generally, it would be interesting to consider the various tradeoffs in coverage, overlap, and balance between the other pairs of categories in the 4-quadrant model (Happy and Sad, Happy and Angry, Happy and Calm, Angry and Sad, Angry and Calm). More generally, it would also be useful to know the coverage, overlap, and balance between other pairs of categories such as the 14 categories we considered in this paper (e.g., Doubtful and Hopeful, Hopeful and Worried, Doubtful and Shy, Shy and Happy, etc.).

Just as researchers have long-sought to chart the multidimensional timbre space of instruments [55–71], it would be fascinating to chart the space of emotional characteristics for different types of music and instruments. What are the shapes of these characteristics, how do they overlap, and what are their symmetries? How do they differ in different genres such as pop music ballads and orchestral music? Investigations into these aspects will shed light on some of the fundamental issues in automatic music emotion recognition and music emotion recommendation.

4 ACKNOWLEDGMENTS

Thanks very much to the anonymous reviewers for their very helpful comments and suggestions.

5 REFERENCES

[1] Y. Song and D. Simon, “How Well Can a Music Emotion Recognition System Predict the Emotional Responses of Participants?” *Sound and Music Computing Conference (SMC),* pp. 387–392 (2015).

[2] Y. Yi-Hsuan, L. Yu-Ching, S. Ya-Fan, and H. C. Homer, “A Regression Approach to Music Emotion Recognition,” *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 16, no. 2, pp. 448–457 (2015). https://doi.org/10.1109/TASL.2007.911513

[3] K. F. MacDorman and S. O. C.-C. Ho, “Automatic Emotion Prediction of Song Excerpts: Index Construction, Algorithm Design, and Empirical Comparison,” *J. New Music Res.,* vol. 36, no. 4, pp. 281–299 (2007). https://doi.org/10.1080/09298210801927846

[4] K. Markov and T. Matsui, “Speech and Music Emotion Recognition Using Gaussian Processes,” *Modern Methodology and Applications in Spatial-Temporal Modeling* (Springer Japan, 2015), pp. 63–85. https://doi.org/10.1007/978-4-431-55339-7_3

[5] K. Markov and T. Matsui, “Music Genre and Emotion Recognition Using Gaussian Processes,” *IEEE Access,* vol. 2, pp. 688–697 (2014). https://doi.org/10.1109/ACCESS.2014.2333095

[6] E. M. Schmidt, D. Turnbull, and Y. E. Kim, “Feature Selection for Content-Based, Time-Varying Musical Emotion Regression,” *Proceedings of the International Conference on Multimedia Information Retrieval,* pp. 267–274 (ACM, 2010). https://doi.org/10.1145/1743384.1743431

[7] B. Schuller, J. Dorfner, and G. Rigoll, “Determination of Non-Prototypical Valence and Arousal in Popular Music: Features and Performances,” *EURASIA J. Audio, Speech, and Music Processing*, vol. 1, pp. 1–19 (2010).

[8] A. Hanjalic, “Extracting Moods from Pictures and Sounds: Towards Truly Personalized TV,” *IEEE Signal Processing Magazine,* vol. 23, no. 2, pp. 90–100 (2006). https://doi.org/10.1109/MSP.2006.1621452

[9] B.-j. Han, S. Rho, and B. Roger E. H. Dannenberg, “SMERS: Music Emotion Recognition Using Support Vector Regression,” *10th International Society for Music Information Retrieval Conference,* pp. 651–656 (ISMIR 2009).

[10] C. Baume, G. Fazekas, M. Barthet, D. Marston, and M. Sandler, “Selection of Audio Features for Music Emotion Recognition Using Production Music,” presented at the *AES 53rd International Conference, Semantic Audio* (2014 Jan.), conference paper P1-3.

[11] R. Parke, E. Chew, and C. Kyriakakis, “Multiple Regression Modeling of the Emotional Content of Film and Music,” presented at the *123rd Convention of the Audio Engineering Society* (2007 Oct.), convention paper 7173.

[12] A. den Brinker, R. van Dither, and J. Skowronek, “Prediction of Valence and Arousal from Music Features,” presented at the *131st Convention of the Audio Engineering Society* (2011 Oct.), Engineering Brief 39.

[13] Y. Song, S. Dixon, and M. Pearce, “Evaluation of Musical Features for Emotion Classification,” *13th International Society for Music Information Retrieval Conference (ISMIR),* pp. 523–528 (2012).

[14] D. Su and P. Fung, “Personalized Music Emotion Classification via Active Learning,” *Proceedings of the Second International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies,* pp. 57–62 (2012). https://doi.org/10.1145/2390848.2390864

[15] Y. Feng, Y. Zhuang, and Y. Pan, “Popular Music Retrieval by Detecting Mood,” *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 375–376 (ACM, 2003). https://doi.org/10.1145/860435.860508

[16] J. Wagner, J. Kim, and E. André, “From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification,”*IEEE International Conference on Multimedia and Expo*, pp. 940–943 (2005). https://doi.org/10.1109/ICME.2005.1521579

[17] K. Bischoff, C. S. Firan, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo, “Music Mood and Theme Classification—A Hybrid Approach,” *10th International Society for Music Information Retrieval Conference (ISMIR),* pp. 657–662 (2009).

[18] S.-W. Bang, J. Kim, and J.-H. Lee, “An Approach of Genetic Programming for Music Emotion Classification,” *Int. J. Control, Automation and Systems,* vol. 11, no. 6, pp. 1290–1299 (2013). https://doi.org/10.1007/s12555-012-9407-7

[19] L. Mion and G. De Poli, “Score-Independent Audio Features for Description of Music Expression,” *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 16, no. 2, pp. 458–466 (2008). https://doi.org/10.1109/TASL.2007.913743

[20] Y.-H. Chin, C.-H. Lin, E. Siahaan, I.-C. Wang, and J.-C. Wang, “Music Emotion Classification Using Double-Layer Support Vector Machines,” *Orange Technologies*

(ICOT), 2013 International Conference on, IEEE, pp. 193–196 (2013).

[21] X. Hu, and J. S. Downie, "When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis," *11th International Society for Music Information Retrieval Conference (ISMIR),* pp. 619–624 (2010).

[22] C. Laurier, J. Grivolla, and P. Herrera, "Multimodal Music Mood Classification Using Audio and Lyrics," *Machine Learning and Applications, (ICMLA'08). Seventh International Conference on IEEE,* pp. 688–693 (2008). https://doi.org/10.1109/icmla.2008.96

[23] M. Naji, M. Firoozabadi, and P. Azadfallah, "Emotion Classification During Music Listening from Forehead Biosignals," *Signal, Image and Video Processing,* vol. 9, no. 6, pp. 1365–1375 (2015). https://doi.org/10.1007/s11760-013-0591-6

[24] S. Pouyanfar and H. Sameti, "Music Emotion Recognition Using Two Level Classification," *Intelligent Systems (ICIS), 2014 Iranian Conference on. IEEE,* pp. 1–6 (2014). https://doi.org/10.1109/iraniancis.2014.6802519

[25] Y.-J.n Hsu and C.-P. Chen, "Going Deep: Improving Music Emotion Recognition with Layers of Support Vector Machines," *Applied System Innovation: Proceedings of the 2015 International Conference on Applied System Innovation (ICASI), May 22-27, 2015*, Osaka, Japan, pp. 209–212 (CRC Press, 2015). https://doi.org/10.1201/b21811-46

[26] X. Hao, L. Xue, and F. Su, "Multimodal Music Mood Classification by Fusion of Audio and Lyrics," *International Conference on Multimedia Modeling*, pp. 26–37 (Springer International Publishing, 2015). https://doi.org/10.1007/978-3-319-14442-9_3

[27] J.-M. Ren, M.-J. Wu, and J.-S. R. Jang, "Automatic Music Mood Classification Based on Timbre and Modulation Features," *IEEE Transactions on Affective Computing,*vol. 6, no. 3, pp. 236–246 (2015). https://doi.org/10.1109/TAFFC.2015.2427836

[28] L. Lu, D. Liu, and H.-J. Zhang, "Automatic Mood Detection and Tracking of Music Audio Signals," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, no. 1, pp. 5–18 (2006). https://doi.org/10.1109/TSA.2005.860344

[29] M. B. Mokhsin, N. B. Rosli, W. A. W. Adnan, and N. A. Manaf, "Automatic Music Emotion Classification Using Artificial Neural Network Based on Vocal and Instrumental Sound Timbres," *New Trends in Software Methodologies, Tools, and Techniques*, pp. 3–14 (2014).

[30] J. Kim and L. Larsen, "Music Emotion and Genre Recognition Toward New Affective Music Taxonomy," presented at the *128th Convention of the Audio Engineering Society* (2010 May), convention paper 8018.

[31] J. Skowronek, M. McKinney, and S. Van De Par, "A Demonstrator for Automatic Music Mood Estimation," *8th International Conference on Music Information Retrieval (ISMIR),* pp. 345–346 (2007).

[32] S.-H. Chen, Y.-S. Lee, W.-C. Hsieh, and J.-C. Wang, "Music Emotion Recognition Using Deep Gaussian Process," *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA),* pp. 495–498 (2015). https://doi.org/10.1109/apsipa.2015.7415321

[33] M. Plewa and B. Kostek, "A Study on Correlation between Tempo and Mood of Music," presented at the *133rd Convention of the Audio Engineering Society* (2012 Oct.), convention paper 8800.

[34] K. Hevner, "Experimental Studies of the Elements of Expression in Music," *Amer. J. Psych.,* vol. 48, no. 2, pp. 246–268 (1936). https://doi.org/10.2307/1415746

[35] E. E. P. Mint, and M. Pwint, "An Approach for Multi-Label Music Mood Classification," *Signal Processing Systems (ICSPS), 2010 2nd International Conference on. Vol. 1. IEEE,* pp. 290–294 (2010).

[36] P. Dunker, S. Nowak, A. Begau, and C. Lanz, "Content-Based Mood Classification for Photos and Music: A Generic Multi-Modal Classification Framework and Evaluation Approach," *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, pp. 97–104 (2008). https://doi.org/10.1145/1460096.1460114

[37] Y.-H. Yang, C.-C. Liu, and H. H. Chen, "Music Emotion Classification: A Fuzzy Approach," *Proceedings of the 14th ACM International Conference on Multimedia*, pp. 81–84 (ACM: 2006). https://doi.org/10.1145/1180639.1180665

[38] Y. Hu, X. Chen, and D. Yang, "Lyric-Based Song Emotion Detection with Affective Lexicon and Fuzzy Clustering Method," *10th International Society for Music Information Retrieval Conference (ISMIR),* pp. 123–128 (2009).

[39] T.-L. Wu, and S.-K. Jeng, "Automatic Emotion Classification of Musical Segments," *Proceedings of the 9th International Conference on Music Perception & Cognition*, Bologna, pp. 385–393 (2006)

[40] M. Barthet, G. Fazes, and M. Sandler "Multidisciplinary Perspectives on Music Emotion Recognition: Implications for Content and Context-Based Models," *9th International Symposium on Computer Music Multidisciplinary Research (CMMR),* pp. 492–507 (2012).

[41] J. Liebetrau and S. Schneider, "Music and Emotions: A Comparison of Measurement Methods," presented at the *134th Convention of the Audio Engineering Society* (2013), convention paper 8875.

[42] C. Baume, "Evaluation of Acoustic Features for Music Emotion Recognition," presented at the *134th Convention of the Audio Engineering Society* (2013 May), convention paper 8811.

[43] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music Emotion Recognition: A State of the Art Review," *11th International Society for Music Information Retrieval Conference (ISMIR),* pp. 255–266 (2010).

[44] R. Panda and R. P. Paiva, "Using Support Vector Machines for Automatic Mood Tracking in Audio Music," presented at the *130th Convention of the Audio Engineering Society* (2011 May), convention paper 8378.

[45] J. A. Russell, "A Circumplex Model of Affect," *J. Personality & Social Psych.,* vol. 39., no. 6, pp. 1161–1178 (1980). https://doi.org/10.1037/h0077714

[46] M. Zentner, D. Grandjean, and K. R. Scherer, "Emotions Evoked by the Sound of Music: Characterization, Classification, and Measurement," *Emotion,* vol. 8, no. 4, pp. 494–521 (2008). https://doi.org/10.1037/1528-3542.8.4.494

[47] T. Eerola and J. K. Vuoskoski, "A Comparison of the Discrete and Dimensional Models of Emotion in Music," *Psychology of Music*, pp. 18–49 (2010).

[48] L.-L. Balkwill and W. F. Thompson, "A Cross-Cultural Investigation of the Perception of Emotion in Music: Psychophysical and Cultural Cues," *Music Perception: An Interdisciplinary J.,* vol. 17, no. 1, pp. 43–64 (1999). https://doi.org/10.2307/40285811

[49] J. Liebetrau, S. Schneider, and R. Jezierski, "Application of Free Choice Profiling for the Evaluation of Emotions Elicited by Music," *Proc. 9th Int. Symp. Comput. Music Modeling and Retrieval (CMMR): Music and Emotions*, pp. 78–93 (2012).

[50] E. Bigand, S. Vieillard, F. Madurel, J. 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,* vol. 19, no. 8, pp. 1113–1139 (2005). https://doi.org/10.1080/02699930500204250

[51] I. Lahdelma and T. Eerola, "Single Chords Convey Distinct Emotional Qualities to Both Naïve and Expert Listeners," *Psychology of Music,* vol. 44, no. 1, pp. 37–54 (2016). https://doi.org/10.1177/0305735614552006

[52] M. M. Bradley and P. J. Lang, "Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings," Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, pp. 1–45 (1999).

[53] R. B. Dietz and A. Lang, "Affective Agents: Effects of Agent Affect on Arousal, Attention, Liking and Learning," *Cognitive Technology Conference*, 61–72 (1999).

[54] P. J. Lang, "Behavioral Treatment and Bio-Behavioral Assessment: Computer Applications," Technology in Mental Health Care Delivery Systems, pp. 119–l37 (1980).

[55] A. Caclin, S. McAdams, B. K. Smith, and S. Winsberg, "Acoustic Correlates of Timbre Space Dimensions: A Confirmatory Study Using Synthetic Tones," *J. Acoust. Soc. Amer.,* vol. 118, no. 1, pp. 471–482 (2005). https://doi.org/10.1121/1.1929229

[56] A. Horner, J. Beauchamp, and R. So, "Detection of Random Alterations to Time-Varying Musical Instrument Spectra," *J. Acoust. Soc. Amer.,* vol. 116, no. 3, pp. 1800–1810 (2004). https://doi.org/10.1121/1.1778741

[57] J. W. Beauchamp and S. Lakatos, "New Spectro-Temporal Measures of Musical Instrument Sounds Used for a Study of Timbral Similarity of Rise-Time-and Centroid-Normalized Musical Sounds," *Proc. 7th Int. Conf. Music Percept. Cognition (ICMPC),* pp. 592–595 (2002).

[58] G. R. Charbonneau, "Timbre and the Perceptual Effects of Three Types of Data Reduction," *Computer Music J.,* vol. 5, no. 2, pp. 10–19 (1981). https://doi.org/10.2307/3679875

[59] J. M. Grey, "Multidimensional Perceptual Scaling of Musical Timbres," *J. Acoust. Soc. Amer.,* vol. 61, no. 5, pp. 1270–1277 (1977). https://doi.org/10.1121/1.381428

[60] J. M. Grey, and J. W. Gordon, "Perceptual Effects of Spectral Modifications on Musical Timbres," *J. Acoust. Soc. Amer.,* vol. 63, no. 5, pp. 1493–1500 (1978). https://doi.org/10.1121/1.381843

[61] J. M. Grey, and J. A. Moorer, "Perceptual Evaluations of Synthesized Musical Instrument Tones," *J. Acoust. Soc. Amer.,* vol. 62, no. 2, pp. 454–462 (1977). https://doi.org/10.1121/1.381508

[62] P. Iverson and C. L. Krumhansl, "Isolating the Dynamic Attributes of Musical Timbre," *J. Acoust. Soc. Amer.,* vol. 94, no. 5, pp. 2595–2603 (1993). https://doi.org/10.1121/1.407371

[63] R. A. Kendall and E. C. Carterette, "Difference Thresholds for Timbre Related to Spectral Centroid," *Proceedings of the 4th International Conference on Music Perception and Cognition (ICMPC)*, Montreal, Canada, pp. 91–95 (1996).

[64] J. Krimphoff, "*Analyse Acoustique et Perception du Timbre*," *unpublished DEA thesis*, Université du Maine, Le Mans, France (1993).

[65] Jo. Krimphoff, S. McAdams, and S. Winsberg, "Caractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique," *Le Journal de Physique IV*, C5, pp. 625–628 (1994). https://doi.org/10.1051/jp4:19945134

[66] C. L. Krumhansl, "Why Is Musical Timbre so Hard to Understand," *Structure and Perception of Electroacoustic Sound and Music*, pp. 43–53 (1989).

[67] S. Lakatos, "A Common Perceptual Space for Harmonic and Percussive Timbres," *Perception & Psychophysics,* vol. 62, no. 7, pp. 1426–1439 (2000). https://doi.org/10.3758/BF03212144

[68] S. McAdams, J. W. Beauchamp, and S. Meneguzzi, "Discrimination of Musical Instrument Sounds Resynthesized with Simplified Spectrotemporal Parameters," *J. Acoust. Soc. Amer.,* vol. 105, no. 2, pp. 882–897 (1999). https://doi.org/10.1121/1.426277

[69] R. Plomp, "Timbre as a Multidimensional Attribute of Complex Tones," *Frequency Analysis and Periodicity Detection in Hearing,* pp. 405–408 (1970).

[70] D. L. Wessel, "Timbre Space as a Musical Control Structure," *Computer Music J.,* pp. 45–52 (1979). https://doi.org/10.2307/3680283

[71] J. C. Hailstone, R. Omar, S. M. D. Henley, C. Frost, M. G. Keyword, and J. D. Warren, "It's Not What You Play, It's How You Play It: Timbre Affects Perception of Emotion in Music," *Quarterly J. Experimental Psych.,* vol. 62, no. 11, pp. 2141–2155 (2009). https://doi.org/10.1080/17470210902765957

[72] T Konstantinos and S Lui, "Modeling Affective Responses to Music Using Audio Signal Analysis and Physiology," *Music, Mind, and Embodiment*, pp. 346–357 (Springer, 2016).

[73] I. Edman and R. Kajastila, "Localization Cues Affect Emotional Judgments—Results from a User Study on Scary Sound," presented at the *AES 35th Int.*

Conference, Audio for Games (2009 Feb.), conference paper 23.

[74] B. Kostek and M. Plewa, "Parametrization and Correlation Analysis Applied to Music Mood Classification," *Int. J. Computational Intell. Studies*, vol. 2, no. 1, pp. 4–25 (2013). https://doi.org/10.1504/IJCISTUDIES.2013.054734

[75] G. Leslie, R. Picard, and S. Lui, "An EEG and Motion Capture Based Expressive Music Interface for Affective Neurofeedback," *The 1st International Workshop on Brain-Computer Interfacing*, Plymouth (2015).

[76] S. Lui, "Generate Expressive Music from Picture with a Handmade Multi-Touch Music Table," *The 15th International Conference on New Interfaces for Musical Expression (NIME)*, Baton Rouge, pp. 374–377 (2015).

[77] C. Lee and S. Lui, "Visualization of Time-Varying Joint Development of Pitch and Dynamics for Speech Emotion Recognition," *167th Meeting of the Acoustical Society of America (ASA)*, Providence, Rhode Island, pp. 2422–2422 (2014). https://doi.org/10.1121/1.4878044

[78] S Lui, "A Preliminary Analysis of the Continuous Axis Value of the Three-Dimensional PAD Speech Emotional State Model," *The 16th International Conference on Digital Audio Effects (DAFx)*, Maynooth, Ireland, Paper Number #6, (2013).

[79] M. Ron, R. H. Y. So, and A. Horner, "An Investigation into How Reverberation Effects the Space of Instrument Emotional Characteristics," *J. Audio Eng. Soc.*, vol. 64, pp. 988–1002 (2016 Dec.).

[80] C. J. Chau, M. Ron, and A. Horner, "The Emotional Characteristics of Piano Sounds with Different Pitch and Dynamics," *J. Audio Eng. Soc.*, vol. 64, pp. 918–932 (2016 Nov.). https://doi.org/10.17743/jaes.2016.0049

[81] M. Ron, C. G. Lam, L. Chung, and A. Horner, "The Effects of MP3 Compression on Perceived Emotional Characteristics in Musical Instruments," *J. Audio Eng. Soc.*, vol. 64, pp. 858–867 (2016 Nov.).

[82] M. Ron, W. Bin, and A. Horner., "The Effects of Reverberation on the Emotional Characteristics of Musical Instruments," *J. Audio Eng. Soc.*, vol. 63, pp. 966–979 (2015 Dec.). https://doi.org/10.17743/jaes.2015.0082

[83] C. J. Chau, W. Bin, and A. Horner, "The Emotional Characteristics and Timbre of Nonsustaining Instrument Sounds," *J. Audio Eng. Soc.*, vol. 63, pp. 228–244 (2015 Apr.). https://doi.org/10.17743/jaes.2015.0016

[84] W. Bin, L. Chung, A Horner, "The Correspondence of Music Emotion and Timbre in Sustained Musical Instrument Sounds," *J. Audio Eng. Soc.*, vol. 62, pp. 663–675 (2014 Oct.). https://doi.org/10.17743/jaes.2014.0037

[85] C. J. Chau and A. Horner, "The Emotional Characteristics of Mallet Percussion Instruments with Different Pitches and Mallet Hardness," *International Computer Music Conference (ICMC)*, Utrecht, Netherlands, pp. 401–404 (2016 Sep.).

[86] C. G. Lam, L. Chung, M. Ron, and A Horner, "The Effects of MP3 Compression on Emotional Characteristics," *International Computer Music Conference (ICMC)*, Utrecht, Netherlands, pp. 411–416 (2016 Sep.).

[87] M. Ron and A. Horner, "The Effects of Reverberation Time and Amount on the Emotional Characteristics," *International Computer Music Conference (ICMC)*, Utrecht, Netherlands, pp. 12–15 (2016 Sep.).

[88] S. Gilburt, C. J. Chau, and A. Horner, "The Effects of Pitch and Dynamics on the Emotional Characteristics of Bowed String Instruments," *International Computer Music Conference (ICMC)*, Utrecht, Netherlands, pp. 405–410 (2016 Sep.).

[89] C. J. Chau and A. Horner, "The Effects of Pitch and Dynamics on the Emotional Characteristics of Piano Sounds," *International Computer Music Conference (ICMC)*, Denton, Texas, pp. 372–375 (2015 Sep.).

[90.] C. J. Chau, B. Wu, and A. Horner, "The Effects of Early-Release on Emotion Characteristics and Timbre in Non-Sustaining Musical Instrument Tones," *International Computer Music Conference (ICMC)*, Denton, Texas, pp. 138–141 (2015 Sep.).

[91] B. Wu, A. Horner, and C. Lee, "Emotional Predisposition of Musical Instrument Timbres with Static Spectra," *International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan, pp. 253–258 (2014 Nov.).

[92] B. Wu, E. Zhong, A. Horner, and Q. Yang, "Music Emotion Recognition by Multi-Label Multi-Layer Multi-Instance Multi-View Learning," *ACM International Conference on Multimedia (ACM MM)*, Orlando, FL, pp. 117–126 (2014 Nov.).

[93] B. Wu, A. Horner, and C. Lee, "Musical Timbre and Emotion: The Identification of Salient Timbral Features in Sustained Musical Instrument Tones Equalized in Attack Time and Spectral Centroid," *International Computer Music Conference (ICMC)*, Athens, Greece, pp. 928–934 (2014 Sep.).

[94] C. J. Chau, B. Wu, and A. Horner, "Timbre Features and Music Emotion in Plucked String, Mallet Percussion, and Keyboard Tones," *International Computer Music Conference (ICMC)*, Athens, Greece, pp. 982–989 (2014 Sep.).

[95] C. Lee, A. Horner, J. Beauchamp, and L. Ayers, "Discrimination of Sustained Musical Instrument Tones Resynthesized with Piecewise-Linear Approximation of Harmonic Amplitude Envelopes," *International Computer Music Conference (ICMC)*, Perth Australia, pp. 100–107 (2013 Aug.).

[96] B. Wu, E. Zhong, D. H. Hu, A. Horner, and Q Yang, "SMART: Semi-Supervised Music Emotion Recognition with Social Tagging," *SIAM International Conference on Data Mining (SDM)*, Austin, pp. 279–287 (2013 May).

[97] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator," *The Annals of Mathematical Statistics,* vol. 27, no. 3, pp. 642–669 (1956). https://doi.org/10.1214/aoms/1177728174

[98] T. W Anderson, "Confidence Limits for the Value of an Arbitrary Bounded Random Variable with a Continuous Distribution Function," *Bulletin of the International and Statistical Institute,* vol. 43, pp. 249–251 (1969).

APPENDIX

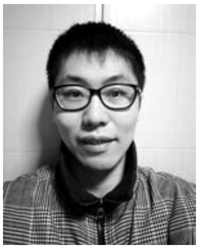
Table A1. The 100 pieces/movements we selected for our listening tests.

Composer	Piece/movement
Bach	Italienisches Konzert in F
	The Well-Tempered Clavier, Book 1: Fugue in C Sharp Minor
	Prelude in E Flat Minor, Prelude in B Flat Minor
	The Well-Tempered Clavier, Book 2: Fugue in E
Beethoven	Piano Sonata #3, #4, #5, #7, #8, #14, #21, #23, #26, #28, #29, #30, #32
Brahms	Ballade #2 in D
	Piano Sonata #3 in F Minor
	Fantasies, Op. 116, No. 4
	Intermezzo in E Flat, Op. 117
	6 Piano Pieces, Op. 118, No. 5, No. 6
	4 Piano Pieces, Op. 119, No. 1
Chopin	Nocturne #2, #8, #11, #13, #17, #20
	Ballade #1, #2
	Fantasy in F Minor
	Waltz #3 In A Minor
Debussy	Images Oubliées - Lent
	Suite Bergamasque - Clair De Lune
	Nocturne, L 82
	Pour Le Piano, L 95 - Sarabande
	Images #1, #2
	Preludes, Book 1, La Cathédrale Engloutie
	Épigraphes Antiques
	La boite a joujoux
Dvorak	In The Old Castle
	Reverie
Grieg	Melancholic, No. 5
	Bell Zringing
	Home Sickness
	Phantom
	Lyric Pieces, Book 8, Ballad
	Lyric Pieces, Book 10, Peace Of The Woods
Liszt	Ballade #2 in B Minor
	Legendes, St. Francois De Paule Marchant Sur Les Flots
	Präludium Nach Johann Sebastian Bach, S 179
	Variationen Über das Motiv von Bach, S 180
	La Notte
	Grosses Konzertsolo
	Weihnachtsbaum
	Via Crucis, Station #12, #13, #14
	Offertorium Aus der Ungarischen Kröningsmesse
	Nuages Gris
	Wiegenlied (Chant Du Berceau)
	Fünf Klavierstücke, No. 1
	Am Grabe Richard Wagners
	Mosonyis Grabgeleit
	À La Chapelle Sixtine
Mozart	Piano Sonata #6, #13, #15, #18
	Fantasy in C Minor
Poulenc	Les Soirées De Nazelles, No. 5, No. 11
	Pieces, No. 1 (Pastorale)
	Theme Varie
Rachmaninov	Preludes, No. 10
	Variations On A Theme Of Corelli
Ravel	Miroirs - Oiseaux Tristes
	Pavane Pour Une Infante Défunte
	Gaspard De La Nuit - Le Gibet
	Prelude
Satie	Gnossiennes
	Chapitres Tournés En Tous Sens
	Gymnopédies
	Embryons Desséchés

Table A1. Continued.

Composer	Piece/movement
Schubert	Piano Sonata in B-flat major
	Piano Sonata in C Minor
Schumann	Kinderszenen - Traumerei, Der Dichter Spricht
	Kreisleriana, No. 4, No. 6
	Fantasy in C, No. 1, No. 3
Shostakovich	24 Preludes & Fugues: Fugue in E Minor, Fugue in F Sharp
	Prelude in F

THE AUTHORS


Yu Hong

Chuck-jee

Andrew Horner

Yu Hong is a Ph.D. student in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology (HKUST). His research focuses on music emotion recognition. He obtained his B.Eng. in software engineering from Tsinghua University (THU).

Chuck-jee Chau is a Ph.D. student in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology (HKUST). His research focuses on timbre analysis and music emotion. During his master studies he developed the timbre visualization tool pvan+ for phase vocoder analysis. He obtained his B.Eng. in computer engineering from the Chinese University of Hong Kong (CUHK) with a minor in music. Besides computer music research, he is also a versatile collaborative pianist and mallet percussionist active in chamber music performances.

Andrew Horner is a professor in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology. His research interests include music analysis and synthesis, timbre of musical instruments, and music emotion. He received his Ph.D. in computer science from the University of Illinois at Urbana-Champaign.