What really matters? A study into people's instinctive evaluation metrics for continuous emotion prediction in music

Vaiva Imbrasaitė
Computer Laboratory
University of Cambridge
Vaiva.Imbrasaite@cl.cam.ac.uk

Tadas Baltrušaitis
Computer Laboratory
University of Cambridge
Tadas.Baltrusaitis@cl.cam.ac.uk

Peter Robinson
Computer Laboratory
University of Cambridge
Peter.Robinson@cl.cam.ac.uk

***Abstract*—Continuous emotion prediction in the arousal-valence space is now being used in various modalities: music, facial expressions, gestures, text, etc. In order to be able to compare the work of different research groups effectively, we believe it is necessary to set certain guidelines for how to conduct research—the choice of evaluation metrics of emotion recognition algorithms in particular. In this paper we focus on the field of musical emotion recognition and describe a study designed to discover people's instinctive preference among the most commonly used evaluation techniques. We gather strong evidence that root mean squared error or Kullback-Leibler divergence should be used for regression based approaches. The raw study data we collected is made publicly available.**

I. Introduction

Continuous emotion representation in the arousal-valence space is widely used in affective computing. Unfortunately, as is often the case with new disciplines, so far there is a noticeable lack of agreed guidelines for conducting experiments and evaluating algorithms.

Moreover, many datasets are used, which have been collected in distinct ways and for various purposes. As researchers make different choices, comparing their work is difficult.

Even though the same problem exists in most if not all subfields of affective computing dealing with continuous emotion representation, in this paper we focus on music. We describe a study we conducted in order to find out how people perceive the "goodness" of different evaluation metrics, and which of them most closely match people's own instincts. The results allow us to suggest several guidelines with confidence—most importantly, which evaluation metric to choose for optimizing emotion prediction algorithms.

II. Background

A. Emotion in music

Even though the majority of work in the field of emotion recognition in music is done on emotion classification, there is already a significant body of research done on continuous emotion prediction in the dimensional space (using regression).

By far the most common choice for axes for the dimensional representation of musical emotion is arousal (describing how active/passive emotion is) and valence (how positive/negative emotion is). It has been repeatedly shown that adding a third axis (e.g. dominance, tension, etc.) gives little or no benefit to a model [1], it has also been reported that the participants find the addition of a third axis confusing or difficult to deal with [2].

Within the research on dimensional musical emotion prediction, there is a wide range of evaluation metrics used. Starting with standard metrics such as root mean squared error and correlation [3], but also including Kullback-Leibler divergence [4], [5], average Euclidian distance [6], [7], Earth mover's distance [7].

B. Other areas

1) Emotion recognition from audio/visual clues: The idea to model emotion in terms of several latent dimensions is not exclusive to music. Such representation of affect is used when modeling external expressions of emotions such as emotional speech, facial expressions, head gestures, and body posture.

When the problem is formed in continuous space (such as Audio/Visual Emotion Challenge 2012 [8] and 2013[1]) metrics such as average RMSE [9], [10], correlation [10] and sign-agreement [10] per sequence are used. The metrics are usually reported per dimension (separate scores for valence, arousal etc.). Unfortunately, many authors fail to make it clear whether the evaluation metrics they report are averaged across sequences or computed from a single concatenated sequence (as is more common in music community), making it more difficult to compare different work.

2) Emotion recognition from physiological clues: Emotion recognition based on the analysis of physiological measurements could provide a way of evaluating the felt emotion, as opposed to the expressed emotion (or the mixture of both). There is a variety of measurements that such a system could be based on: EKG, skin conductivity, heart-rate variability, EEG, etc. Classifiers instead of regressors are often used, with accuracy as the evaluation metric [11]. Even when regressors are used initially, the final outcome is commonly converted to a class by using a set of bins for the labels and accuracy as the evaluation metric [12].

[1] http://sspnet.eu/avec2013/

978-0-7695-5048-0/13 $26.00 © 2013 IEEE
DOI 10.1109/ACII.2013.106

IEEE computer society

3) Sentiment analysis in text: In the field of sentiment analysis in text, the majority of work tends to focus on the valence axis only [13]. Even though the task of inferring how positive a piece of text is would lend itself naturally to regression, it is often approached as or converted to a classification problem (binary or ordinal) [14] [15]. In the case of classification, accuracy is used as the evaluation metric, sometimes with the addition of root mean squared error [16]. For tasks defined as regression, correlation is used [17][18].

III. Method

Sections III-A to III-E explain the different evaluation metrics considered and the reasons for choosing a particular set of them. Section III-G describes the design of our study.

A. Metrics considered

1) One dimensional case: The simplest approach is to consider each affective attribute as a separate dimension. As seen in the previous work section (Section II) there are a multitude of metrics used to evaluate the machine learning algorithms for the task of dimensional emotion prediction. If we consider a sequence of length n with a ground truth $g(x)$ and prediction $p(x)$ per time-step x, we can define the most common metrics used in the field.

Average Euclidean distance:

$$E_{\mathrm{Eucl}}(g,p) = \frac{1}{n}\sum_{i=1}^{n} ||g(i) - p(i)|| \tag{1}$$

Root mean square error (RMSE), here defined for both single and multi-dimensional cases:

$$E_{\mathrm{RMSE}}(g,p) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{d=1}^{\mathrm{dims}}(g_{\mathrm{d}}(i) - p_{\mathrm{d}}(i))^2} \tag{2}$$

Pearson correlation coefficient:

$$E_{\mathrm{Corr}}(g,p) = \frac{\sum_{i=1}^{n}[(g(i)-\overline{g})(p(i)-\overline{p})]}{\sqrt{\sum_{i=1}^{n}(g(i)-\overline{g})^2}\sqrt{\sum_{i=1}^{n}(p(i)-\overline{p})^2}}, \tag{3}$$

where $\overline{g}$ and $\overline{p}$ are the mean ground truth and predictor values for the sequence of interest. Some authors use squared correlation coefficients instead of non-squared ones, we choose not to do so. Squaring the correlation coefficient, can hide the fact that the predictions are inversely correlated with ground truth, which is not a desired behaviour of a predictor.

We use the definition of the average sign agreement (SAGR) from Gunes *et al.* [10]:

$$E_{\mathrm{SAGR}}(g,p) = \frac{1}{n}\sum_{i=1}^{n} s(g(i),p(i)), \tag{4}$$

$$s(x,y) = \begin{cases} 1, & \mathrm{sign}(x) = \mathrm{sign}(y) \\ 0, & \mathrm{sign}(y) \neq \mathrm{sign}(y) \end{cases} \tag{5}$$

The Kullback-Leibler divergence (KL) is defined in the following section in Equation (7). In one dimensional case we would use scalars instead of vectors, and the covariance matrix just becomes the variance.

2) Two dimensional case: A stronger approach to two-dimensional models is to consider two affective attributes (typically valence and arousal) simultaneously. We have two predictors p_1 and p_2 (or a single non-correlated predictor for both dimensions $\mathbf{p}$), we also have the ground truth for both dimensions as well g_1 and g_2.

Average Euclidean distance is defined in Equation (1) and RMSE in Equation (2).

Average correlation across dimensions:

$$E_{\mathrm{Corr}}(g,p) = \frac{1}{2}(E_{\mathrm{Corr}}(g_1,p_1) + E_{\mathrm{Corr}}(g_2,p_2)) \tag{6}$$

Average KL-divergence for Normal distributions is a metric that measures the difference between two probability distributions, and is often suitable for the task at hand.

$$E_{\mathrm{Mean-KL}}(\mathbf{g},\mathbf{p}) = \frac{1}{n}\sum_{i=1}^{n} E_{\mathrm{KL}}(\mathbf{p}(i),\mathbf{g}(i),\Sigma_{p(i)},\Sigma_{g(i)}) \tag{7}$$

The predictor could provide an estimate together with uncertainty (Σ_p) and our ground truth can be modeled as a Normal distribution as well (centered on mean with Σ_g calculated from the labels from multiple people).

$$E_{\mathrm{KL}}(\mathbf{g},\mathbf{p},\Sigma_p,\Sigma_g) = \frac{1}{2}(\Sigma_g^{-1}\Sigma_p + (\mathbf{p}-\mathbf{g})\Sigma_g^{-1}(\mathbf{p}-\mathbf{g}) - d - \ln(\Sigma_g^{-1}\Sigma_p)) \tag{8}$$

Above Σ_p is a diagonal matrix as we assume our predictor is uncorrelated, Σ_g is a per time step covariance derived from labels given for that timestep by multiple people; d is the number of dimensions considered.

Finally we define a combined version of sign agreement:

$$E_{\mathrm{Sign\ agr}} = E_{\mathrm{Sign\ agr}}(p_1,g_1) + E_{\mathrm{Sign\ agr}}(p_2,g_2) \tag{9}$$

B. Defining a sequence

All of the above metrics except for the correlation coefficient are performed on a per time-step basis, and are then averaged across the whole sequence. Correlation coefficient relies on the mean value of the sequence as well (in calculating $\overline{p}$ and $\overline{g}$), so it becomes important how such a sequence is defined. In the Audio/Visual emotion recognition community the sequence is defined as a recording (or a part of a recording). A correlation score is then calculated for each of the recordings (short correlation). This is averaged across all of the sequences to provide a final evaluation metric. In the music community, however, it is more common to concatenate all of the individual songs into one long sequence and then compute the correlation (long correlation).

At first glance, whether short correlation or long correlation is used does not seem to have much of an impact. However, computing long correlation score might hide bad per sequence predictions. For example, a predictor that is good at predicting the average position in valence space for a song can still get a high correlation score, even though it is very bad at predicting change within a sequence (which is particularly interests us).

It is very important that research workers in the field make clear which of the averaging methods are being used,

especially if they are using correlation coefficients. But we advise against using overall correlation in general because of its tendency to hide information, especially if researchers are interested in capturing changes within a sequence.

C. Generating predictions

In order to evaluate how well a certain metric represents people's perception of emotion in music, we needed to be able to present the participants of our study with several different emotional traces that optimise a particular metric. We chose to use a hypothetical predictor that always predicts the trace as centered around the ground truth but with added Gaussian noise (the standard deviation matching that of human labelers of this dataset). We justify this amount of noise as we expect a statistical approach to perform within the boundaries of human variation. An example of such noisy trace can be seen in Figures 1, 3, and 5.

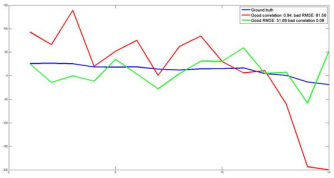


Fig. 1: Sample synthetic traces. Blue is ground truth, Red has a great correlation score, and green has a low RMSE and vice versa

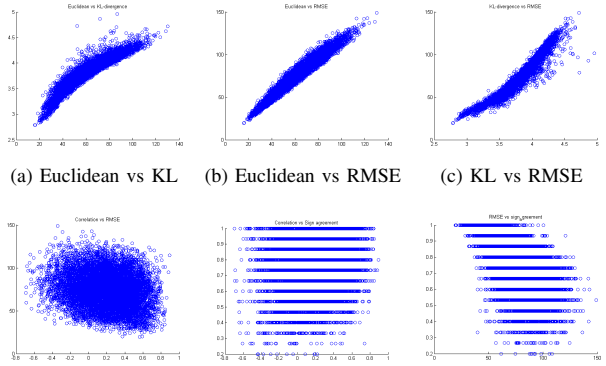
D. Optimising one metric over another

Once we generate a sufficient number of noisy predictions using our hypothetical predictor (we used 10^5 predictions) we can choose a prediction for a sequence that has the best score in a metric of interest when compared to the ground truth. So for example from the 10^5 generated noisy sequences pick one that has the best correlation coefficient with the ground truth and use that for the further experiment. We do this for every metric we are interested in. This allows us to pick traces which best represent a certain metric.

For the metrics we have chosen for our experiment (correlation, RMSE (1D), sign agreement, and KL-divergence (2D)) a sequence of predictions that optimised one metric never happened to be the one with the best score in another, hence just by generating noisy data we were able to pick predictions that have very different scores for different metrics. For example: in Figure 1 both of the predicted traces have been generated by adding the same type of noise, however resulting in two very different traces with very different metric scores.

E. Choosing evaluation metrics for our experiment

Ideally, we would use all of the five previously defined metrics for our experiment. We chose not to do this for two reasons. Firstly, this would have made the task more difficult for our participants (apparent from the pilot study, Section III-G1). Secondly, some of the metrics are redundant in the presence of others. This is particularly true for average Euclidean distance, RMSE, and average KL-divergence which are very similar (that is a prediction with a small average Euclidean distance will have small RMSE and small average KL-divergence)—see Figure 2. Thus, an approach which optimises any of these metrics would produce very similar results.



(a) Euclidean vs KL (b) Euclidean vs RMSE (c) KL vs RMSE

(d) Corr vs RMSE (e) Corr vs sagr (f) RMSE vs sagr

Fig. 2: Scatter plots of relationships between metrics when comparing a noisy synthetic prediction with ground truth. Notice how Euclidean, KL-divergence and RMSE are related.

Correlation and sign agreement metrics differ markedly from the other three and each other. Correlation is particularly distinct, and does not seem to be related to the RMSE and sign agreement metrics at all (this is partly because of several songs being somewhat stable over periods of 15 seconds, and correlation is not suited for evaluating stable sequences). Hence, we include the most widely differing and the most popular metrics in our further analysis.

F. Dataset

The dataset that we have based our emotion traces on is, to our knowledge, the only publicly available emotion tracking dataset of music extracts labeled on the arousal-valence dimensional space. The data [19] has been collected using Mechanical Turk (MTurk)[2], asking paid participants to label 15-second long excerpts with continuous emotion ratings on the AV space, with another 15 seconds given as a practice for each song. The songs in the dataset cover a wide range of genres—pop, various types of rock, hip-hop/rap, etc, and are drawn from the "uspop2002"[3] database containing Western popular songs. The dataset consists of 240 15-second clips (without the practice run) with 16.9 ± 2.7 ratings for each clip, where each clip has been randomly chosen within a song with no particular focus on a change in emotion.

G. Experimental design

There were several different questions that we wanted to answer. First of all, we wanted to see whether people differentiate between or have preference for a particular way of optimizing (or evaluating) emotion traces. If so, we were interested to see if the preferred evaluation technique depended on a choice of a song. We were also interested to see if the

[2]http://mturk.com

[3]http://labrosa.ee.columbia.edu/projects/musicsim/ uspop2002.html

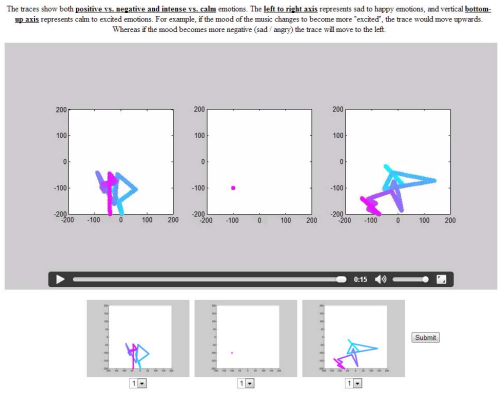


Fig. 3: Screenshot of the study page. Instruction at the top, followed by a video and the static emotion traces.

preferred evaluation metric depended on the axis (arousal or valence) or the number of dimensions (one or two).

To achieve this goal, we designed the following study. Each participant was presented with 56 15-second extracts from a subset of songs used in MTurk dataset (see Section III-F). For each song we had a video that displayed several emotion traces at the same time, synchronised with the audio extract (Figure 3). The participants were allowed to re-watch the video as many times as they wanted. Underneath the video, all the traces were presented in the static form (as they appear at the very end of the video) with a drop-down selection for ordering them. The participants were forced to give a unique ordering for the traces, i.e. they were not allowed to say that any two or all three traces were equally good.

Each trace for a song was based on a different evaluation metric - one optimized for correlation (best correlation, but higher RMSE, and lower sign agreement), one for sign agreement, RMSE and KL-divergence (see Section III-D for more details). The order in which those traces were presented was randomized for each song.

The songs were split into three groups, and therefore each participant was presented with three different tasks. In the first part of the experiment, we had 18 songs with a focus on the arousal axis. The songs were chosen with as much change in the arousal values and little change in the valence values, based on the labels in MTurk dataset. The participants had to order the arousal traces only (see Figure 5). The second group of 12 songs had the exact opposite properties—some change in the valence and little chance in the arousal values. The participants were presented with traces of affect on the valence axis. The third task was focused on the change in emotion on both axes. The 26 traces used in the last part were shown in 2D and were colour-coded to represent time (Figure 3).

The songs within each task were presented in a random order for each participant, but the order of the tasks remained the same. We hypothesized that one-dimensional emotion traces are easier to understand and deal with than two-dimensional ones and that arousal is easier to judge than valence. This way the participants would have time to practice on an easier task before moving on to a more difficult one.

1) Pilot study: To evaluate the suitability of our experimental design, we first ran a pilot study. We recruited two participants from our research group. They were not aware of the design of the study and its purpose.

As explained in the Section III-E, we first used 4 different evaluation metrics—correlation, RMSE, sign agreement and KL-divergence. The design of the experiment followed the description above.

Both participants did the study individually. They were provided with a pair of headphones each and did the study in their own time. The instructions were given on the screen. The experiment lasted approximately 30 mins for each participant.

The comments we received after the study confirmed that the task of evaluating 2D emotion traces was more difficult than 1D. The results also confirmed the appropriateness of our experimental design—there was a clear difference between the average rank for each of the evaluation metric.

2) Changes in the final study: After the success of our pilot study, we conducted the actual experiment with several changes—all based on the comments we received.

In the pilot study we found that even with only two participants, it was already clear that KL-divergence and RMSE achieve the same average rank—both per participant and overall. This, together with the theory described in Section III-E and the comments from the participants that it was often difficult to order 4 different traces, led us to decide to remove one of them. As RMSE is generally used for models dealing with one axis at a time, we kept RMSE as the third evaluation metric for the first two (one-dimensional) tasks. Similar reasoning led us to remove RMSE and keep KL-divergence for the third, two-dimensional, task.

We also made several changes to the instructions provided at the beginning of the study, making them a bit more informative and clear. In addition, we provided the participants with a sheet explaining the meaning of arousal and valence axes, as a reference throughout the experiment.

We had 20 participants (13 female and 7 male), recruited through a local ad-website and the graduate-student newsletters. Each participant was required to come to our lab for the study and received a £10 Amazon voucher for their time. We had up to 5 participants doing the study at the same time, all in the same room, each provided with a pair of headphones and doing the study in their own time. All of the instructions were given on the screen, and apart from 2 participants, none of them required extra explanations.

IV. Results

For the purpose of this study, we use the rankings for each song and each metric as numerical values ranging from one to three—1 being the most and 3 being the least preferred choice. This allows us to compute average ranking for each metric for each song, participant, or task. It also allows us to compare the means and check if any differences are statistically significant.

We split the analysis into two parts—one dimensional tasks, and the two dimensional task. There are two reasons for this. Firstly, since we have used slightly different metrics for the two types of tasks, it was impossible to combine all of

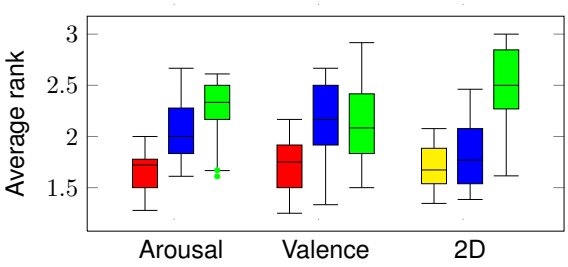


Fig. 4: Box-and-whiskers diagrams for the three tasks.
RMSE Correlation Sign-agreement KL-divergence

the data we had into one analysis. Secondly, we expected similar results/conclusions from the two one-dimensional tasks, while we expected the results to possibly differ between one-dimensional and two-dimensional tasks.

A. One-dimensional tasks

There are several questions we wanted to answer when looking at the data from the one-dimensional tasks. First of all, we wanted to check if there is any effect of the dimension on the average rank. Then within each dimension we want to check if the ranks are significantly different from each other, and if so, which one of them is preferred.

1) Normality: In order to answer these questions, we needed to check that our data is normally distributed, as many statistical tests require this. We calculated the average rank for each metric and each dimension per participant, i.e. we computed a 20x6 table (20 participants, 2 dimensions, 3 metrics) of mean ranks.

All but one (sign-agreement for arousal) of the distributions are approximately normally distributed. This is confirmed by Kolmogorov-Smirnov test—there is a statistically significant difference between the sign-agreement data for arousal and the normal distribution ($D(20) = 0.21, p = 0.023$). On the other hand, there is no statistical difference between the normal distribution and any other datasets ($D(20) = 0.19, p > 0.05$ for RMSE for arousal and $D(20) = 0.12, p > 0.05$ for valence axes, $D(20) = 0.15, p > 0.05$ for correlation for arousal and $D(20) = 0.12, p > 0.05$ for valence axes, and $D(20) = 0.14, p > 0.05$ for sign-agreement for valence axis).

When the data is aggregated over the two dimensions (giving 20x3 values), all three distributions show no statistically significant difference from the normal distribution.

2) ANOVA: A repeated measures within-subject factorial ANOVA with dimensions (2 levels) and metrics (3 levels) as factors show a small significant effect of dimension on the average rank ($F(1, 19) = 5.5, p = 0.030$). The effect of metrics, on the other hand, is much stronger ($F(2, 38) = 16.39, p < 0.001$), with no interaction between the two ($F(2, 38) = 1.785, p > 0.05$).

The pairwise comparison (with Bonferroni adjustment for multiple comparisons) reveals that there is a statistically significant difference between the average ranks for RMSE and correlation ($t(19) = -5.39, p < 0.001$), and RMSE and sign-agreement ($t(19) = -5.68, p < 0.001$), but no significant difference between correlation and sign-agreement ($t(19) = 0.75, p > 0.05$). The same conclusion can be observed in the box-and-whisker plot showing all 6 distributions (Figure 4).

B. Two-dimensional task

The questions we want to answer when looking at the two-dimensional task are the same as the ones from one-dimensional tasks. Mainly we are interested in seeing if there is a statistically significant difference between the average ranks of the different metrics. And if so, which is the preferred one.

1) Normality: Again, we first check if our data is normally distributed. We aggregate data in the same way as for the one-dimensional tasks—average the rank for each metric for each participant. This time all three datasets are normally distributed—the Kolmogorov-Smirinov test showed no statistically significant difference between the three sets and the Normal distribution ($D(20) = 0.09, p > 0.05$ for correlation, $D(20) = 0.13, p > 0.05$ for sign-agreement and $D(20) = 0.12, p > 0.05$ for KL-divergence).

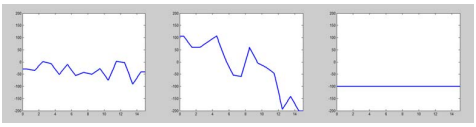
2) ANOVA: One-way repeated-measures ANOVA with metrics (3 levels) as factors show that there is a strong statistically significant effect of metrics on the average rank ($F(2, 38) = 28.55, p < 0.001$). Pairwise comparison between the three metrics (with Bonferroni adjustment for multiple comparisons) reveal that the average rank for sign-agreement is statistically significantly different from correlation ($t(19) = 4.89, p < 0.001$) and KL-divergence ($t(19) = 6.60, p < 0.001$). However, there is no statistically significant difference between the average ranks of KL-divergence and correlation ($t(19) = 1.29, p > 0.05$). This can also be observed in a visual inspection of the box-and-whisker plot (Figure 4).

C. Further analysis

As explained in Section III-B, there is a notable difference between short and long correlation. As a post-hoc analysis, we looked at the long correlation of the traces from people's top choices for each song, comparing it to the long correlation of the traces from each evaluation metric. For arousal, the correlation of the top choice reached 0.87, while RMSE optimized traces had correlation of 0.93. The lowest one was from short correlation optimized traces (0.77), with even sign-agreement scoring higher (0.82). Similar results are seen for valence (top choice–0.80, short correlation–0.72, RMSE–0.89 and sign-agreement–0.89).

We also wanted to consider whether or not the preferred choice of evaluation metric might depend on a song in question. To investigate this question, we take the average rank for each metric over each song, rather than over the participants. We then inspect the results to see if there are any exceptions.

Even though the majority of songs seem to follow the trends described in Sections IV-A and IV-B, there are some examples of songs with a different preference for evaluation metric. Occasionally, participants were choosing sign-agreement over the other two metrics. As can be seen from an example in Figure 5, these songs tend show less variation in the expressed emotion.



(a) RMSE (b) Correlation (c) Sign-agreement

Fig. 5: Example valence trace of a song used in the experiment

V. Discussion

The conclusions of this paper can be split into two parts. The results from the study allow us to propose which metrics to use when evaluating music emotion prediction algorithms (Section V-A). The analysis of literature and minor observations from the study also encourage us to suggest some further guidelines for future work (Section V-B).

A. Choice of evaluation metrics

Our study indicates that RMSE should be used for optimizing algorithms to estimate one-dimensional models (Section IV-A), and that it is also the most appropriate metric for reporting results.

For two-dimensional models the situation is less straightforward. The analysis of the results from the third task (Section IV-B) indicate that both correlation and KL-divergence were equally preferred by our participants. As a choice between the two still needs to be made, we would suggest using KL-divergence, as it is more similar to the preferred choice for one-dimensional models.

All of the data we collected in this study is made available at http://www.cl.cam.ac.uk/ vi206/evaluation/

B. Other considerations

There are several other issues that might be worth considering when approaching the problem of emotion prediction.

First of all, the fact that RMSE was the preferred choice as an optimization metric identifies two things participants cared about. It seems that when judging the emotional content of a song, participants expect to see not only the relative change of emotion within a song (as correlation would suggest), but also the absolute position in the arousal-valence space. This has implications not only on the choice of evaluation metrics to use, but also on the kind of models that should be investigated in future work.

Another observation is that there was a (small) number of songs where sign-agreement was preferred over the other metrics (Section IV-C). It only seems to occur when there is little change in the expressed emotion of a song—in which case sign-agreement displays a flat line, while other metrics fluctuate around it. This suggests that a level of smoothing might be preferable when predicting emotion or when displaying the results. We also urge against using long correlation as an evaluation metric, as it hides important information about the performance of an algorithm (Section III-B) and does not seem to relate well to people's preferences (Section IV-C).

As it is possible to achieve good results in one metric while bad results in other metrics, we advise reporting the results using several metrics. This would give a better understanding of the general behaviour of an algorithm. In addition to that, we urge researchers to give the formulas of the metrics used in the evaluation. It is often not clear which exact evaluation metrics are used to describe the results (short versus long correlation, etc.), making it more difficult to compare different algorithms.

The conclusions we have reached and suggestions we have made can obviously only be directly applied to the field of emotion prediction in music. We believe that similar studies could and should be used to check if the same trends occur in other fields of affective computing. We expect that similar conclusions will be drawn, but the comparison across different fields will provide results that are interesting either way.

References

[1] S. O. C.-C. H. K. F. MacDorman, "Automatic Emotion Prediction of Song Excerpts: Index Construction, Algorithm Design, and Empirical Comparison," *Journal of New Music Research*, pp. 281–299, 2007.

[2] P. Evans and E. Schubert, "Relationships between expressed and felt emotions in music," *Musicae Scientiae*, vol. 12, no. 1, pp. 75–99, 2008.

[3] M. D. Korhonen, D. A. Clausi, and M. E. Jernigan, "Modeling emotional content of music using system identification," *IEEE transactions on systems man and cybernetics Part B Cybernetics*, pp. 588–599, 2006.

[4] K.-Y. E. Schmidt, Erik M, "Prediction of Time-Varying Musical Mood Distributions Using Kalman Filtering," *9th ICMLA*, pp. 655–660, 2010.

[5] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions from audio," in *Proc. of ISMIR*, 2010, pp. 465–470.

[6] E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *Proc. of ISMIR*. ACM, 2010, pp. 267–274.

[7] E. M. Schmidt and Y. E. Kim, "Modeling musical emotion dynamics with Conditional Random Fields," *Proc. of ISMIR*, pp. 777–782, 2011.

[8] B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "Avec 2012: the continuous audio/visual emotion challenge - an introduction," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, ser. ICMI '12. ACM, 2012, pp. 361–362.

[9] M. a. Nicolaou, H. Gunes, and M. Pantic, "Output-associative rvm regression for dimensional and continuous emotion prediction," *Image and Vision Computing*, vol. 30, no. 3, p. 186196, 2012.

[10] H. Gunes, M. A. Nicolaou, and M. Pantic, "Continuous analysis of affect from voice and face," in *Computer Analysis of Human Behaviour*, A. A. Salah and T. Gevers, Eds., 2011, pp. 255–291.

[11] R. A. Calvo and S. D'Mello, "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications," pp. 18–37, 2010.

[12] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based computer interfaces," *Journal of Neural Engineering*, vol. 4, no. 2, pp. R1–R13, 2007.

[13] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, 2008.

[14] ——, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," *Proc. 43st ACL*, 2005.

[15] Y. Mao and G. Lebanon, "Isotonic Conditional Random Fields and Local Sentiment Flow," *Advances in Neural Information Processing Systems 19*, vol. 19, no. April 2008, pp. 961–968, 2007.

[16] T. Wilson, J. Wiebe, and R. Hwa, "Just how mad are you? Finding strong and weak opinion clauses," *Science*, vol. 04, pp. 761–769, 2004.

[17] Y. Bestgen, "Can emotional valence in stories be determined from words?" *Cognition & Emotion*, vol. 8, no. 1, pp. 21–36, 1994.

[18] F. Dzogang, M.-j. Lesot, M. Rifqi, and B. Bouchon-meunier, "Analysis of a text's emotional content in a multidimensional space," *International Conference on Kansei Engineering and Emotion Research*, 2010.

[19] J. A. Speck, E. M. Schmidt, B. G. Morton, and Y. E. Kim, "A comparative study of collaborative vs. traditional music mood annotation," *Proc. of ISMIR*, pp. 549–554, 2011.